Resultant

TIMELY, ACTIONABLE INSIGHTS BUILT BY

Scalable Probabilistic Record Linkage

CONTENTS

INTRODUCTION

No business thrives without leveraging data for effective decision making. Without a comprehensive plan in place, data silos evolve separately across departments. Company mergers and acquisitions demand unification of databases. When organizations are unable to easily and accurately combine and analyze data from disparate sources, they don't gain actionable insight for intended outcomes.

Record linkage builds a comprehensive view of all relevant information pertaining to the same entity: person, business, or event.

Many record linkage solutions don't have the scalability to keep up with data that continues to grow at an aggressive pace. Often systems aren't structured to generalize and rely on overly complex, cumbersome, and incomplete rules.

Our unique approach to scalable, probabilistic record linkage frees organizations from this burden of insufficiency and opens a clear path to data-based actionable insights.

PROBABILISTIC VS. DETERMINISTIC RECORD LINKAGE

When joining data within or across databases, exact matches of all table keys are not always possible; however, individuals have a unique identifier, like an email address or driver's license number. Two records containing a common key can be joined by exact matching. In the absence of a unique identifier, records are linked based on the similarity of personal identification information (PII).

The traditional approach of developing rules for matching records between datasets is time-consuming, tedious, inefficient, and impractical as the quantity of records and data sources continually increases. In a system with numerous datasets from multiple sources, the effort required to develop rules that deterministically match all combinations grows exponentially. Then as new datasets are added, old rules must be revisited, making the overall process fragile.

Probabilistic record linkage does not require exact value matching. Our unique approach instead leverages fuzzy logic to identify matches. This configuration accounts for typos, nicknames, and partially missing PII that exact matching isn't designed to handle.

Scalable record linkage methodology allows organizations with disparate data and varying subsets of PII to find relationships. As new data is entered, duplicates are automatically identified within data silos to minimize error within the system.

We combine a custom algorithm with refined processes to probabilistically link large-scale datasets. New data silos with varying PII populated fields are automatically incorporated while optimizing thresholds within every relationship, whether at data silo or table level. When added to the system, they're leveraged not only to form new linkages but to add additional power and information to existing matches.

Probabilistic record linkage holds distinct advantages over deterministic matching because it takes into account

- Data quality issues including typos and misspellings
- Data incompleteness
- PII mismatch, as when one database collects date of birth while another records age
- Lifestyle changes in PII such as marriage or change of address

Accuracy and **Performance**

The probabilistic matching algorithm utilizes fuzzy hashing to link approximately equal PII records together. Hashing is a function that maps arbitrary-sized inputs to a structure with fixed-sized components; fuzzy is a way of compressing looking at byte-level similarities. The threshold of fuzzy hashing is adjusted to optimize both positive and negative uncertainty.

FALSE POSITIVE:

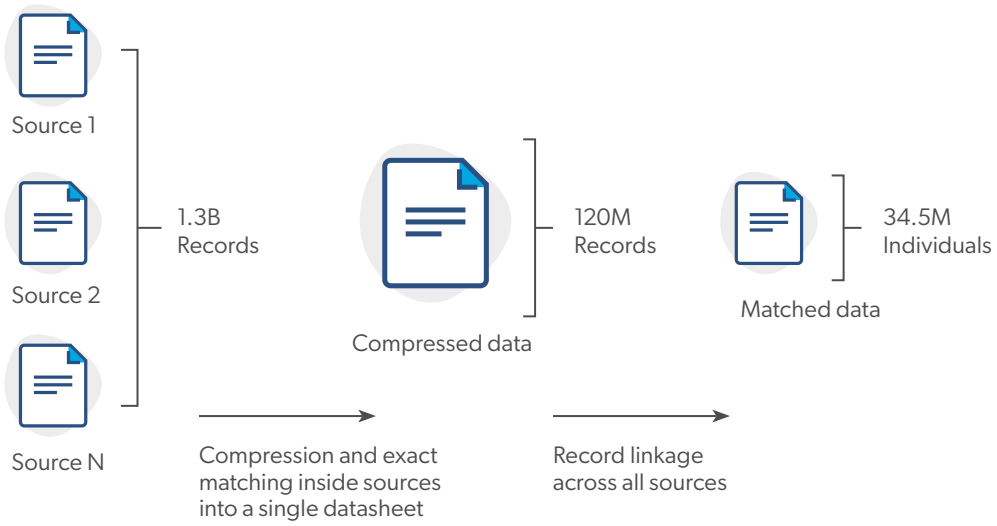
groups of records matched together yet having different unique identifiers like SSN and DOB. They're matched but are, in fact, separate entities.

FALSE NEGATIVE:

records that have matching unique identifiers but aren't connected. They're counted separately but should be matched.

These categories are interdependent: reducing one beyond a certain threshold results in increasing the other over the unacceptable limit.

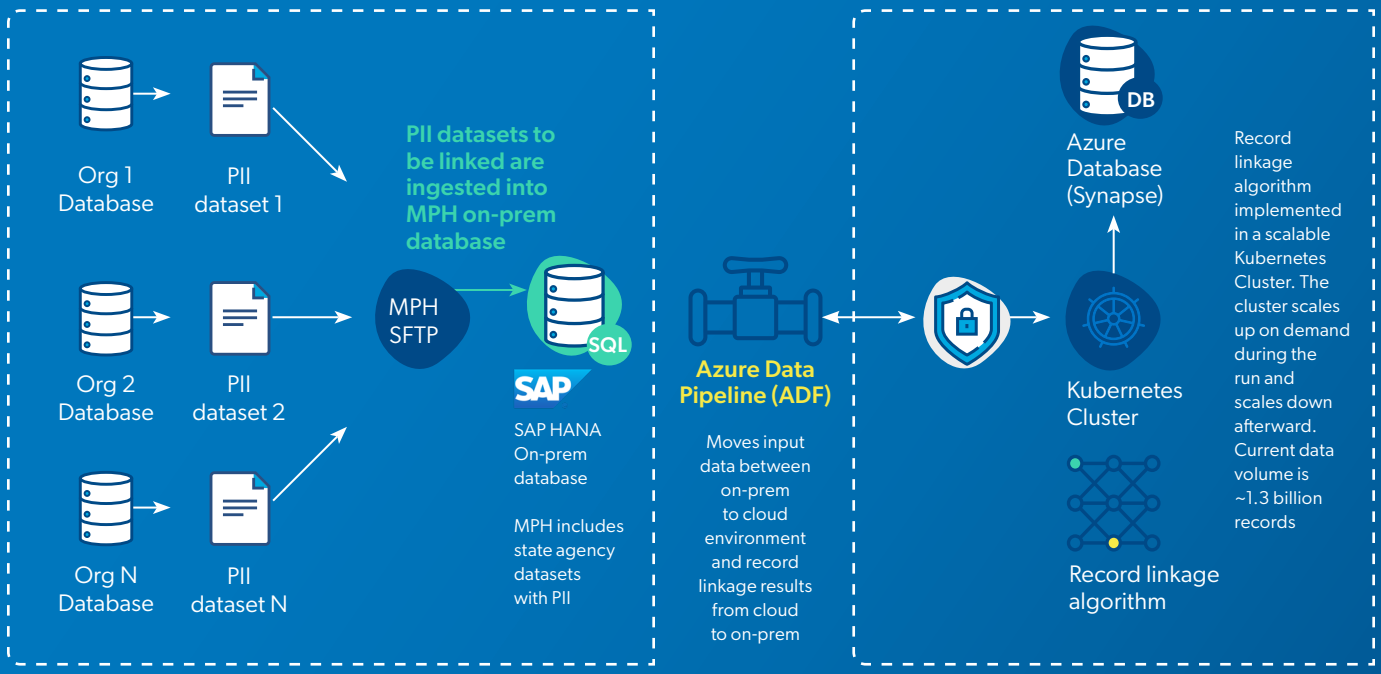
In the example of Indiana's Management Performance Hub (MPH), created with Resultant's collaboration, we adjusted the matching threshold so both false negative and false positive rest in an acceptable range. In its current environment, MPH has approximately 1.3 billion PII records that, when exact matching rules are applied, reduce to about 120 million unique records. Through record linkage, these are mapped to approximately 34.5 million unique individuals. Cloud record linkage provides a significant performance boost. To match 120 million records, preparation and running the algorithm takes less than fifty minutes. The end-to-end process, including automated QA, relevant metric generation, and environment scale-up is under an hour and a half.



ERROR TYPE	%	PROCESS	TIME
False Positive	**1.04%**	Algorithm implementations	**50 MIN**
False Negative	**0.83%**	Total end-to-end time	**90 MIN**

Architecture and **Methodology**

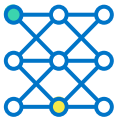
Where on-prem servers are utilized such as in Indiana's MPH, a hybrid structure leverages the existing security of the on-prem server and the flexibility of cloud record linkage. Data files from internal or selected external sources are transferred to an SFTP folder then ingested into the HANA server for staging. Staged PII data to be linked with other PII datasets is transferred to the Azure cloud server via a secure Azure data factory pipeline. Cloud server memory adjusts automatically as needed to accommodate the record linkage algorithm. The algorithm is applied to the data, identifying all PII records deemed to be of the same entity, and assigns a single unique identifier called a primary key. PII records not matching any other records or lacking necessary information to pass minimum criteria are referred to as singletons and remain as separate entities.



Records with updated primary keys are then transferred back to the on-prem data warehouse. For desired outcomes such as a deduplicated PII record or cross reference file, a custom use case view is prepared manually using these longitudinal data.

ORGANIZATION	ID	FIRST	LAST	DOB	SSN	SSN4	ZIP
Stage Agency 1	1234	John	Doe	11/25/80	123456789	6789	47908
Stage Agency 2	2345	John	D	11/25/80	123456789	6789	46240
Stage Agency 3	3456	John	Doe	12/25/80	123456789	6789	46240
Stage Agency 4	4444	John	D	NULL	NULL	NULL	NULL
Stage Agency 1	5555	Jane	Doe	8/9/82	23456789	7890	45346
Stage Agency 2	6666	J	NULL	1/1/80	NULL	NULL	47805

Input records include PII and system IDs



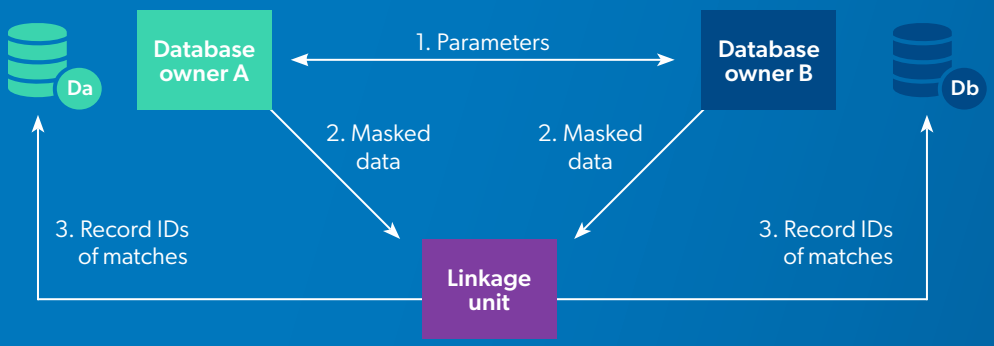
Record linkage algorithm

ORGANIZATION	PRIMARY KEY	ID	FIRST	LAST	DOB	SSN	SSN4	ZIP
Stage Agency 1	ID_1234	1234	John	Doe	11/25/80	123456789	6789	47908
Stage Agency 2	ID_1234	2345	John	D	11/25/80	123456789	6789	46240
Stage Agency 3	ID_1234	3456	John	Doe	12/25/80	123456789	6789	46240
Stage Agency 4	ID_9999	4444	John	D	NULL	NULL	NULL	NULL
Stage Agency 1	ID_9990	5555	Jane	Doe	8/9/82	23456789	7890	45346
Stage Agency 2	ID_9991	6666	J	NULL	1/1/80	NULL	NULL	47805

Output of record linkage: Similar records are grouped together and assigned the same primary key, and distinct and missing information records are assigned their own primary key

Hashed Secured **Record Linkage**

In some use cases, PII cannot be disclosed and needs to be obfuscated before sharing. A separate process for hashed record linkage makes this possible. A hashing tool is distributed to the outside entity to one-way hash their records using a salt, which adds data to the input for security; the hashed dataset and salt passes to the original system, where the same salt hashes repository records. The hashed version of the algorithm then matches the records. The outside agency does not have to disclose any information about its population unless a match is found in the main database. Post-linkage, the crossmatch along with any relevant metrics is prepared and sent.



Accuracy of hashed record linkage compared with non-hashed record linkage is quite similar: hashed record linkage has a slight drop in accuracy, as expected, but it's not significant enough to impact most use cases. For example, in a comparison test run at MPH data sources, the uncertainty is bounded at 2%, making the results as reliable as unhashed record linkage with an additional level of security. Unfiltered results of hashed record linkage should be used for research purposes, not for transactional system input.

A transactional system by purpose requires zero false positive matching; by design, generalized record linkage has minimal—though not zero—false positives. To utilize results from probabilistic record linkage in a transactional system requires an additional layer of data management implementing domain-based rules, thereby reducing false positives to that system's tolerance level.

Applications of Probabilistic **Record Linkage**

A probabilistic record linkage system provides a fast, accurate, and secure way to match nearly unlimited records over multiple organizations or departments. The results allow inferential analysis across populations. They can be used to develop longitudinal records of a person with different points of interaction, understand behavioral patterns, and conceptualize policy and practices implications on an aggregate level.

COVID-19 PANDEMIC RESPONSE, ADVANCED ANALYTICS, BI DASHBOARDS AND COLLABORATION

Resultant has played a crucial role in Indiana's COVID response, performing all advanced analytics projections for decision support, producing and maintaining public-facing and internal dashboards on the state of the crisis and response, building and maintaining the data back-end, and facilitating cross-organizational partnership for research and analysis—all possible because of probabilistic records linkage.

INDIANA DEPARTMENT OF CORRECTION

Corrections organizations have historically focused on the punitive aspects of incarceration, contributing to high recidivism rates—in many places, 60% or greater. Rehabilitation and training programs are now becoming a key component of incarceration. Resultant leveraged advanced machine learning and artificial intelligence techniques to enable Indiana's Department of Correction to use data on delivered programming and long-term outcome information to optimize programming for current offenders. The insight is delivered directly to case managers through an interactive software application, and relies on probabilistic record linkage through Indiana's MPH.

MODERNIZATION OF MEDICAID MANAGEMENT SYSTEM

Indiana's Family and Social Services Administration (FSSA) partnered with Resultant to ensure it achieved success of Core Medicaid Management Information System (CoreMMIS). The system is transformational for FSSA due to its broad-reaching nature, ensuring accurate information capture and reporting, controlling and administering costs, and serving recipients and providers. Because of the Resultant team's collaboration and implementation of a probabilistic records linkage solution, the system successfully launched on the date recommended following the assessment.

Ohio's Data Integration and Controlled **Substance Monitoring**

7,900%

increase in OARRS queries, according to the OARRS 2018 Annual Report Executive Summary.

After a successful pilot linking a large grocery store chain's pharmacy fulfillment system with statewide data, Resultant helped develop a statewide approach for the implementation of a controlled prescription database that utilizes probabilistic record linkage. Ohio now has the largest utilization of a controlled prescription database, allowing the state to take a proactive approach against the controlled substance crisis. They continue to carefully track progress with the Ohio Automated Rx Reporting System (OARRS). The OARRS 2018 Annual Report Executive Summary noted the following highlights:

- Decrease of 325 million doses of opioids dispensed.
- Decrease of 100 million doses of benzodiazepines dispensed.
- 7,900% increase in OARRS queries.
- 41,000 prescribers and pharmacists have integrated access to OARRS.
- 89% decrease in doctor shoppers.

ABOUT RESULTANT

Our team believes solutions are more valuable, transformative, and meaningful when reached together. Through outcomes built on solutions rooted in data analytics, technology, and digital transformation, Resultant serves as a true partner by solving problems with our clients, rather than for them.

DATA ANALYTICS

We help organizations understand their data landscape and solve problems by turning data into insight. While data can be dense, our team's empathetic approach to problem solving creates meaningful solutions with deep technical foundations.