

The vendor *name problem*.

One supplier, forty-three spellings, and a finance team certain they know their **top ten**. Why **entity resolution** is the least glamorous and most valuable thing in the pipeline.

Riverton & Associates, Inc. · 2 June 2026

Executive summary

Every spend analysis rests on a question that sounds trivial: *who did we pay?* In practice the answer is scattered across dozens of spellings of the same name, in several systems, in records typed by people who had no reason to care about consistency.

Until those records are resolved to one supplier, every number downstream is wrong. Category totals are wrong. Negotiating leverage is understated. Duplicate payments hide. The top-ten supplier list — the one report every finance team trusts — is confidently, quietly incorrect.

Entity resolution is the work of turning many records into one supplier. It has no demonstration value, it produces no chart anyone wants to look at, and it is the single highest-return step in a spend pipeline.

The forty-three spellings

One supplier generates variation from a dozen independent sources, and they compound.

- **Legal form.** Inc , Inc . , Incorporated , LLC , L.L.C. , Ltd , GmbH , Pty Ltd — with the suffix, without it, and punctuated three ways.
- **Abbreviation.** MKTG and Marketing . INTL , MFG , SVCS . Whoever set up the vendor master had a forty-character field and used it.

- **Case, punctuation and damage.** Trailing whitespace, control characters that survived an export, encoding damage where an accent met the wrong codepage.
- **Legal entity versus trading name.** The invoice says one thing, the contract another, and the cheque is made out to a third.
- **Corporate structure.** Dell Marketing L.P., Dell Products L.P., EMC Corporation, VMware — four legal entities, one ultimate parent, and no string similarity between the last two and the first two.
- **Card and expense channels.** Purchasing-card lines arrive as merchant descriptors: DELL 800-999-3355 TX . Same supplier, not one shared field with the AP record.
- **Multiple systems.** Two ERPs from an acquisition, a procurement tool, an expense system — each with its own vendor master, IDs, and duplicates within it.
- **Time.** Companies are acquired, renamed and divested. A record correct in 2019 describes an entity that no longer exists.

Forty-three spellings is not an exaggeration. It is what you get when a dozen sources of variation multiply across a few hundred thousand transaction lines.

Why the top ten is wrong

The failure is not that the numbers are slightly off. It is that they are wrong in a direction nobody checks. Fragmentation always *understates* a supplier: spend that belongs to one relationship is divided among its variants, and each variant ranks lower than the whole.

AS RECORDED	SPEND	RANK
DELL MKTG L.P.	\$1.6M	7
Dell Marketing LP	\$1.1M	12
DELL PRODUCTS L.P.	\$0.8M	19
Dell Mktg L.P. (duplicate vendor ID)	\$0.4M	31
card and expense lines	\$0.2M	—
Resolved	\$4.1M	2

Illustrative. Every row is a real supplier with real invoices.

Nothing in the unresolved view suggests an error. The report is internally consistent, reconciles to the ledger, and is wrong about the thing it exists to tell you. The consequences are specific:

- **Leverage is left on the table.** You negotiate as a \$1.6M customer when you are a \$4.1M customer. The supplier's account team knows the real number. You do not.
- **Tail spend looks like tail spend.** Fragments fall below the threshold where anyone looks, so a major relationship is managed as thirty small ones.
- **Duplicate payments survive.** Detection keys on supplier plus invoice number. Two spellings are two suppliers, so the duplicate never surfaces.
- **Contract coverage is overstated.** Spend under contract looks compliant because the off-contract variants count as different suppliers.
- **Risk and concentration are invisible.** Concentration limits, sanctions screening and supplier risk scoring all assume you know the counterparty.

Why this is harder than string matching

The instinct is to reach for fuzzy matching. It is not sufficient, and it is not even the main event.

High similarity, different entities. Dell Marketing L.P. and Dell Products L.P. are almost identical strings and are genuinely different legal entities with different contracts and payment terms. Whether they should be merged depends on the question being asked — a business decision, not a distance threshold.

Zero similarity, same entity. IBM and International Business Machines Corporation. Alphabet and Google. Meta and Facebook. No string metric will connect them. Only reference data will.

Name similarity is therefore one signal among several, and the others carry more weight than people expect: tax identifiers, registered address, bank remittance details, the GL codes a supplier is normally posted to, the buying entity, what was bought, and the behavioural fingerprint of invoice cadence and amount distribution. A match supported by four weak signals is stronger than one supported by a single strong name score.

Scale. Naïve comparison is quadratic — a few hundred thousand vendor records is tens of billions of pairs. The work has to be blocked, and blocking is where most of the recall is won or lost: a pair that never enters a block can never be matched, no matter how good the scorer is.

The grain problem

There is no single correct answer to *who is this supplier*, because the right grain depends on the question.

 What leverage do we have in a negotiation?

Ultimate parent

 Who do we owe, and on what terms?

Legal entity

 Who actually delivers, and what is our risk?

Operating entity or site

 Which contract governs this line?

Contracting entity

A layer that flattens everything to the ultimate parent cannot answer the second and third questions. One that stops at the legal entity cannot answer the first. The resolution is to keep the hierarchy rather than choose a level: resolve each record to a canonical **legal entity** with a stable identifier, then attach parentage above it, so reports roll up to whichever level the question needs. Choosing a single grain early is the most common irreversible mistake in this work.

The asymmetry that should govern the design

False merges and false splits are not equally bad.

A **false split** leaves two records unmerged. The number is understated, the fragment is visible, and someone eventually notices a familiar name in the wrong place. It is a known unknown, and it is recoverable.

A **false merge** combines two genuinely different suppliers. The number is now confidently wrong, it reconciles, nothing looks unusual, and it has corrupted a figure someone is about to act on. It is very hard to detect after the fact, and it destroys trust in the whole layer when it is found.

This asymmetry should shape every threshold. Bias toward precision. Route the uncertain band to human review rather than guessing. And never let an automated merge be unexplainable: when a business owner challenges a number — and they will, usually the number that matters most — the answer has to be a specific reason, not a shrug.

How it is actually done

- **Normalise deterministically first.** Case, whitespace, control characters, encoding repair, legal-suffix handling, an abbreviation dictionary. Dull, rule-based and testable — and it removes a large share of the variation before any model sees the data. Probabilistic matching on unclear input wastes its discriminating power on noise.
- **Block.** Partition into candidate groups on cheap keys — name prefixes, phonetic keys, tax ID, postcode, domain. Use several passes with different keys, because any single key has a blind spot.
- **Score on multiple signals.** Combine them into a calibrated score, not an arbitrary weighted sum. A score that means something as a probability is what lets you set thresholds honestly.
- **Resolve to a canonical entity** with a stable identifier that survives re-runs. A report whose supplier IDs change every month cannot be compared to itself.
- **Attach the hierarchy** — parentage, ultimate parent, firmographics — so the grain question is answered by rolling up rather than re-resolving.
- **Queue the uncertain band for review.** Make it small enough to actually clear, and capture every decision.
- **Persist decisions and feed them back.** Human adjudications are the most valuable training data available. A layer that discards them re-asks the same questions every month and never improves.

How to tell whether it worked

Resist reporting *percentage matched*. It rewards over-merging, which is exactly the failure mode with the worst consequences. Better measures:

- **Spend under management** — the share of total spend attached to a resolved supplier above the review threshold. This is the number that ties to value.
- **Precision on a sampled gold set** — stratified sample, adjudicated by hand, reported honestly with the sample size.
- **Distinct supplier count before and after** — a sanity signal, not a target. A large drop is expected; an implausibly large one means over-merging.
- **Stability across runs** — the share of records whose supplier ID is unchanged month on month. Instability destroys trend reporting even when each run is accurate.
- **Review queue burn-down** — if the uncertain band never clears, the thresholds are wrong.

Set these criteria before building, not after. A resolution layer with no agreed measure of success cannot be said to have succeeded.

Conclusion

Entity resolution produces nothing anyone wants to demonstrate. There is no visualisation, no model architecture worth discussing, and no moment where it looks impressive. It is normalisation rules, blocking keys, calibrated thresholds and a review queue.

It is also the step on which everything else depends. Classification applied to fragmented suppliers classifies fragments. Anomaly detection on fragmented suppliers finds nothing, because the baseline is split across the variants. Every downstream number inherits the quality of this one.

The finance team is not wrong to trust their top ten. They are wrong about which suppliers are in it — and they have no way to know that from the report itself. Resolving the vendor name is what turns spend data into an answer somebody can defend.

Riverton & Associates implements the CXO Nexus enterprise spend analytics platform, including the resolution layer described here.

Riverton & Associates, Inc. · Austin, Texas · riverton-assoc.com