Senzing

UNIQUELY SENZING

WHITE PAPER

NOTE: This paper is for readers familiar with entity resolution technologies and their specific use cases.

CONTENTS

Senzing® software is the first real-time, purpose-built artificial intelligence[1] (AI) for entity resolution[2] (ER). The plug-and-play Senzing ER autonomously discovers common entities and relationships within your data to provide you with a complete inventory of every record related to each person and company. The AI in Senzing software learns through experience and improves over use and time.

Senzing software makes ER easy and affordable. With Senzing ER, you can quickly provide single customer views while significantly reducing false positives and false negatives. Senzing ER allows your organization to improve financial and other compliance operations, watchlist processing, and fraud and insider threat detection, as well as marketing and other customer experience initiatives.

If you want to try Senzing software on your data, you can download the Desktop Eval Tool or developer SDK for free at senzing.com.

Senzing, Inc. receives no personal data.
Our software runs entirely on your cloud or on-premise computers.

Why Senzing

Organizations that choose Senzing ER will realize a rapid return on investment (ROI), achieve better outcomes and gain competitive advantages:

- Rapid ROI – Most organizations realize an ROI within the first year, some in less than 30 days. Senzing ER achieves significant cost savings because:
  - Fewer analysts are required as false positives (alerts) are reduced by up to 90%
  - No ER expertise is needed to train, tune and manage
  - Fewer engineering resources are required to cleanse, prepare and label data, up to 85% fewer
  - Affordably priced, up to one-tenth the cost of alternatives, and includes technical support
- Better outcomes – Senzing ER typically produces superior results to homegrown and other commercial ER systems. Culturally-aware name matching, Entity Centric Learning™ [3] and nonobvious relationship awareness are some of the powerful features that result in higher quality business decisions and outcomes (see below for more details).
- Competitive advantages – Senzing ER allows organizations to more easily innovate and leapfrog the competition. For example, you can implement real-time customer onboarding systems with continuous risk monitoring, or deploy privacy-enhanced information sharing that uses selective field hashing to reduce the risk of unintended disclosure.

1 We use the term AI to mean "systems that act smart." Some AI systems use machine learning (ML) meaning "learning through experience" and some don't. The AI in Senzing software learns through experience.

2 Entity resolution is the process of recognizing when two records relate to the same entity, despite having been described differently. And conversely, recognizing when two records do not relate to the same entity, despite having been described similarly. For more about entity resolution, read the blog https://senzing.com/to-know-entity-resolution-is-to-love-er/ or watch the video about ER explained step by step https://senzing.com/entity-resolution-explained/

3 https://senzing.com/entity-centric-learning-explained/

More than 250 years of our team's combined ER experience is built into Senzing software.

Organizations get significantly more advanced ER capabilities at one-tenth of the cost.

What Makes Senzing ER Unique

We have built more than 250 years of our team's combined ER experience into Senzing software. Below are details about what makes Senzing ER most unique and more advanced than other ER systems, specifically:

- Purpose-built AI for Entity Resolution
- Real-time operations
- Minimal data preparation
- Built-In Privacy by Design (PbD)
- Nonobvious relationship awareness
- Speed and scalability

Purpose-Built AI for Entity Resolution

Entity Resolution Market Overview

Entity resolution is essential to almost every industry, including banking, securities, insurance, healthcare, manufacturing, retail, marketing, telecommunications, energy and government.

Until Senzing ER, organizations that required ER had three suboptimal choices:

- Acquire Lightweight Data Quality and Integration Software – These technologies use rudimentary techniques, such as basic matching on exact ID numbers e.g., passport and social security numbers (SSNs). ER accuracy declines significantly when fuzzy matching is required e.g., on names and addresses. These solutions are often inexpensive, but still require significant configuration and tuning, ideally by an expert.
- Build Proprietary Entity Resolution Software – Developing robust ER systems in-house is expensive. Projects might require dozens of engineers or more, depending on data volumes, number of data sources and attributes. Some organizations we know of have more than 100 engineers working on in-house ER systems.
- Buy Legacy Commercial Entity Resolution Software – These systems require large budgets for software, hardware, and the professional services needed to configure, train and tune them. Deployments typically begin as consulting projects with budgets of $100,000 or more for proof of concepts. Production-ready systems start at $1M for software, hardware and services. It's not uncommon for organizations to pay many millions of dollars and spend a year or more to operationalize one of these systems.

With Senzing software, organizations get significantly more advanced ER capabilities at one-tenth of the cost of in-house or legacy commercial options. You can be up and running on Senzing ER in days or weeks, depending on the complexity of your data and interfaces.

For a list of important capabilities to consider when evaluating whether to buy or build enterprise-grade ER technology, check out our ER Capabilities to Consider paper.

The Senzing software team created a purpose-built AI for ER that includes two unique properties:

- The ability to make human-intelligent decisions on extremely small and extremely large data sets, without any pretraining or pretuning
- Gets smarter over time, as it autonomously learns and adapts in real time, without reloading

The AI in Senzing software is composed of two tightly coupled classes of algorithms: common sense and real-time machine learning. A 2018 story in *Wired*, How to Teach Artificial Intelligence Common Sense,[4] discusses the importance of this duality.

The AI is composed of two tightly coupled algorithm classes: common sense and real-time machine learning.

Common Sense in the Senzing AI

Unlike many AI machine learning techniques that must be initially trained using extremely large data sets, Senzing software comes pre-built with common sense that includes principle based entity resolution[5] and advanced knowledge.

Common sense allows Senzing ER to be smart on day one, even with data sets as small as two records. Common sense also helps Senzing ER ensure its real-time learning is not fooled by newly introduced anomalies e.g., mismapped fields or other errors.

Principle Based Entity Resolution

Principles are a special form of generalized knowledge that draw on common attribute behaviors.

The use of principles is a key reason Senzing software does not need training, tuning or experts to deploy into new domains or to add new data sets, new languages, etc.

The difference between the rules in some other ER systems and the principles in Senzing ER are distinct. Imagine telling your child to quit throwing rocks at cars. Only to realize the next day you have to tell him to quit throwing baseballs at SUVs. Then, a few days later, you have to tell him not to throw golf balls at trucks, fire engines and ambulances. Instead of all these rules, why not one simple principle: "Don't throw things at other people's stuff."

The principles in Senzing ER are based on expected attribute behaviors. For example, only one person should have an SSN, while many people can share the same date of birth (DOB), even though each person should only have one. Senzing software assigns these common-sense behaviors to attributes based on the following three expected behavior settings:

Senzing software comes pre-built with common sense AI that includes principle based ER and advanced knowledge.

- Frequency – does one, few or many entities generally share the same value e.g., an SSN is commonly used by one entity, an address is shared by a few, and a date of birth (DOB) is shared by many
- Exclusivity – does an entity typically have only one such value e.g., an entity should only have one SSN or DOB, but an entity could rightfully have more than one credit card number
- Stability – is the value typically stable over the lifetime of an entity or not e.g., an SSN and DOB are typically stable over a lifetime, but a home address is usually not

Notably, Senzing software recognizes that messy, real-world data may not always behave as expected. For example, if multiple people reportedly have the same SSN or one person has multiple DOBs, Senzing ER automatically detects these anomalies and adjusts accordingly.[6]

4 https://www.wired.com/story/how-to-teach-artificial-intelligence-common-sense/

5 https://senzing.com/principle-based-matching/

6 See the Anomaly Detection section below for more information.

Senzing software ships with approximately 30 default ER principles it uses to determine when entities are the same, possibly the same or related. Each principle considers the three feature behaviors described above plus names. Here are two examples of the types of principles built into Senzing ER:

- If entities have a close name and the same frequency one feature e.g., an SSN, they are likely the same entity
- If entities have a close name and a frequency few feature e.g., address, but have a contradictory exclusive feature e.g., different DOBs, they are considered related (not the same)

In a radical departure from other ER methods, the single set of default principles Senzing software provides automatically work as delivered for a wide range of entity types e.g., people, companies, vessels and planes.

The single set of default principles automatically work as delivered for a wide range of entity types.

Advanced Knowledge

The AI in Senzing software includes more than 10 pre-built comparison routines containing deep knowledge about specific attributes such as phone numbers, SSNs, dates, etc. Since culturally-aware name recognition and global address matching are most critical for achieving high quality ER, the comparators Senzing software uses for these attributes are particularly advanced.

- Global name recognition – IBM InfoSphere Global Name Management[7] comparison technology is built into Senzing software. This culturally-aware name library, pretrained on 800M global names, was created over decades by a team of linguists at a cost of tens of millions of dollars.

  Senzing AI immediately understands synonyms e.g., Bob and Robert or Elizabeth and Liz, and transliterations e.g., Mohamed, Mohammed, Mhd and dozens of other spellings. It also resolves names across different alphabets and scripts[8] e.g., Arabic, Mandarin and Roman.

- Global address comparison – Senzing ER uses libpostal[9], an open source library for global address parsing and normalization, to assess address similarity with uncanny precision. This library for programmers was trained, using machine learning, on the hundreds of millions of global addresses in the OpenStreetMap database.

  libpostal is embedded into Senzing software and wrapped with custom logic that provides exceptional matching accuracy and eliminates the need to pre-parse address data prior to loading.

More than 10 pre-built comparison routines contain deep knowledge about specific attributes.

The Senzing architecture allows advanced users to add additional attributes, such as height, weight, hair color, eye color, voice, fingerprints, etc., by writing custom plug-ins that standardize, express and compare new attribute data. For example, one source system may store height data in inches and another in centimeters. A set of custom plug-ins could automatically standardize height data into centimeters, create an expression of the data to the nearest tenth to help with matching.

7 https://www.ibm.com/us-en/marketplace/ibm-infosphere-global-name-management

8 https://senzing.zendesk.com/hc/en-us/articles/234766668-Globalization

9 https://senzing.com/what-is-libpostal/

Senzing ER gets smarter over time with real-time algorithms for entity-centric learning, anomaly detection and sequence neutrality.

Senzing software is always to up to date, overall error rates decrease over time, and reloading is never required.

Real Time Machine Learning in the Senzing AI

The AI in Senzing ER uses real-time machine learning (ML) to get smarter over time. The real-time algorithms deliver entity centric learning, anomaly detection, and sequence neutral processing.[10]

Entity Centric Learning

Senzing ER retains history and attribute variations for each entity as it resolves new records against existing entities e.g., learning every name, address and phone variation. Over time, based on the accumulated variations, the software learns nicknames, alternative email addresses, common typographical errors, etc., including intentionally fabricated information.

Senzing software uses its entity centric learning when comparing records during ER. Entity centric learning is what allows it to make higher quality ER decisions[11] than most other systems that use the more popular, but very basic, record-to-record matching. This is critical for catching clever criminals.[12]

Anomaly Detection

Senzing ER actively tracks feature statistics in real time using anomaly detection[13] as it resolves and relates entities. Based on the information it has seen to date, the software keeps detailed statistics about its entity repository, e.g., it contains approximately 150M males, 500 people with the same DOB, and exactly seven people who have lived at 123 Main Street.

By comparing actual statistics to expected feature behaviors, Senzing ER detects anomalies such as garbage values[14] e.g., if the SSN value 121212121 is used by hundreds of entities, the software recognizes this as an exception, since SSNs generally belong to one person. When such anomalies are detected, the software automatically self-tunes to account for them going forward by either assigning them less value or disregarding them altogether.

Sequence Neutrality (Self-Correcting the Past)

Based on what it learns about entities and anomalies, Senzing ER continuously evaluates its earlier assertions to determine if they need to be corrected. Sequence neutrality allows the software to self-correct the past in real time, whether it received record A first then B, or record B first then A.[15]

Humans self-correct all the time e.g., you think you know what someone means, but as they keep talking you realize they meant something else. The ability of Senzing ER to fix the past in real time at scale, as new data streams in, is extremely difficult to achieve.

Without sequence neutrality, the error rates of ER systems increase between the periodic reloads required to bring them up to date. With the sequence-neutral processing in Senzing software, your system is always to up to date, overall error rates decrease over time as new information reverses earlier assertions, and reloading is never required.

10 https://senzing.com/sequence-neutrality/

11 Watch the video about ER explained step by step https://senzing.com/entity-resolution-explained/

12 https://senzing.com/catching-bad-guys-with-entity-centric-learning/

13 https://senzing.com/purpose-built-ai/

14 Garbage values might include phone numbers or addresses with values such as S300, NA , or "no known value."

15 Review ER step by step for record 7 in the video https://youtu.be/VFE3kGdoXzA?si=yZOwpoK0Yr86Pld1&t=246

Real-Time Operations

The Senzing team made significant design decisions early on to ensure the software natively supports real-time operations, including the following:

- Real-time adds and changes – immediately resolves new data as it is received
- Real-time queries – instantly delivers resolved entity data resulting from user queries
- Real-time decision systems – supports real-time business transactions by instantaneously providing systems with resolved entity data
- Real-time deletion – immediately removes data and the consequences of that data e.g., right to be forgotten requests required by emerging privacy regulations
- Real-time maintenance – performs scheduled or emergency maintenance on live systems to eliminate downtime or the need to run two instances of an operational system
- Real-time replication and publishing – replicates its database of resolved entities to data marts or data warehouses

Senzing ER is a true on-line transaction processing (OLTP) engine that reduces operational risks and helps ensure organizations never have to make decisions based on outdated entity information.

The Senzing team made significant design decisions early on to ensure the software natively supports real-time operations.

Minimal Data Preparation

Senzing ER is more tolerant of messy and structurally inconsistent data than traditional systems. You just map data as you find it in your source systems and the software takes care of the rest. For example, some source systems may standardize names and addresses while others don't, or DOB formats may be different across source systems.

Senzing ER has its own data standardization and parsing routines that perform comparisons on the same data attributes, even if those attributes are structurally inconsistent within source systems. The software even matches across different scripts such as Roman, Mandarin, Cyrillic, etc.

These capabilities deliver significant advantages. Some customers report that Senzing ER is eight times faster than other legacy commercial ER systems when preparing and loading data. We estimate that Senzing software eliminates up to 85 percent of the typical data cleansing, preparation and transformation tasks required by other ER systems.

We estimate that Senzing software eliminates up to 85 percent of the typical data cleansing, preparation and transformation tasks.

Built-in Privacy by Design

Senzing technology was built with Privacy by Design[16] (PbD) in mind from inception and includes:

- Full attribution – Senzing software retains each original record, so it always knows where every record came from and when. Each record is stored with a pointer to its source system and record identifier. Full attribution ensures system-to-system reconciliation and audits, and enables other PbD features.
- False negative favoring – Senzing software is designed to favor false negatives i.e., err toward not asserting a match or relationship that may actually exist. This approach to ER, from a civil liberties standpoint, is preferred to the alternative of favoring false positives, i.e., inadvertently making incorrect ER assertions.

  Why? False positives can adversely affect people's lives, e.g., an innocent person is investigated. The Senzing algorithms favor false negatives by default, but you can adjust as needed e.g., for less sensitive marketing use cases or investigations where humans are in the loop.
- Self-correcting false positives – When new records arrive that demonstrate prior assertions are no longer correct, Senzing software automatically self-corrects them in real time. For example, if two records are initially asserted as one entity, because they shared the exact same name, address and home phone number, but later new information is received that confirms the records are actually for a junior and senior, the software instantly self-corrects the earlier assertion by separating the single entity into two.
- Data tethering – Senzing software processes adds, changes and deletes made in source systems in real time to ensure data is always current. Support for data tethering is especially important when Senzing entity data is used to make decisions that affect people's freedoms or privileges. For example, if someone is removed from a watch list, their name is cleared in a downstream Senzing system in real time.
- Selective field hashing – Senzing software can perform ER on attributes that are cryptographically hashed at the source system, before information is transferred. The software is able to perform fuzzy-like ER, with similar results to ER using clear text, because it parses and standardizes attributes before hashing them. The ability to use selective field hashing to reduce the risk of unintended disclosure is unique to Senzing ER.

As privacy laws become more pervasive, PbD will increasingly be legally required for new technology deployments. Senzing software is the only ER technology with built-in PbD. For more information, read the blog post Privacy by Design (PbD) and Senzing.[17]

Full attribution ensures system-to-system reconciliation and audits, and enables other PbD features.

Senzing software is the only ER technology with built-in Privacy by Design to meet new legal requirements.

16 "Privacy by design is an approach to systems engineering initially developed by Ann Cavoukian and formalized in a joint report on privacy-enhancing technologies by a joint team of the Information and Privacy Commissioner of Ontario (Canada), the Dutch Data Protection Authority and the Netherlands Organisation for Applied Scientific Research in 1995 … Privacy by design calls for privacy to be taken into account throughout the whole engineering process."

17 https://senzing.com/privacy-by-design-pbd-in-senzing/

Relationship Awareness

Not only does Senzing software resolve entities, it also identifies, maintains and manages the following types of relationships between entities:

- Disclosed natural relationships, based on provided information e.g., a guarantor on a credit application or an emergency contact on an employment application
- Derived relationships e.g., family members sharing an email address
- Possible matches where there is not enough information yet to establish certainty
- Ambiguous matches, an exotic type of possible match unique to Senzing software, where records have more than one perfectly matching entity. For example, a record containing a name and address for Pat Jones when there is a Patrick Jones and a Patricia Jones at the same address (in this example, Patrick or Patricia could be the right answer, but picking one would be arbitrary)

As record volumes grow to billions, it is easy to scale out Senzing software across heterogeneous clusters.

Speed and Scalability

The Senzing team spent 12 months, over 2009 and 2010, designing and proving out a database schema to ensure Senzing ER supports unprecedented speed, scalability and flexibility. Because of this work, Senzing software runs up to 400M records on a $5,000 commodity server and performs millions of new entity resolutions a day in real time with sub-second response rates, without ever reloading.

Senzing ER is specifically designed to scale vertically and horizontally in cloud computing infrastructures. As record volumes grow to billions, it is easy to scale out across heterogeneous clusters.

Getting Started with Senzing ER

Does your organization desperately need to resolve entities to gain new insights and make better decisions faster? Senzing makes it possible with the first plug-and-play, real-time AI for ER and the most advanced ER software available.

Don't take our word for it. See for yourself. Download the Senzing[18] Desktop Eval Tool or developer SDK today for free. With Senzing ER, it's easy to get started and quickly see the results of ER on your own data.

You can use Senzing ER to enhance and transform your big data analytics, fraud operations, insider threat, marketing intelligence, risk mitigation and other operations. Senzing software requires minimal data preparation and transformation, and no training or tuning is required. And, you can rest assured that your system will scale up from a laptop to huge cloud clusters to support your most demanding, mission critical needs.

You don't need a million-dollar-plus budget, expensive ER experts, or a large number of IT resources to deploy Senzing ER .

You don't need a million-dollar-plus budget, expensive entity resolution experts, or a large number of IT resources to deploy Senzing ER. Senzing software is subscription based, so there are no lock ins, you can pay as you go, and subscriptions include technical support.

For more information visit www.senzing.com

18 https://senzing.com/try-senzing/