Senzing

ENTITY RESOLUTION CAPABILITIES TO CONSIDER

This technical overview provides a list of important capabilities to consider when evaluating whether to buy or build enterprise-grade entity resolution (ER) technology. While many of these capabilities are obvious, some are often overlooked. The table below summarizes the capabilities covered in this document and the benefits of each. In the three columns on the right:

- Accuracy indicates fewer false positives, fewer false negatives, or a higher overall F1 score.
- Operations refers to efficiencies in day-to-day administration and the ability to easily adapt to evolving requirements.
- Cost savings designates a faster return on investment (ROI).

SUMMARY OF ER CAPABILITIES TO CONSIDER

CATEGORY	CAPABILITY	ACCURACY	OPERATIONS	$ SAVINGS
EASE OF USE	Plug & Play ER		✓	✓
	No ER Experts Required		✓	✓
ACCURACY	Advanced Name Comparison	✓		
	Advanced Address Comparison	✓		✓
	Entity-Centric Learning	✓		
	Auto-Generics Detection	✓	✓	✓
	Ambiguous Detection	✓		
	Sequence Neutrality	✓	✓	✓
REAL TIME	Streaming ER	✓	✓	✓
	Low Latency		✓	
	Active Maintenance	✓	✓	✓
PRIVACY BY DESIGN	Full Attribution		✓	
	Selective Field Hashing		✓	
DEVELOPER FOCUSED	No New Moving Parts		✓	✓
	Cloud-Ready by Design		✓	✓
	User-Extensible Comparators		✓	
	Open Source		✓	✓
OPERATIONAL IMPACT	Minimal Data Preparation		✓	✓
	Explainable Matching		✓	
	No System Reload	✓	✓	✓
RELATIONSHIP AWARENESS	Disclosed Relationships	✓	✓	
	Derived Relationships	✓	✓	
	Find Path		✓	
GLOBALIZATION	Unicode Support	✓		
	Culturally Aware Analytics	✓		

Ease of Use

ER technology has historically been difficult to build and use. This was partially due to the fact that a diverse range of expertise was required, including knowledge of statistics, linguistics and performance engineering to name a few. With easy-to-use ER, organizations of all sizes – not just the elite with huge budgets – can affordably evaluate and deploy ER.

CAPABILITY	BUSINESS VALUE	@SENZING
Plug & Play ER Run ER out-of-the-box with little or no configuration changes – regardless of industry or use case.	Eliminates the time consuming and expensive training and tuning most ER systems require. Training data, which many organizations lack in sufficient quantities, is also not required.	Senzing can perform ER on most data sources automatically. Users can get great results on both small and large data sets without any training or tuning.
No ER Experts Required Validate, deploy and operate ER without ER experts.	Reduces the cost of deploying and operating ER by eliminating the need for expensive and scarce ER experts.	No ER experts are needed to validate, deploy or operate Senzing.

Senzing uses a radically different approach to ER that delivers unmatched ease of use and lightning fast ROI. The technology is a result of a 2009 IBM skunkworks initiative, code-named G2, that was subsequently spun-out of IBM in 2016 to form Senzing.

Accuracy

The number of false positives and false negatives and their combined F1 score defines the accuracy of ER systems. Robustness of a system's name, address, and other feature comparison techniques and the underlying matching method are two of the main contributors to ER accuracy.

CAPABILITY	BUSINESS VALUE	@SENZING
Advanced Name Comparison Compare names in a culturally-aware way including support for transliteration and script differences.	Sophisticated handling and comparison of personal and company names results in more accurate ER.	Best-in-class IBM Global Name Management[1] software is built into Senzing.
Advanced Address Comparison Match messy addresses in both structured and unstructured formats.	Sophisticated handling and comparison of addresses results in more accurate ER.	Libpostal,[2] an open source library for global address parsing and normalization, is built into Senzing.
Entity-Centric Learning Resolved entities are treated holistically as a single entity during ER.	Enables ER to get more accurate over time e.g., learning every name and address variation. Essential for detecting channel separation,[3] the primary tradecraft of clever bad actors.	Senzing has spent decades perfecting entity-centric learning.[4]
Auto-Generics Detection Detect widely used erroneous data during real-time ER.	Eliminates overmatching when data contains generic values e.g., an 800-555-1212 phone number. Real-time detection ensures best levels of accuracy 7x24 by avoiding the typical need for retraining, retuning and reloading to correct for generic values.	Senzing learns generic values in real time, automatically reevaluates and corrects past decisions as needed, and makes better decisions in the future.
Ambiguous Detection Handle new, uncertain records with special care.	Detection and proper handling of ambiguous records is essential for accuracy. For example, George Foreman had five sons who were all named George, so any record containing the name George Foreman with a home address and home phone could belong to one of the sons or the father. Most ER systems are less accurate because they arbitrarily resolve ambiguous records to any suitable match.	Senzing identifies ambiguous[5] conditions in real time. This is nontrivial when combined with sequence neutrality at massive scale.
Sequence Neutrality Regardless of the order data is loaded, every new record is used to reevaluate and improve prior ER decisions. Handle new, uncertain records with special care.	Improves accuracy in real time, e.g., decoupling previously resolved records after a new record reveals it was actually a father and son. Without sequence neutrality, accuracy degrades over time and periodic maintenance, e.g., database reloading is required.	Sequence neutrality[6] is the most sophisticated form of real-time learning Senzing offers. This capability is computationally nontrivial to perform in real time at massive scale.

When used with real data, Senzing routinely produces more accurate results than humans. Senzing is so accurate, organizations can also use it to quickly audit the accuracy of their existing ER algorithms.

Real Time

Many organizations benefit from batch-based ER today but recognize they will need real-time ER in the future to remain competitive. Transforming a homegrown batch-based ER system into a real-time system is not possible without significant reengineering. Since real-time systems also support batch data, we recommend choosing an ER system that is natively real time, even if real-time capabilities are not needed today. This will ensure readiness for real-time ER at a moment's notice.

CAPABILITY	BUSINESS VALUE	@SENZING
Streaming ER Load and resolve new records transactionally in real time.	Allows organizations to make higher quality decisions based on real-time ER e.g., while onboarding a customer or as a money wire is being created.	Senzing both resolves and relates new records in in real time.
Low Latency Perform ER fast enough to be used in real-time decisioning systems.	Higher quality decisions are made when ER is performed while transactions are happening, e.g., during a point of sale transaction.	Senzing typically performs ER in sub-200 milliseconds and delivers lookup times even faster. Optionally, users can realize sub-10 millisecond latency using the Senzing real-time data mart replication architecture.
Active Maintenance Handle maintenance activities in real time while the system is supporting normal operations.	Eliminates the need to go offline to perform maintenance e.g., loading new data sets, deleting historical data, retraining models. Offline maintenance, required by most ER systems, is expensive e.g., may require mirrored systems – one active while the other reloads.	Senzing supports active maintenance, so streaming ER, interactive query and other operational workloads can remain online while maintenance is performed.

The engineering effort to create Senzing, the first real-time AI for ER, began with a one-year project to design its underlying database schema. Senzing's unique schema allows our real-time learning algorithms to deliver unmatched accuracy over billion-record systems with low latency.

Privacy by Design

Privacy by design[7] (PbD) is an approach to systems engineering initially developed by Ann Cavoukian.[8] The framework was published in 2009 and adopted as a standard by the International Assembly of Privacy Commissioners and Data Protection Authorities in 2010. PbD calls for the consideration of privacy throughout the entire engineering process.

CAPABILITY	BUSINESS VALUE	@SENZING
Full Attribution Details about where every record came from (source system and record locator) and when. No data survivorship i.e., records or fields are never discarded.	Enables active maintenance and sequence neutrality, as described above, and ensures an ER system can be 100% reconciled with source systems.	Senzing uses full attribution.[9]
Selective Field Hashing One-way hashed[10] attributes created at the system of record before being submitted to the ER process.	Reduces the risk of unintended disclosure because hashed values submitted to and stored in the ER database are cryptographically obscured.	Senzing supports selective field hashing.[11]

The Senzing team has been dedicated to PbD since inception.[12] The underlying technology in Senzing (code-named G2) was first announced publicly in 2011 on Data Privacy Day.

Developer Focused

Commercially available ER has not historically focused on developers. Today, many developers, data engineers and data scientists are looking for ways to quickly add ER to their projects with easy to use, componentized technology.

CAPABILITY	BUSINESS VALUE	@SENZING
No New Moving Parts Link ER directly into applications.	Enables active maintenance and sequence neutrality, as described above, and ensures an ER system can be 100% reconciled with source systems.	Senzing is a lightweight embeddable library that provides C, Java, and Python interfaces to enable developers to instantly add ER into their projects.
Cloud-Ready by Design Deploy and scale ER in modern cloud infrastructures.	Reduces the risk of unintended disclosure because hashed values submitted to and stored in the ER database are cryptographically obscured.	Senzing supports a wide range of cloud deployment options, ranging from REST APIs to Kubernetes, AWS ECS, OpenShift, etc. These building blocks are available in source code for use, modification, security scans, etc.
User-Extensible Comparators Add new comparison functions.	Enables rapid responses to changing market needs, product innovation, etc. by enhancing or adding feature comparison functions e.g., adding support for comparing height or voice.	Senzing allows developers to use C to create new feature comparators.
Open Source Leverage and contribute to the open source community to increase freedom of action.	Accelerates ROI by providing assets that can be immediately viewed, used, modified or extended to decrease deployment times. Allows developers and integrators to make changes on their timetables rather than being dependent on a vendor's roadmap.	Unlike traditional open source, Senzing provides open source "inside out." The Senzing core is lightweight, configurable, extensible and proprietary, but 100+ assets are publicly available on GitHub[13] e.g., the REST API server. This approach allows developers and other users to leverage the specific assets they need to accelerate application development and success.

Senzing is unique in that it makes the complicated task of ER easy for developers, whether it is deployed in the cloud or on-prem using Kubernetes or Docker, or on bare metal.

Operational Impact

Human resources are required to operate production ER deployments e.g., for database maintenance, onboarding new data sources, and technical support to users. Some ER capabilities reduce operational expenses while making ER more agile and responsive to shifting markets and enterprise innovation goals.

CAPABILITY	BUSINESS VALUE	@SENZING
Minimal Data Preparation Prepare data for ER processing with less effort.	Reduces costs by eliminating most tasks normally required to pre-structure and pre-clean data e.g., formatting names and addresses or filtering out such values as "UNK"	Senzing eliminates up to 85% of the time it takes to prepare data for ER.
Explainable Matching Simple to review and understand reasons why records matched or did not.	Answers inquiries quickly and clearly from end users, auditors, senior management, regulators or attorneys who may want to know exactly why records matched or did not match.	The Senzing API includes "Why" and "Why Not" functions that can be called natively by applications or summoned via our open source Exploratory Data Analysis (EDA)[14] tools.
No System Reload No need to go offline for routine maintenance.	Eliminates the need to maintain a mirrored system, which is a costly necessity if one ER system needs to be online while another is reloaded.	Senzing is a true online transaction processing[15] (OLTP) engine that never has to be reloaded.

The operational cost of Senzing is dramatically less than other ER technologies. Sometimes operational cost savings alone justify the entire return on investment (ROI) for Senzing. An example is ERIC[16], a nonprofit organization modernizing voter registration in America. ERIC has a Senzing ER system that contains more than 350M records representing two thirds of America's voters in 30 states. Until recently, ERIC had an IT department of just one person managing Senzing and all other IT requirements. More details on ERIC can be found in the New York Times story Another Use for A.I.: Finding Millions of Unregistered Voters.[17]

Relationship Awareness

Most entity resolution methods perform only basic ER to determine who is who. When a system also identifies relationships, or who is relate to whom, the ER results are much more useful.

CAPABILITY	BUSINESS VALUE	@SENZING
Disclosed Relationships Relationships between entities that are certain because the knowledge of their existence comes from a statement of fact.	Makes it easier to accurately estimate opportunity and risk by providing a better understanding of how entities relate to each other e.g., business ownership hierarchies disclosed when a business opens a bank account.	Senzing supports disclosed relationships.
Derived Relationships Undisclosed, sometimes hidden, relationships between entities detected during ER.	Increases opportunities and reduces risk by identifying when entities are likely to know each other in an intimate way e.g., members of the same household.	Senzing identifies derived relationships.
Find Path Explore the entity graph to reveal connections between distant entities.	Enhances investigations and decision-making by allowing users or analytics to uncover hidden relationships that are impractical to discover manually.	Senzing provides several Find Path API calls.

Senzing builds, persists, and manages relationships between entities in an entity-resolved graph in real time at scale. Relationships are available transactionally during streaming ER or ad hoc via the Senzing API.

Globalization

Data in an ER system rarely involves a single culture. It is essential, an ethical obligation, to be able to perform accurate ER over culturally-diverse data.

CAPABILITY	BUSINESS VALUE	@SENZING
Unicode Support Handle international character sets including double-byte characters.	Enables data using multiple scripts, including non-Roman scripts e.g., Chinese, Arabic, to be resolved within the same system.	Senzing stores data in UTF-8 to provide full support for Unicode.
Culturally Aware Analytics Leverage culture when performing entity analytics.	Allows international data to be resolved accurately. One example of culturally-specific analytics is parsing addresses in Singapore differently from addresses in India.	Senzing maintains and leverages regional domain knowledge during ER processing e.g., name culture, address parsing.

The Senzing team has been supporting culturally-aware ER for decades. Our focus is reflected in the accurate outcomes Senzing API delivers from culturally diverse data.

Conclusion

An organization may not require all of the ER capabilities, or the scale and performance, discussed above. Yet all organizations benefit to some degree when such capabilities are available in an easy-to-use offering at an affordable price.

At Senzing, we are making this possible. We have literally dedicated most of our lives to this mission.

We hope you enjoy our technology.

[1] https://www.ibm.com/products/ibm-infosphere-global-name-management
[2] https://www.mapzen.com/blog/inside-libpostal/
[3] https://senzing.com/channel-separation-the-primary-tradecraft-of-clever-bad-guys/
[4] https://senzing.com/entity-centric-learning-vs-record-matching/
[5] https://senzing.com/ambiguous-conditions-in-entity-resolution-systems/
[6] https://senzing.com/sequence-neutrality/
[7] https://en.wikipedia.org/wiki/Privacy_by_design
[8] https://en.wikipedia.org/wiki/Ann_Cavoukian
[9] https://senzing.com/privacy-by-design-pbd-in-senzing/
[10] https://en.wikipedia.org/wiki/Cryptographic_hash_function
[11] https://senzing.com/discovery-without-disclosure/
[12] https://medium.com/@jeffjonas/privacy-by-design-pbd-and-senzing-eac8deb6c11f
[13] Open source made available under the Apache 2 license.
[14] https://senzing.zendesk.com/hc/en-us/articles/360050643034-Exploratory-Data-Analysis-4-Comparing-ER-results
[15] https://en.wikipedia.org/wiki/Online_transaction_processing
[16] http://www.ericstates.org/
[17] https://www.nytimes.com/2018/11/05/technology/unregistered-voter-rolls.html