

FREE PREVIEW

Identity Resolution Playbook

for Securities & Brokerage — unify investor records
across any CDP, warehouse or MDM.

This preview covers the first three sections — the problem, the maturity model, and the method at a glance. The full 13-section toolkit is available as the complete product.

Section 01 — The problem: one investor, scattered across your systems

One relationship, many records

In private markets a single investor almost never arrives as a single record. A principal typically commits through several separate legal entities — each with its own tax ID, account and address — so the firm's systems hold several investor records with nothing to say they belong to the same person.

Take one principal, *J. Whitmore*. He is in the books five ways:

- A direct holding in his own name
- The *Whitmore Family Trust*
- *Atlas Holdings LLC* (a deal co-invest)
- The family-office foundation
- A leftover prospect record from when he was first courted

To anyone looking at the register, that is five investors. It is one relationship.

Why the records don't line up: each system knows a different "who"

This is hard not because of sloppy data entry, but because the pieces of a person's identity are split across systems that were each built to answer a *different* question. None was built to answer "who is the human behind all of this?"

System	What it stores	Which "who" it knows
Fund-administration register	Subscription registers, capital accounts, cap tables	The investing entity — LLC, trust or vehicle, keyed by entity tax ID and subscription number. Not the person.
KYC / AML onboarding	Identity verification, beneficial-ownership records	The natural person — the human behind each entity, with passport/SSN and DOB. The <i>only</i> system that records the human.
Investor / LP CRM	Relationships, contacts, commitments pipeline	The relationship as the deal team sees it — often a contact or adviser, not reconciled to the register.
Investor portal	Logins, statements, capital-call & distribution docs	The login identity — whoever accesses the account, maybe a family-office administrator.
Subscription-document store	Sub-docs, side letters	The signatory on each vehicle's paperwork.
Custody / treasury	Wire instructions, settlement	The bank / settlement identity behind the money movement.

Read down that last column and the problem is plain:

- The register knows Whitmore's **entities**.
- KYC knows the **person**.
- The CRM knows a **contact**. The portal knows a **login**.
- Each system holds one true fragment. **No system holds the whole answer.**

The identifier that ties the entities together — the human — is recorded only in KYC, and the register was never joined to it. The entity and the person live in different systems, under different keys, with nothing crossing between them. That is the problem this playbook exists to solve.

What it costs you while the person stays scattered

Leaving those records unlinked is not neutral. It costs on two sides.

Commercial — you mis-read your own book.

- Split across five records, Whitmore looks like five mid-sized investors, so he sits mid-pack on every ranked list — the re-up call sheet, the co-invest allocation, the roadshow priority list.
- Your biggest backer gets a generic touch.
- You can't see concentration: five "small" positions that are really one large exposure to one person is a risk you carry blind and can't report.

Compliance — you meet obligations in fragments.

- AML and beneficial-ownership rules require knowing, and keeping current, the natural person behind each entity — not just checking the LLC and the trust.
- While the person is unlinked, the obligation is satisfied only in pieces, and a high-risk individual can sit behind a benign-looking holding company unseen.
- You also duplicate work — a separate KYC refresh, statement run and FATCA/CRS filing for one human — each a cost and a fresh chance to file wrong.

Non-advice caveat: these are the kinds of obligations that typically apply to a private-fund adviser; confirm the specifics with your compliance / DPO function.

What is impossible today because of it

The sharpest way to feel the cost is not an abstract "worse data quality" — it is the ordinary, high-value things the firm simply *cannot do* while an investor is scattered. Each of these is a routine private-markets motion, and each is dead on arrival without a resolved investor.

Call your largest relationships first for the next fund

When a firm raises its next fund, the IR team works a **ranked call list** — biggest relationships first, principal-to-principal, because those are the calls that convert re-ups and anchor the first close in a narrow fundraising window. That list is built by *totalling each investor's commitment across the firm*.

Whitmore, split into five mid-sized vehicles, never totals high enough to reach the top — so your single largest backer receives a mass email or a junior associate's call instead of a partner's. Worse, you may spend your most senior outreach on an investor who merely *looks* big because their money sits in one large vehicle, while Whitmore — actually larger — is skipped. The cost is direct: lower re-up conversion on your most valuable relationships, and your scarcest asset during a raise, partner time, aimed at the wrong names.

Size a co-invest offer to what an investor actually has with you

Co-invest — the chance to invest directly alongside the fund in a specific deal — is the most prized and most rationed offer a firm extends, and it is sized to the strength of the relationship. Offer it against a *single vehicle* and you either give Whitmore a fraction of what his true relationship warrants, or pass him over because no one vehicle looks big enough on its own.

The consequence is a relationship you actively weaken: your whale feels under-valued, and a competitor who *can* see his whole wallet courts him with a right-sized allocation. You also leave co-invest capital unplaced — the very capital that gets a deal filled.

Spot a large-but-concentrated investor and invite them to diversify

Once you can see an investor's true total *and* how it's split across strategies, you can find someone whose overall commitment is large but sits almost entirely in one fund, and proactively invite them into an adjacent strategy — growing wallet share while reducing their concentration.

While Whitmore is scattered you can compute neither his total nor its distribution, so this entire play is invisible. The cost is a missed cross-sell into your other funds and an unseen concentration you cannot manage on either side.

The common root

Every one of these is blocked by the same thing: **the person is not assembled anywhere**. Resolving identity is not a data-hygiene nicety — it is the *precondition* for the firm's highest-value commercial motions. Until the investor is whole, the firm is flying blind on exactly the relationships it can least afford to misjudge.

Asset — Systems & Data Map

Before reading on, map your own estate. For each system you run, write down:

- **Which fragment of investor identity it holds** (entity, person, contact, login, signatory, settlement)
- **The key it is stored under**

That single page shows you, concretely, where the person is scattered — and it is the starting point for everything the rest of this playbook does.

Section 02 — Three levels of identity resolution (L1, L2, L3)

Now that the problem is clear — one investor scattered across systems — there is a ladder of how firms actually pull those records back together. It has three rungs. Naming them gives you a shared language for the rest of this playbook and, more usefully, lets you point at exactly where your firm sits today.

A level is a data footprint, not an ambition

Each level is a concrete, auditable **footprint** made of four choices:

- **Fields** — which attributes you match on
- **Sources** — which systems those attributes come from
- **Match method** — how you compare them (exact, fuzzy, model-scored)
- **Cadence** — how often resolution runs (batch, or continuous / real-time)

Change any one of those and you have moved between levels. That is what makes this a diagnostic, not a slogan.

	Fields	Sources	Match method	Cadence
L1	Strong keys only — entity tax ID, LEI, subscription number	The fund-admin register alone	Deterministic, exact	Batch
L2	Strong keys plus beneficial-owner ID, entity name, name + DOB, contact, advisor ID	Widens into KYC and the LP CRM	Deterministic and fuzzy, tiered	Batch
L3	L2's fields plus model / probabilistic scoring	Adds event & streaming feeds	Model-scored	Real-time

L1 — deterministic, strong keys only

- **What it does:** matches records only when they carry the *same* hard identifier — the same entity tax ID, the same subscription number.
- **What happens with Whitmore:** his five records each carry a *different* entity key, so L1 correctly leaves them as five. The only thing it merges is an exact collision — e.g. *Atlas Holdings LLC* accidentally entered twice with the same tax ID.
- **The verdict:** correct, cheap, and small. It does not link across keys because it was never designed to. Every firm should have L1; almost no firm is fully served by it.

L2 — add softer signals, still in batch

- **What it does:** keeps L1's strong-key matches and *adds* new fields from *new systems* — most importantly the beneficial-owner ID from KYC, plus corroborating signals like name + DOB, shared contact, and advisor ID. It matches these with a mix of exact and fuzzy logic, and it still runs in a batch window.

- **What happens with Whitmore:** KYC records the same human behind the trust, the LLC and the direct holding. L2 uses that beneficial-owner link — confirmed by name, DOB and a shared family-office contact — to pull his five records into **one** relationship.
- **The verdict:** this is the level that recovers the scattered person. It gains coverage by doing two things at once — more fields *and* more sources — while staying inside a batch job. **This playbook is about getting you from L1 to L2.**

L3 — streaming and model scoring

- **What it does:** adds continuous, event-driven resolution and model scoring, so identity is resolved the moment something happens rather than in an overnight batch.
- **What happens with Whitmore:** the instant a new subscription is signed or he logs into the portal, the system scores it against the resolved graph and links it in real time — no waiting for the batch run.
- **The verdict:** powerful, but the change from **batch to real-time is a re-architecture** — new pipelines, new infrastructure, new failure modes — not a few more rules. That is why L3 is **out of scope** here. L2 is drawn to stop deliberately before that line.

The data-strategy angle — climb the ladder, don't jump it

The three levels are a ladder, not a menu. You do not pick L3 because it sounds best; you *earn* your way up, because **each level depends on the maturity of the one below it.**

- **L2 is only as trustworthy as L1.** Fuzzy, multi-signal matching amplifies whatever it is built on. If L1's strong keys are dirty, duplicated, or your system-of-record is unreliable, L2 does not fix that — it spreads the error across more records. Strong L1 foundations (clean, deduplicated strong keys, a trusted register, working governance) are the precondition for L2, not an optional nicety.
- **L3 is only as safe as L2.** Real-time model scoring on top of rules that were never proven at precision just makes bad merges faster. You cannot go real-time until your batch rules are live, measured, and stable.
- **The strategic implication:** the goal *is* to move toward L3 over time — but the route is to make each rung solid before standing on the next. Skipping a rung builds on sand.

Define the gates before you climb

Moving up a level is a decision that should be made on evidence, not ambition. Every firm should set **specific, measurable thresholds** that must be met to warrant a move — and only move when they are hit.

- **L1 → L2 readiness, for example:** strong-key fill rate is high and audited; the L1 match rate has plateaued; reference data is clean; the unresolved population has been sized and is materially large enough to justify the work.
- **L2 → L3 readiness, for example:** L2 rules are live at target precision; the coverage lift is realised and monitoring is stable; and a genuine business need for real-time exists that batch cannot serve.
- **The discipline:** without a gate, "let's move to L2/L3" is a wish. With a gate, it is a funded, defensible decision.

This is not left as advice. **Dedicated assets later in this workbook give you the tools to measure these gates** — the baseline-and-gap calculator that sizes the L1→L2 case, the precision scorecard that proves rules are safe, and the exit criteria that confirm a level is genuinely complete.

The scope line — hold it

The single most common way an L2 programme goes wrong is scope creep into streaming.

- **L2 adds fields and sources. It does not change cadence.** Resolution stays batch.
- The moment someone asks for "real-time resolution," that is L3 — a different, bigger programme.
- Keep the goal fixed: recover the scattered person, in batch, safely.

The method, at a glance

The full playbook works through six disciplined steps — sizing the gap, choosing the fields, designing the rules, proving precision, governing every merge, and going live. Here is the shape of it. The detail of each step is in the complete toolkit.



This is where the preview ends.

You've seen the problem, the maturity model, and the shape of the method. The rest of the playbook is where the work gets done — the candidate fields, the fuzzy match rules, the held-out-key precision test, the use-code governance, and the data architecture — plus every calculator, guidance note and a full worked example.

[View the full playbook →](#)

Questions, or want to discuss access options? Get in touch at datawhistle.com/contact.