A DATAWALK SOLUTION GUIDE

The Village Always Knew.

A solution guide to Native Entity Resolution inside the investigation graph.



DataWalk resolves the subject inside the enterprise knowledge graph investigators already use

DataWalk

2026 edition

SUMMARY

The Risk Is Between The Records

One version of the name is in one system. The old address is in another. A second passport is attached to a third record.

1 question underneath every case, every alert, and every screening run: which records refer to the same real person or organization?

That is where risk hides. One person becomes several customers. One company becomes separate counterparties. Alerts, investigations, and decisions run against fragments while the connection between them stays unseen.

Entity resolution closes that gap. It decides which records refer to the same real person or organization. But the match is not the destination.

Many specialist entity resolution products stop once they decide which records belong together. DataWalk carries the resolved subject into the graph where investigators search, follow relationships, assess risk, and act. That is what native means here: the resolved entity lives inside the platform where the investigation happens.

It keeps pace with changing data.
Every match can show its work.
It lives where the investigation lives.

This guide explains those three claims, and how to test any vendor against them.

A note on claims

Every performance figure in this guide describes a measured DataWalk run or deployment, sourced on the back page. Results vary with data, hardware, and configuration. Where the honest answer is a limitation rather than a promise, we state it.

WHY IDENTITY MATTERS

The Village Always Knew

A knowledge graph, without the jargon.

A village did not know people as records. It knew who married whom, whose son worked at the bank, who owned both shops, and which unfamiliar name belonged to an old neighbor.

Nobody reconstructed those relationships for each question. The knowledge was shared, updated as life changed, and available the moment a decision had to be made.

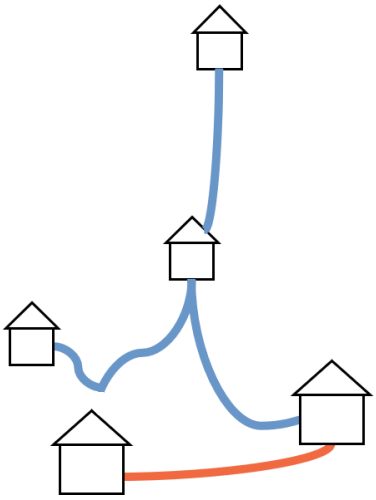
Your organization holds the same kind of knowledge, but it has filed it apart in artificial silos. The account sits in one system. The directorship sits in another. The address sits in a third. Ask about a name and each system returns its own record. None returns the neighborhood around the person.

A knowledge graph restores that neighborhood. It holds people, companies, accounts, addresses, and events together through their relationships.

But the knowledge graph has one dependency. It must recognize that Maja Nowak, Dr M. M. Nowak, and Maja Marie Nowak are the same person. Otherwise it builds three neighborhoods around one subject and leaves the real connections divided between them.

The village never separated identity from the relationships around it. Identity and relationships belong in the same investigation platform.

The facts were not missing. The subject had been split apart.



WHY IDENTITY MATTERS

The Village Got Big

One person can now arrive under seven names, three systems, and two alphabets.

At enterprise scale, identity splinters continuously. A name is shortened in one source and transliterated in another. A customer changes address. A company trades under a different name. An old passport stays attached to one record while the new number appears somewhere else.

The organization still knows the subject. It is the conflicting systems that no longer agree on who the subject is.

What replaced village knowledge is a collection of coping strategies: investigators searching systems one at a time, stewards merging duplicates pair by pair, and matching scripts maintained long after their author has left.

Entity resolution can identify which records belong together. The more important question is what happens to that answer next.

Does it remain current as new data arrives?

Can somebody explain why the records matched?

Does the resolved subject become part of the investigation graph, or return as another file, identifier, or database to reconcile?

That is the distinction this guide examines.

Not entity resolution versus no entity resolution.
Entity resolution as a separate match engine versus entity resolution as a native property of the platform where intelligence is gathered, evidence is explored, and decisions are made.

Three questions to ask any ER vendor

- Where does the resolved entity live?
- What happens when the data or the matching rules change?
- Can the system explain both a match and a non-match?

THE QUESTION AFTER THE MATCH

Most ER Ends At The Match

An entity resolution engine can tell you that three records belong to the same person.

Useful. But that is where the investigation begins.

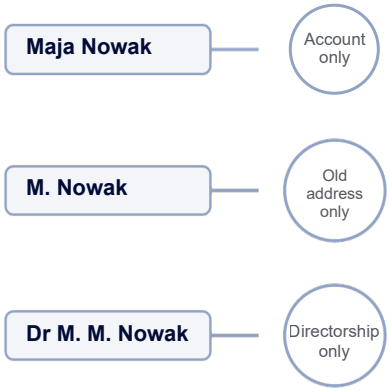
Which accounts belong to the subject? Which companies share the old address? Was the person named in a prior case? Does the new record expose a relationship that was previously split across two identities?

A separate match engine typically produces a cluster, an identifier, or a file. That result still has to be carried into the investigation model, governed there, connected to the surrounding records, and kept synchronized as the data changes.

DataWalk resolves the entity inside the enterprise knowledge graph investigators already use.

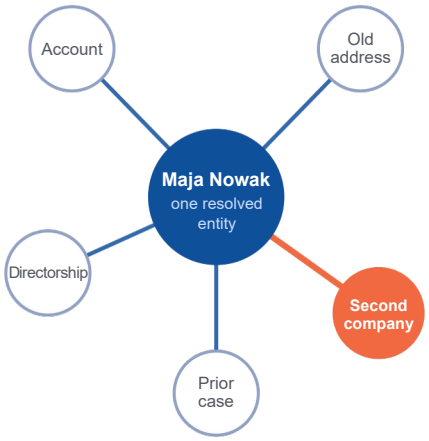
Maja Nowak, M. Nowak, and Dr Maja Marie Nowak become one governed entity. Their accounts, addresses, directorships, alerts, and prior cases now form one connected neighborhood. A relationship that previously required three searches becomes one traversal.

BEFORE



Three records,
Three partial neighborhoods.

AFTER



One resolved entity, one connected neighborhood.
Orange: the link nobody saw.

The result is not handed to the investigation. It is already inside it.

The match identifies the subject. The knowledge graph reveals the case.

Start With The Subject, Not The Search

Most investigation tools assume that identity has already been prepared

DataWalk starts earlier.

A new name enters the investigation. Instead of searching each source, trying alternative spellings, and hoping the decisive record appears, the investigator checks the subject against the resolved estate.

Candidate identities return with the evidence behind them. If the subject is already known, the wider neighborhood is available immediately: accounts, addresses, companies, alerts, and prior cases.

The same change improves screening. Seven spellings no longer behave like seven unrelated people. Alerts can be evaluated against the resolved entity while every underlying record remains available behind it.

The investigator still decides what the connections mean. DataWalk makes sure the investigation begins with what the organization already knows.

WHAT IF YOU DIDN'T HAVE TO SEARCH EVERY SYSTEM?

Enter a name. DataWalk checks it against the resolved data already in the platform.

Before the investigation begins, the investigator sees possible matches, why they may be the same person, and the records connected to them.

80% fewer false alerts, retaining more than 90% of true alerts: independent research on real-world banking data, using entity and network context. [5]

6.6–18 seconds server-side in single-request tests against a 212.8-million-entity resolved corpus.[3]

One subject. One place to start.

ONE ENTITY. EVERY CLUE.

Resolve The Subject. Keep The Evidence.

Bringing records together should not mean flattening them.

A golden record chooses the preferred name, the current address, and the document judged most authoritative. That is useful when the purpose is to service a customer through one clean profile.

An investigation has a different requirement.

A previous name or alternate spelling may connect the subject to an earlier filing. The old address may link two companies. The second passport may explain why a search under the current document found nothing. The values rejected by survivorship rules can be the values that reveal the network.

DataWalk keeps the source records linked through the resolved entity. Investigators work with one subject, but can still open the aliases, historical values, documents, and source evidence that contributed to it.

The same principle applies when a watchlist or external file is checked. Each candidate returns with a verdict, a score, and the evidence behind it. An uncertain result remains available for review rather than being silently merged or discarded.

Operations get a usable entity. Investigators keep the trail.

WHAT IF NOTHING HAD TO BE THROWN AWAY?

DataWalk resolves the subject without flattening the evidence.

The preferred view stays simple. Behind it, every name variant, address, document, and source record remains available for investigation.

When a list is checked, each result returns with a verdict and the evidence behind it. “No match” means no supported match was found under the configured rules.

One subject. The source trail stays visible.

RESOLUTION THAT KEEPS UP

The Data Changes. The Answer Changes With It.

Entity resolution fails when it is treated as a cleanup project.

A batch produces a clean view of the data on Friday. Then a new source arrives, a customer moves, a company changes its trading name, or another version of the subject enters through a different system.

By Monday, the result has already started to drift.

DataWalk treats resolution as maintained state. The first run resolves the estate. After that, change detection identifies what changed and reprocesses the delta rather than rebuilding the complete base. New records extend the entities already in the graph. They do not begin another reconciliation project.

The matching logic can evolve too. Rules, evidence weights, and thresholds are configuration rather than code buried inside a one-off implementation. Teams can test a changed setup against a sample, compare the results, and publish it when the new configuration performs better.

The result is not a queue that periodically gets cleared. It is a resolved graph that moves with the data.

WHAT IF YOU COULD TRUST THE RESOLVED VIEW TO STAY CURRENT?

New data updates the resolved view instead of creating another cleanup job.

DataWalk reprocesses only the records that changed. Existing entities are updated, and new matching rules can be tested without rebuilding the system.

~44 minutes for a routine daily change set of roughly 3,100 records on a 40-million-record base, touching about 105,000 of 549 million candidate pairs.[1]

The resolved view keeps pace with the estate.

WHY NATIVE MATTERS

No Second ER Stack

A separate match engine creates a second operating model.

With a separate ER engine, data is typically copied out of the platform, mapped to another schema, matched there, and returned for use elsewhere. That creates another data store, another security boundary, often another license, and another version of the resolved entity to keep synchronized. Every new source, corrected match, and changed rule must pass through that architecture again.

DataWalk resolves people and organizations inside the platform that already holds and connects the data. Relational, search, and graph capabilities operate on the same schema, model, database, and compute layer. The resolved entity and its evidence are written directly into the investigation graph.

Search, link analysis, scoring, screening, casework, and AI all consume the same governed entity.

This changes the role of entity resolution. It is no longer a preparation step completed before the real work begins. It becomes part of the model the real work runs on. A corrected match does not have to be reconciled with a separate graph. A new record extends the entity already used by investigators. AI agents can consume the same resolved entities through the platform's APIs and MCP.

The match is not another output. It becomes part of the operating state of the platform.

The one-question test

Ask any entity resolution vendor: after the match is made, where does the resolved entity live? When the answer is another store, another model, or another file, identity still has to be moved, mapped, and reconciled before the investigation can use it.

When the data cannot leave

Resolution stays inside the customer's DataWalk environment. Nothing is exported to a separate match engine, and no second copy of the data exists in one. The same architecture applies to on-premises and air-gapped deployments.

A separate match engine delivers a result. But DataWalk creates a resolved entity the platform can act on.

EVERY MATCH CAN SHOW ITS WORK

A Confidence Score Is Not An Explanation

A score tells you how strongly the system leans. It does not tell you why.

DataWalk preserves the path behind the decision. An investigator can inspect the values that were normalized, the clue that brought the records together, the evidence that contributed to the score, and any deterministic rule that promoted, demoted, or vetoed the result.

The same applies when records do not resolve. The system can show whether the pair was compared and rejected, or whether a shared value was too common to justify comparison. For evaluated pairs, DataWalk can show why the records matched, were rejected or remained possible. It can also show when an over-common value was suppressed before pairing.

This matters because similarity is not always identity. Two companies may share an address. Two people may share a name. An otherwise convincing pair may carry conflicting exclusive identifiers.

DataWalk retains that uncertainty rather than forcing false certainty. Possible remains possible until a person reviews the evidence.

That makes the process usable by investigators, tunable by data teams, and reconstructable when an auditor asks what happened.

What a reviewer can inspect in DataWalk

- The cleaned values actually compared
- The clue and rule that created the candidate pair
- The evidence contributors and their weights
- The score before and after any override
- The rule that promoted, demoted, or rejected
- The reason an over-common value was excluded

Matches and evaluated non-matches can show their work.

THE EVIDENCE

Measured At Enterprise Scale

The DataWalk Native Entity Resolution has incredible performance, has been measured at the scale enterprise investigations actually face, with the method and limits stated plainly here.

DataWalk resolved 212.8 million records end to end in approximately 13 hours on three servers. That means even a very large data estate can be resolved in a single scheduled run, ready for investigators the next day rather than after a multi-day processing cycle.

In a separate run across 162-million organization records, DataWalk used blocking to narrow roughly 13-quadrillion possible record pairings to 3.69 billion plausible candidates. Every record remained in scope; compute was concentrated on pairs with credible evidence of a match, making resolution at this scale practical.

Once the estate is resolved, routine updates are considerably smaller. A daily change set on a 40-million-record base completed in approximately 44 minutes.

On-demand checks against 212.8 million resolved entities returned in 6.6 to 18 seconds per server-side request.

Actual results may vary by deployment. Performance and match quality depend on the data, configuration, hardware, and mapping to the ontology. Precision and recall should be validated separately against the customer's own records.

13.08 hours

to resolve 212.8million records end to end on three servers.[1]

3.69B

plausible candidate pairs generated from ~13 quadrillion possible pairings across 162-million organization records.

~44min

for a routine daily change set on a 40-million-record base [1]

Honest bounds

Resolution is scheduled and incremental, not sub-second streaming. It currently resolves records representing people and organizations. Blocking makes comparison at this scale practical, but records that share no usable matching signal may never be compared. Suppressed values and the reasons for suppressing them remain inspectable. Possible matches are not merged automatically and should be reviewed by a person. On-demand checks use the most recently completed batch and do not indicate how current it is. DataWalk processes data in many languages and scripts, but its name matching is tuned primarily to Western naming conventions. Names that follow different conventions, or have been heavily transliterated, should be validated against the customer's own data.

HOW DATAWALK RESOLVES ENTITIES

Seven Inspectable Steps From Record To Resolved Entity

The DataWalk matching process; each candidate-pair decision follows a deterministic, inspectable path. The same evidence and configuration produce the same verdict, and each decision can be traced through seven stages.

1. **Detect change and exact duplication**
   Records are fingerprinted so unchanged data is not reprocessed. Byte-identical records link as exact matches before comparison begins.

2. **Normalize**
   Names, addresses, phones, dates, and identifiers are converted into consistent comparable forms, including transliteration and checksum validation.

3. **Create features**
   Normalized values become individual or combined comparison signals, one feature value per entity per distinct value.

4. **Block**
   Deterministic clues, such as a phone number or a name-and-postcode key, select the records worth comparing. Over-common values are suppressed.

5. **Score**
   Candidate pairs accumulate positive and negative evidence across controlled evidence families, landing in a match, possible, or non-match band.

6. **Apply deterministic rules**
   Decisive evidence overrides the band: conflicting exclusive identifiers veto, strong combinations promote, weak name-only evidence holds for review.

7. **Publish results**
   The verdict, score, contributors, and applied rules are written back into the knowledge graph as ER Results Links.

Going deeper
Identifier coverage, parser internals, and the named rule set are documented in the technical evaluation paper, DataWalk Native Entity Resolution, version 1.2.[4]

The process is technical. The answer does not have to be mysterious.

BEFORE YOU BUY ENTITY RESOLUTION

Do Not Evaluate The Match In Isolation

A separate match engine may produce excellent matches and still leave the organization with another database, another integration, another security boundary, and another version of the subject to reconcile. Before comparing algorithms, ask what you will have to build, operate, and govern around them.

1. Where does the resolved entity live?

GOOD ANSWER Inside the governed model used for investigation, analytics, and AI.

RED FLAG In another database, file, or identifier store that must be synchronized with downstream systems.

2. Does the data leave your existing environment?

GOOD ANSWER Resolution runs where the governed data already resides.

RED FLAG Records must be exported to a separate match engine and copied into its store.

3. What keeps the answer current?

GOOD ANSWER Changed records are detected and resolved incrementally into the existing entity.

RED FLAG Periodic full rebuilds, remediation projects, or batch outputs that begin aging as soon as they are produced.

4. Can the system explain both the match and the non-match?

GOOD ANSWER The evidence, score, candidate-generation logic, and applicable rules are inspectable.

RED FLAG A confidence score, an opaque model output, or an unexplained absence of a match.

5. What else must you buy or build?

GOOD ANSWER Resolution, graph, search, analytics, permissions, and investigation operate in one platform.

RED FLAG The engine still requires a database, integration layer, graph, case-management surface, or separate analytics product.

6. How is performance proved?

GOOD ANSWER Documented workloads, configurations, results, and limitations that can be tested against your data.

RED FLAG "Real time," "best-in-class," or universal accuracy claims without enough method to reproduce them.

7. Can it run under your operating constraints?

GOOD ANSWER Support for your deployment model, authentication, audit and data-residency requirements.

RED FLAG A cloud-only service or separate platform that cannot operate in sovereign, on-premises, or air-gapped environments.

The same one-question test

After the records have been resolved, how much work remains before an investigator can use the answer?

THE HIDDEN STACK BEHIND A SEPARATE MATCH ENGINE

The matching engine may be only one component of what you ultimately have to operate.

- Database for the ER engine
- Data export and ingestion pipeline
- Separate identity model
- Synchronization with the graph
- Security and access controls
- Investigation or casework application
- Audit and evidence layer
- Infrastructure and specialist skills
- Separately licensed graph modules

A lower-cost matcher can become an expensive operating model.

Price the whole path from record to investigation, not only the match.

CONCLUSION

Make Resolution Part Of The Platform

Your organization does not need another place where identity is calculated.

It needs one governed subject that screening, investigation, analytics, and AI can use together.

A separate match engine produces a match that must be moved, mapped, secured, and reconciled with the systems where work happens. Native entity resolution is not another system to operate. It is a property of the platform where intelligence is gathered, evidence is explored, and decisions are made. New records extend the subject. Every decision retains its evidence. The resolved entity is already connected to the accounts, companies, addresses, alerts, and cases around it.

The village always knew because identity was not a separate project. It was part of how the village operated. DataWalk brings that property to the investigation graph.

Before you add a separate match engine

Ask what remains after the match:

- Another database?
- Another data copy?
- Another identity model?
- Another integration?
- Another license?
- Another place where the answer can drift?

DataWalk resolves the subject where the investigation data already lives.

See native entity resolution inside DataWalk

See fragmented records become governed subjects, inspect the evidence behind each decision, and follow the connections that become visible around them. Request a demonstration through your DataWalk representative or at datawalk.com.

See what becomes visible once the entity is resolved.

The Connections Are Already There.

DataWalk links fragmented records into resolved entities, then connects them to everything the knowledge graph already knows.

Native entity resolution runs inside the same governed platform used for search, link analysis, investigations, analytics, and AI. No separate match engine. No second identity model. The evidence behind each match is kept.

Resolve the subject. See the whole case.

For more information, visit datawalk.com.

DataWalk

SOURCES

1. **DataWalk engineering benchmarks.** A 212.8-million-record heterogeneous dataset resolved end to end on a three-server cluster; a separate 162-million-organization record workload producing 3.69 billion candidate pairs; and incremental processing measurements on a 40-million-record base. Results describe the measured runs. Performance varies with data, hardware, and configuration.
2. **DataWalk production banking deployment.** Screening against resolved entities reduced false alerts by 80% while retaining more than 90% of true alerts.
3. **DataWalk on-demand resolution test.** Submitted records checked against a 212.8-million-entity resolved corpus in 6.6 to 18.0 seconds server-side per request, with up to 100 records per request and no writes to the graph.
4. **DataWalk Native Entity Resolution: Native, Explainable, and Built to Scale. Technical evaluation paper, version 1.2, 28 July 2026.** Source for the architecture, deterministic pipeline, explainability, supported entity types, and stated limitations.
5. **"Anti-Money Laundering Alert Optimization Using Machine Learning with Graphs"**; Ahmad Naser Eddin, Jacopo Bono, David Aparício, David Polido, João Tiago Ascensão, Pedro Bizarro, Pedro Ribeiro (arXiv:2112.07508, Dec 2021, revised Jun 2022)

GLOSSARY

Entity resolution: deciding which records refer to the same real person or organization.

Resolved entity: the records identified as one real actor, retained with the evidence behind that decision.

Native entity resolution: resolution performed inside the data and investigation platform itself.

© 2026 DataWalk. All trademarks are the property of their respective owners. This solution guide is provided for general information and describes capabilities and measured results as of its publication date.

2026 edition