



WHITE PAPER

The AI-Native RIA

Rent the model, own the harness: an owner's manual for AI in wealth management

by Carlo Bergonzi · Version 3 · July 2026 · For RIA principals, COOs & Chief Compliance Officers

About this paper. An educational paper about how AI adoption in a fiduciary business **can be done more effectively** — the ownership question, the architecture, and the governance. It is **not** investment, legal, or compliance advice. Figures are industry estimates from the cited sources and will vary by firm. Product screenshots depict a simulated advisory firm. See Sources and Important notices at the end.

Executive summary

Most firms have adopted *some* AI, but there is still huge untapped potential. Two-thirds report only "modest" returns, blocked by fragmented systems and poor data quality. The problem is not how intelligent the models are, or where they score on benchmarks. It is the architecture and the platform the model works within — and above all the **ownership question** many firms are beginning to face.

This paper makes five claims, developed in order:

- 1 **RIAs are uniquely positioned to win with AI.** Revenue scales with assets, not hours — every reclaimed hour compounds into margin, capacity, or growth — and the industry's data already lives in a small set of known systems.
- 2 **The harness, not the model, is the differentiator.** Intelligence should be rented per task, like a commodity. The operating system around it — memory, context, integrations, permissions, audit, compliance — should be built, and owned, by the firm.
- 3 **The way in is a digital twin, built at the edge.** Nothing about the firm's current processes changes on day one; the twin normalizes data from the systems the firm already runs and lets AI work against it safely.
- 4 **The twin's core is an ontology** — a real-world model of the business (objects, relationships, governed actions) that any employee, application, or AI model can speak. It is what grounds AI in the real business and lets it act safely, not just analyze.
- 5 **Compliance must be architectural, not bolted on.** In an SEC-regulated business, post-processing is insufficient; obligations should be translated into engineering primitives beneath the AI itself.

This approach leads not just to increased efficiency, but to **an asset that compounds**. A firm that owns its ontology, its audit history, and its approval data can route ever-cheaper models at every task — and eventually adapt models on its own data — instead of forfeiting that leverage to a platform vendor.

The opportunity is concrete: industry research puts the administrative burden at **10–15 hours per advisor per week**. We conservatively model roughly a third to a half of that as reclaimable — **~230–265 hours per advisor per year**. But the hours are just the entry point. The bigger opportunity is the compounding asset the firm builds along the way.

30–50%

of advisor time on
administrative work

230–265

modeled
hours/advisor/year
reclaimable

~\$10T

in assets moving as
advisors retire

1. Why RIAs

In short: **revenue scales with assets, not hours** — so every hour handed back to an advisor compounds into capacity, better service, or growth. Industry research pegs the administrative load at 10–15 hours per advisor per week. Few business models convert reclaimed time this cleanly.

The moment: four forces, one conclusion

Consolidation is accelerating. Roughly **77% of RIA assets now sit with about 7% of firms**, and 2025 set records for M&A — on the order of **466 deals, up ~27% year-over-year**, with private equity touching the large majority of transactions (Cerulli; Echelon Partners; DeVoe & Company). Sub-\$1B firms increasingly face a scale disadvantage they must close operationally — or sell.

The talent pipeline can't keep up. The median advisor is about **51 years old**; an estimated **~40% will retire within a decade**, with roughly **\$10 trillion** in client assets in motion (McKinsey; Cerulli). There aren't enough new advisors to backfill the work — which means **the work itself has to get lighter**.

Organic growth is thin. Stripped of market appreciation, organic growth for billion-dollar firms has run around **3.9% annually** (Cerulli). With fee pressure persistent, productivity is among the most controllable growth levers a firm has.

AI has arrived — but underwhelmed. Most advisors now have an AI policy and use at least one AI feature, yet **about two-thirds report only "modest" ROI**, citing data quality and fragmented systems as the top barriers (industry surveys; Publicis Sapient). The tools work; the *integration and trust* are missing.

The structural advantages

Four things make an RIA better positioned for AI than almost any other professional-services business:

- 1 Efficiency compounds for a fiduciary.** An RIA charges a fee on assets — commonly blending to roughly **0.75%–1.0%** of AUM — while compensation runs **50–60% of revenue** (Kitces; InvestmentNews). Revenue scales with assets, not hours; paid time freed from administrative work flows almost directly to the bottom line, to capacity, or to growth.
- 2 The data already lives in a few known places.** A typical firm runs a recognizable stack: a CRM, a portfolio system fed by custodial data, a planning tool, and Microsoft 365. An RIA's operational reality can be normalized into a single coherent model — the prerequisite for AI that actually works. The number-one barrier firms cite is **solvable by design**.
- 3 Human-in-the-loop fits the fiduciary duty.** A fiduciary already reviews client-facing work before it goes out. An operating model where agents draft and humans approve isn't a safety compromise bolted on — it mirrors how good firms already work.
- 4 The bar is low.** Because most firms get little from point-solution AI, the competitive bar is not "use AI" — it's "use AI that is *integrated, measured, and defensible*." A firm that can show an examiner a complete audit trail, and show clients faster service, stands apart.

The prize, quantified

- **30–50% of advisor time** goes to administrative, non-client work (Fidelity; Kitces).
- Advisors report on the order of **11 hours/week** on manual tasks like note-taking and meeting prep (industry surveys).
- We conservatively model **roughly a third to a half of the 10–15 hour weekly administrative burden as reclaimable: ~230–265 hours per advisor per**

year.

- **Onboarding** illustrates the prize: studies cite roughly **8 hours of manual labor per new client versus under 2 hours when automated**, with faster onboarding linked to materially higher retention.

Reclaimed hours are where every AI conversation starts, but they should not be where it ends. The rest of this paper is about the bigger idea underneath them: who owns the system doing the work.

2. AI sovereignty: rent the model, own the harness

The first idea to internalize is the distinction between a **model** and a **harness**. Without a harness, an AI model is a set of generalized weights — **a brain in a jar**. The smartest intelligence in the world cannot accurately report on, analyze, or act on your business without the context, tools, and data that are specific to your business.

This is where the harness comes in. It gives the brain everything else it needs to operate inside an enterprise:

- **Memory** — short-term state and long-term canonical knowledge
- **Enterprise system context** — what your firm runs and how it fits together
- **User context** — who is asking, what their role permits
- **Relevant documents and data** — retrieved and packed into context
- **Software integration and tools** — reading from and acting on real systems
- **A runtime environment** — plan, act, observe, continue
- **Permissions and governance guardrails** — scoped to role and task
- **Credential and key management** — delegated, rotated, contained
- **Audit trails** — tamper-evident evidence of what happened
- **Human-in-the-loop approvals** — people where judgment belongs
- and, most importantly, **industry-specific regulatory compliance**

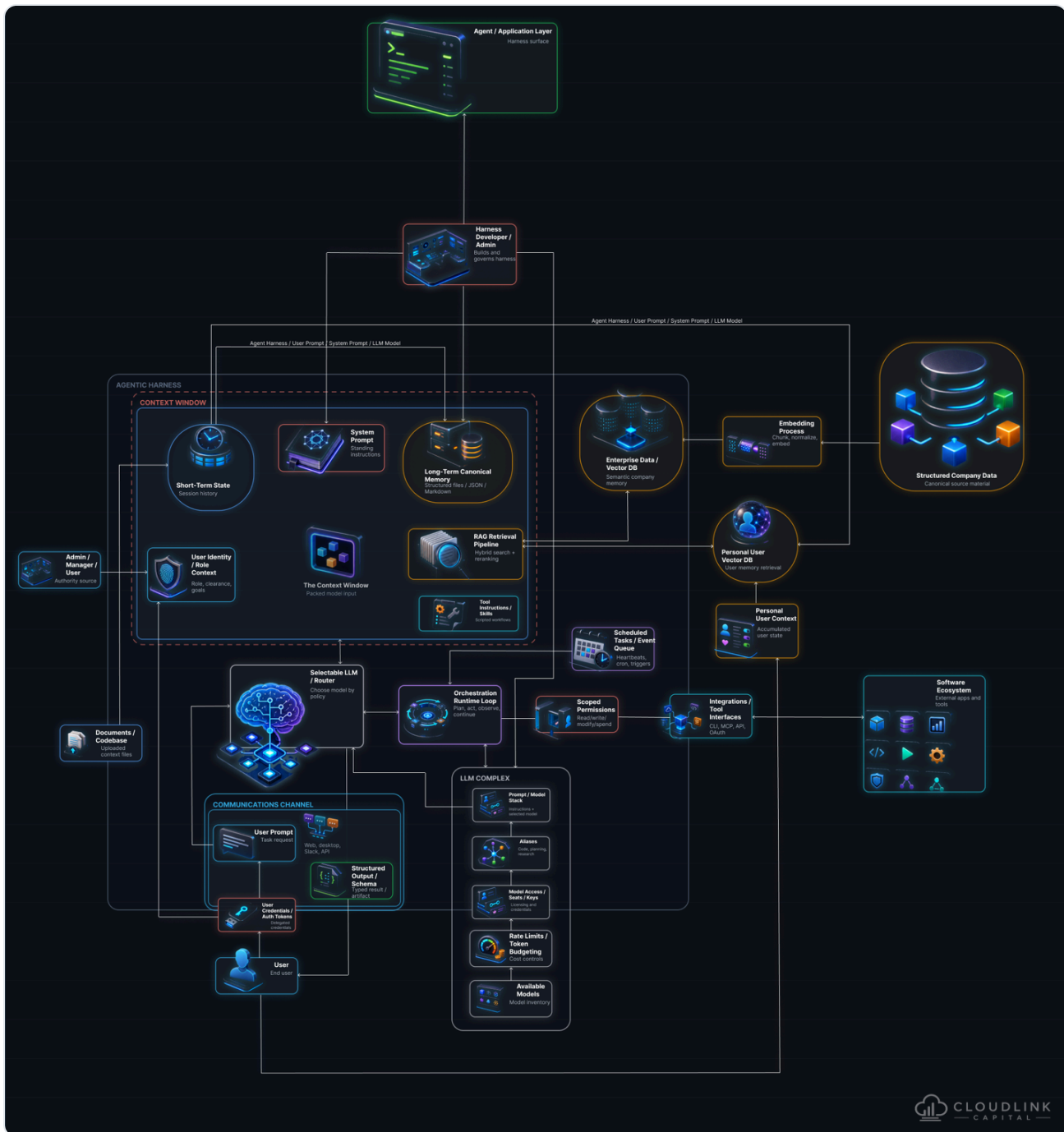


Figure 1 — The anatomy of an AI harness: everything the model needs that the model doesn't come with. The model is one selectable component; the rest is the operating system. *Source: CloudLink Capital.*

This is why a firm cannot simply download Claude Code or Codex, connect HubSpot and Microsoft 365, and expect to have an "enterprise solution." Those are superb *generalized* tools. The enterprise requires a specific operating system that is **shaped like the business**.

Any LLM can be plugged into a harness and swapped back out without disruption — the only question is how much reasoning power a given task requires. The harness cannot be swapped out. It is woven into the fabric of the business: its data, its permissions, its approval flows, its evidence. This creates a situation where the harness — not the model — is the bigger differentiating lever.

Who owns your operating system?

This brings us to build-vs-buy, and the ownership question underneath it: **who owns your operating system and your data?** As firms accumulate real experience with AI, the optimal strategy is getting clearer:

- **Build** (or have forward-deployed engineers build) the harness — the operating system over your data and systems — so the firm has 100% control over an asset optimized to its own business.
- **Buy** the LLM usage — per task, via API, like the commodity it is becoming — matched and tailored to each job being performed.

This gets the firm maximum efficacy, data sovereignty, and industry compliance — and critically, **maximum flexibility in model routing**. Many repeatable enterprise tasks do not require frontier intelligence. Armed with optimized context, smaller or open-weight models handle them day in and day out for a fraction of the cost — think orders of magnitude less per unit of work.

This is why forward-deployed engineers exist, and why firms like Palantir deploy them: not to hand out AI subscriptions to some employees and call it enterprise adoption, but to **convert the business into an AI-native platform**.

3. Becoming AI-native: a digital twin at the edge

Having worked with large, medium, and small businesses, there is an **enterprise inertia** that must be overcome for any platform change. This is the advantage of entrepreneurs and small business owners: they can build a solution before the enterprise can even have a meeting on it.

The only way an established firm keeps up is with an approach that draws from Clayton Christensen's *Innovator's Dilemma*: innovate **at the edge**, where the core business doesn't need to change. Concretely: **a digital twin of the firm and its data layer, built by a dedicated team, beside the business — not inside it.**

This works because it disrupts nothing. Emails still arrive in Microsoft 365. Contacts are still created in the CRM. But the data from all of these systems — AdvicePay, Calendly, DocuSign, eMoney, custodial file feeds from Fidelity and Schwab, Jump, Microsoft 365, Nitrogen, Orion, Smarsh, Wealthbox, Holistiplan, and the rest of the stack — is **normalized and synced into an ontology layer** that lets the entire firm's context and capability be leveraged by AI systems: first reading, then, with governance, reading and writing.

The firm keeps operating exactly as before. The twin becomes the surface on which AI works. And when the twin has earned trust — measured, workflow by workflow — the firm has become AI-native without a single big-bang migration.

4. The ontology: a digital twin that speaks your business

The core of the twin is an **ontology** — the operational semantic layer of the organization. (Palantir proves this pattern at industrial scale, and the concept maps remarkably well onto wealth management.)

Instead of forcing employees, AI models, or applications to interact with messy, siloed database tables and exports (`tbl_cust_v2_final`), the ontology translates raw data into a **real-world, human-readable model of the business.**

The building blocks

- **Object types** — the real-world entities, assets, and events. *Generic examples: an Employee, a Flight, a Manufacturing Plant, a Support Ticket.*
- **Properties** — the attributes that define an object. *A Flight has a number, a departure time, a status.*

- **Link types** — the relationships connecting objects. *An Employee is assigned to a Flight; a Part belongs to a Machine.*
- **Action types** — the rules, business logic, and **governed operations** that let users or AI change objects. *"Cancel Flight," "Reassign Crew," "Approve Purchase Order"* — and when an action runs, it securely updates the underlying source systems.

Translate that into the RIA space and the mapping is direct — the focus shifts from supply chains and machinery to households, custodial workflows, and market risk:

Component	RIA example	Properties / signature
Object type	Household	Total AUM, composite risk score, service tier
Object type	Custodial account	Account number, tax status (Roth, taxable), cash balance
Object type	Security	Ticker, asset class, 30-day yield
Link type	Manages / holds	Advisor <i>manages</i> Household, which <i>owns</i> Accounts, which <i>hold</i> Securities
Action type	Governed verbs	RebalanceToTarget(), FlagForCompliance(), ExecuteTaxLossHarvest() — each carrying its own governance

Here is what an RIA ontology looks like running against a live (simulated) firm — every object with its properties, its authority source, and its PII/NPI classification declared up front:

OBJECT TYPES — WITH LIVE INSTANCES FROM THIS FIRM ⓘ				
OBJECT	INSTANCES	AUTHORITY	WHAT IT IS	PII / NPI PROPERTIES
household > 8 props	75	crm	The client relationship unit	
client > 13 props	144	crm	An individual person or legal entity inside a household. Carries the personally identifiable attributes (PII/NPI)	address, date_of_birth, display_name, email, first_name, last_name, phone, tax_id
relationship > 6 props	144	crm	A role-carrying membership edge between a client and a household (spouse, trustee, POA, beneficiary)	
advisor > 7 props	14	m365	A supervised person serving households. Advisor identity anchors ABAC scope, HITL assignment, and author-of-record on approvals.	
team > 5 props	3	crm	An advisor grouping (team/office) used for org rollups and coverage.	
custodian > 4 props	2	reference	The institution holding client assets (Schwab, Fidelity). Reference data; custodial FILES are a separate asserted-identity surface (custodian_position).	
account > 13 props	222	portfolio	A custodial/investment account owned by a client within a household. Registration type drives tax treatment (RMD, wash-sale relevance); balance is the current market value.	account_number
security > 9 props	110	portfolio	A tradable instrument (reference/market data)	
holding > 9 props	2,138	portfolio	A security position in an account	
tax_lot > 9 props	5,012	portfolio	Purchase-lot detail under a holding	
transaction > 11 props	2,607	portfolio	A money or security movement posted to an account (trade, dividend, contribution, distribution, fee, RMD). The activity record billing, RMD tracking, and reconciliation read.	
model_portfolio > 6 props	5	twin	A firm-configured target allocation ("model"). Accounts target a model; drift is measured against its weights and tolerance.	
financial_plan > 8 props	66	planning	A planning engagement for a household (the eMoney/MoneyGuide artifact). last_reviewed_at is the review-recency signal compliance screens read.	
goal > 9 props	115	planning	A goal within a financial plan (retirement, education, purchase) with target amount and funding progress.	
fee_arrangement > 8 props	75	portfolio	The billing schedule for a household: marginal AUM tiers in bps, billing frequency, household aggregation. What the billing run computes from.	
opportunity > 8 props	13	crm	A pipeline item	
task > 10 props	44	crm	A work item, linkable to a household (null = firm-level). Workflows propose tasks; humans own them.	
document > 8 props	381	m365	A file (statement, plan, agreement, ID, form). The record is metadata; content lives in the object store behind object_ref. Client uploads are externally-authored content.	object_ref
communication > 11 props	997	m365	An email, call, meeting, note, or portal message. The record is the envelope; the body lives behind body_ref in the object store. Inbound bodies are client-authored	body_ref
consent > 7 props	144	crm	A client-granted consent or preference (e-delivery, marketing, data sharing, recording) with capture timestamp	
custodian_position > 11 props	4,276	custodian	The custodian's OWN position snapshot from a T4 file feed	account_number
custodian_transaction > 14 props	404	custodian	The custodian's OWN transaction/activity feed from T4 TRN files	account_number
esign_envelope > 13 props	3	esign	An e-signature envelope	events_ref, recipients
risk_profile > 10 props	60	planning	A risk-tolerance assessment (Nitrogen-style)	questionnaire_ref, score
invoice > 10 props	75	portfolio	A billing-system invoice (AdvicePay-style)	
payment > 8 props	75	portfolio	A payment received against an invoice	
transfer_request > 13 props	4	custodian	An ACAT/asset transfer as the CUSTODIAN reports it (T4 XFR files)	account_number
beneficiary_designation > 9 props	159	crm	A beneficiary designation on an account	designee_name



Figure 2 — Object types from a live firm twin: 75 households, 222 accounts, 2,607 transactions — each type carries its properties, its authoritative source system, and its PII/NPI-sensitive fields. Simulated firm data. *Source: CloudLink Capital.*

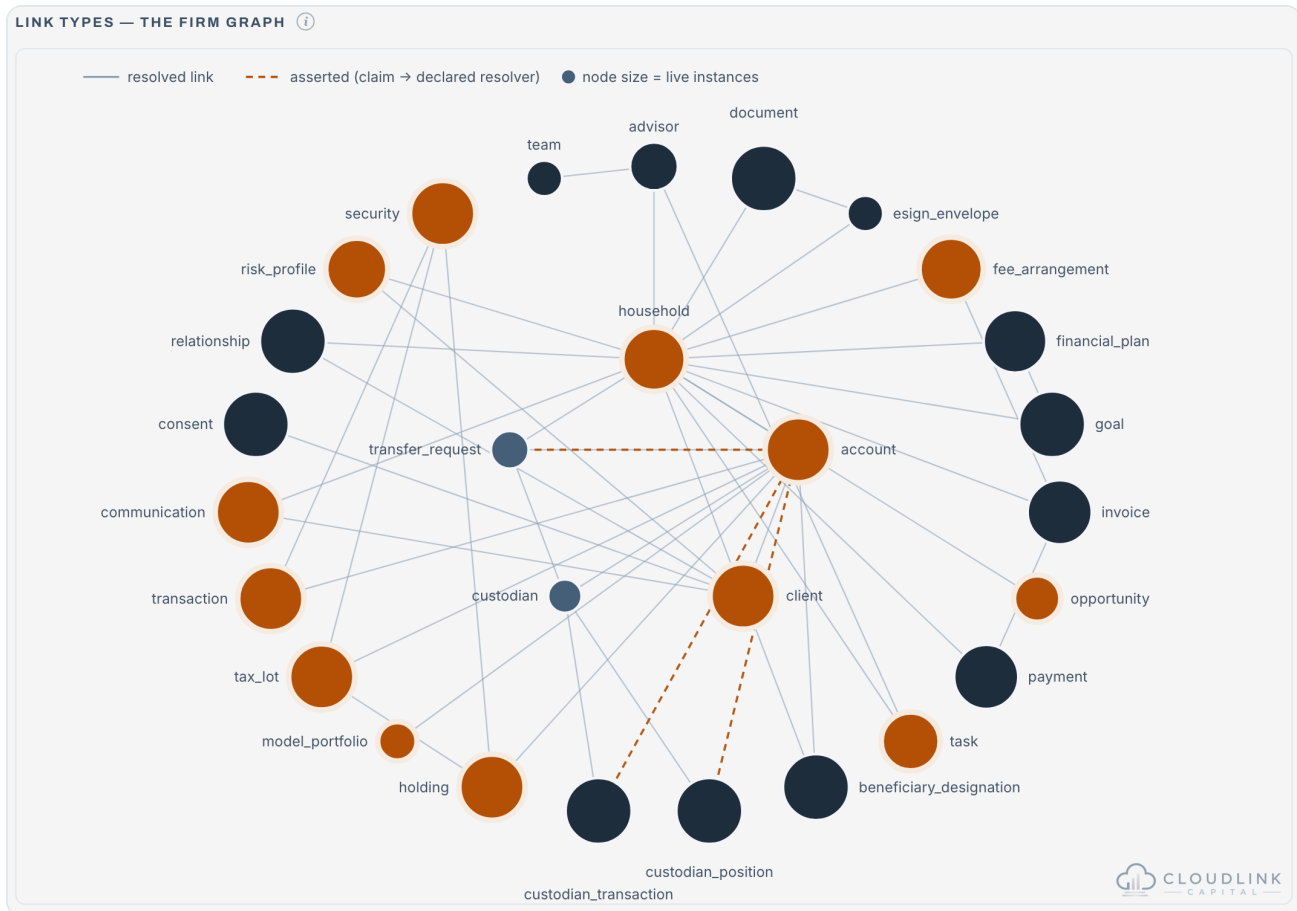


Figure 3 — Link types: the firm as a graph. Solid edges are resolved links; dashed edges are asserted claims (like a custodian file claiming an account) that must pass through a declared identity resolver before the twin treats them as fact. Simulated firm data. *Source: CloudLink Capital.*

What the ontology changes

It provides a common language. Whether you are a data scientist, a client service associate, or the CEO, everyone looks at the same concepts. You no longer have four departments defining a "household" or an "active account" four conflicting ways across four dashboards.

It's built for action, not just analysis. Most data platforms are read-only: you see a problem on a dashboard, then log into a different legacy system to fix it. A good ontology is **kinetic**. When a user or a workflow triggers an action type, the platform safely writes the decision back to the source system — updating the digital twin and the real world together.

It serves as the grounding layer for AI. With LLMs, firms struggle with models hallucinating or misreading database schemas. The ontology is the guardrail and

the context: when an AI agent is asked to work, it doesn't write blind SQL — it reads the defined objects, links, and actions, and it can only take **legally and logically permitted business operations**.

Traditional data stack	Ontology
Manages data (rows, columns, tables)	Manages decisions (entities, relationships, actions)
Hard for non-technical users to query	Real-world language anyone can speak
Custom code for every new dashboard or app	Apps and AI models plug into the pre-mapped model

And because actions are first-class objects, the entire catalog of what AI *may do* is inspectable — with governance in the signature, not in a wiki:

ACTION	KIND	OPERATES ON	GOVERNANCE
accounts_off_model	compute	account	read-only · citable
beneficiary_review	compute	beneficiary_designation, account, household	read-only · citable
billing_run	compute	fee_arrangement, account, transaction	read-only · citable
capture_coverage	compute	communication, document	read-only · citable
carryover_tasks	compute	task	read-only · citable
communications_missing_household	compute	communication	read-only · citable
compliance_calendar	compute	workflow context	read-only · citable
concentration_screen	compute	household, account, holding, security, financial_plan, communication	read-only · citable
drift_vs_model	compute	account, holding, security, model_portfolio	read-only · citable
envelope_status	compute	esign_envelope	read-only · citable
estate_event_screen	compute	client, communication, document	read-only · citable
exception_packet	compute	account, holding, security, custodian_position, custodian, transaction	read-only · citable
households_without_advisor	compute	household	read-only · citable
hygiene_sweep	compute	client, household, account, communication	read-only · citable
instance_health	compute	account, beneficiary_designation, communication, custodian_position, document, household, invoice	read-only · citable
intake_documents	compute	document, client, account · taints document	read-only · citable
invoice_reconciliation	compute	invoice, payment, fee_arrangement, account	read-only · citable
label_household	compute	household	read-only · citable
onboarding_package	compute	household, client, opportunity, account, custodian	read-only · citable
orphan_documents	compute	document	read-only · citable
overdue_invoices	compute	invoice	read-only · citable
plan_progress	compute	financial_plan, goal	read-only · citable
portfolio_rollup	compute	account, holding, security	read-only · citable
proposal_packet	compute	opportunity, household, client, risk_profile, model_portfolio, fee_arrangement	read-only · citable
read_body	compute	communication · taints communication	read-only · citable
rebalance_proposal	compute	account, holding, security, model_portfolio, tax_lot, transaction	read-only · citable
recent_activity	compute	transaction	read-only · citable
reconcile_activity	compute	account, transaction, custodian_transaction, custodian	read-only · citable
reconcile_positions	compute	account, holding, security, custodian_position, custodian	read-only · citable
retirement_accounts_missing_beneficiary	compute	account, beneficiary_designation	read-only · citable
rmc_status	compute	client, account, transaction	read-only · citable
schedule_orchestration	compute	communication · taints communication	read-only · citable
scope_to_accounts	compute	account, holding, transaction, tax_lot	read-only · citable
screen_communications	compute	communication · taints communication	read-only · citable
screen_injection	compute	communication · taints communication	read-only · citable
security_exposure	compute	holding, security	read-only · citable
service_form	compute	task, client, account, custodian, household	read-only · citable
stale_sent_envelopes	compute	esign_envelope	read-only · citable
suitability_check	compute	risk_profile, account, model_portfolio, household	read-only · citable
tlh_scan	compute	tax_lot, holding, security, transaction, account	read-only · citable
transfer_status_chase	compute	transfer_request, account, household	read-only · citable
triage_inbox	compute	communication · taints communication	read-only · citable
upcoming_meeting	compute	communication	read-only · citable

Figure 4 — The action catalog, read-only tier: every compute action declares what it operates on and carries its governance ("read-only · citable") in the signature. Simulated firm data. *Source: CloudLink Capital.*

acat-status-chase	composite	account, household, transfer_request	analysis	L2	HITL: ops
activity-reconciliation	composite	account, custodian, custodian_transaction, transaction	record	L2 (hard floor)	HITL: ops
beneficiary-annual-review	composite	account, beneficiary_designation, household	record	L2 (hard floor)	HITL: advisor
billing-run	composite	account, fee_arrangement, transaction	record	L2 (hard floor)	HITL: ops
capture-coverage-attestation	composite	communication, document	record	L2 (hard floor)	HITL: cco
client-communications	composite	client, communication, household, task · taints communication	client-comms	L2 (hard floor)	HITL: advisor
client-onboarding	composite	account, client, custodian, household, opportunity	record	L2 (hard floor)	HITL: ops
compliance-calendar-adv	composite	communication	record	L2 (hard floor)	HITL: cco
compliance-surveillance	composite	communication · taints communication	record	L2 (hard floor)	HITL: cco
crm-followup-task	composite	household, task · writes task	record	L2 (hard floor)	HITL: advisor
crm-review-note	composite	account, communication, holding, household, security · writes communication	record	L2 (hard floor)	HITL: advisor
custodial-reconciliation	composite	account, custodian, custodian_position, holding, security	record	L2 (hard floor)	HITL: ops
custodian-exception-resolution	composite	account, custodian, custodian_position, holding, security, transaction	record	L2 (hard floor)	HITL: ops
data-hygiene-sweep	composite	account, client, communication, household	record	L2 (hard floor)	HITL: ops
document-intake	composite	account, client, document, household · taints document	record	L2 (hard floor)	HITL: ops
esign-envelope-chase	composite	esign_envelope	analysis	L2	HITL: ops
esign-envelope-send	composite	esign_envelope, household · writes esign_envelope	record	L2 (hard floor)	HITL: advisor
estate-event-response	composite	client, communication, document	record	L2 (hard floor)	HITL: advisor
inbox-triage	composite	communication · taints communication	analysis	L2	HITL: ops
invoice-payment-reconciliation	composite	account, fee_arrangement, invoice, payment	record	L2 (hard floor)	HITL: ops
notetaker-meeting-summary	composite	communication, household · taints communication	analysis	L2	HITL: advisor
portfolio-rebalancing	composite	account, holding, household, model_portfolio, security, tax_lot, transaction	advice	L2 (hard floor)	HITL: advisor
proposal-prospect-packet	composite	client, fee_arrangement, household, model_portfolio, opportunity, risk_profile	client-comms	L2 (hard floor)	HITL: advisor
review-meeting-prep	composite	account, client, communication, financial_plan, goal, holding, household, model_portfolio, security, task, transaction	analysis	L2	HITL: advisor
risk-suitability-review	composite	account, household, model_portfolio, risk_profile	analysis	L2	HITL: advisor
rmd-distributions	composite	account, client, transaction	record	L2 (hard floor)	HITL: ops
scheduling-orchestration	composite	communication · taints communication	analysis	L2	HITL: ops
service-paperwork	composite	account, client, custodian, household, task	record	L2 (hard floor)	HITL: ops
tlh-season-campaign	composite	account, holding, security, tax_lot, transaction	advice	L2 (hard floor)	HITL: advisor

Figure 5 — Composite workflows: each one declares its class (record, analysis, advice, client-comms), its autonomy ceiling — "L2 (hard floor)" means full autonomy is structurally unavailable — and the named human role that approves it. Simulated firm data. *Source: CloudLink Capital.*

The patterns that matter to an RIA

Kinetic onboarding (bi-directional action). In traditional onboarding, an advisor collects data in the CRM, a client service associate re-types it into a custodial portal, then generates an e-signature envelope. With an active ontology this becomes a single action type: an associate triggers "Open custodial accounts" on the Household object; the system pulls the client details, pushes them via API to the custodian to generate the shell accounts, and dispatches the e-signature packet. **No double data entry.**

Git-style branching (planning scenarios). Instead of simulating market crashes for trading, you simulate life events for planning. An advisor **branches** a client's digital twin to model selling the business next year, or retiring early; the ontology projects tax impact and cash flow over twenty years. If the client adopts the plan, the branch **merges** and becomes the new baseline.

Dynamic network traversal (compliance surveillance). When a new SEC rule drops or a private fund changes its fee structure, compliance officers usually scramble through spreadsheets to figure out who is affected. With an ontology, graph traversal answers it instantly: *Fund* → *Holdings* → *Custodial Accounts* → *Households* — a real-time, filtered list of every affected household, ready for a bulk, governed response.

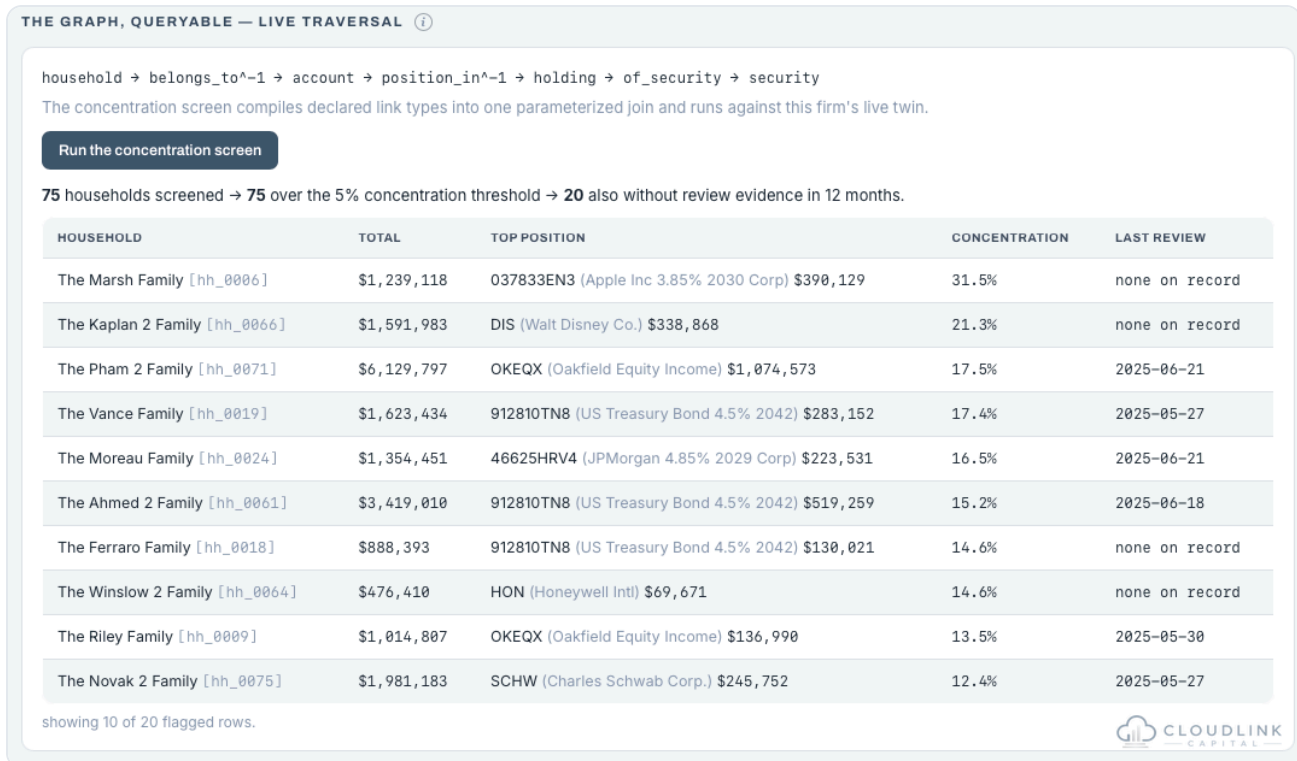


Figure 6 — The graph, queryable: a concentration screen compiled from declared link types (household → account → holding → security) running live — 75 households screened, 20 flagged over threshold with no review evidence in 12 months. Simulated firm data. *Source: CloudLink Capital.*

Security baked into the object (information barriers). If the Client object carries Social Security numbers or net worth, the ontology locks those properties by role. An intern building a marketing campaign sees anonymized tiers; the lead advisor and the CCO see the full picture. The protection travels with the data — dashboard, query, or AI tool alike.

The action layer (money movement). If a client requests a \$50,000 wire for a house purchase, the workflow runs as a structured action, not a manual database update: **verify identity** (is verbal authorization or a dual-factor log linked to the household?) → **check liquidity** (is the cash there?) → **stage the request** (generate a service ticket routed toward the custodian) → **require human sign-off** (a named principal reviews and approves). Note what the action does *not* do: it does not move money. It stages, evidences, and routes to a human. Section 5 goes further — under the Custody Rule, an early deployment shouldn't just gate money movement; it should exclude it outright.

create_note	write	workflow context · writes communication	wealthbox · T1 · reversible	active
create_task	write	workflow context · writes task	wealthbox · T1 · reversible	active
send_envelope	write	workflow context · writes esign_envelope	docusign · T2 · compensable	active



Figure 7 — The write tier: every write action is vendor-scoped and classified before it can run — reversible (a CRM task) or compensable (an e-signature envelope that can be voided). Simulated firm data. *Source: CloudLink Capital.*

The agent harness on top

Put the harness of Section 2 over this ontology and the AI assistant becomes an operational co-pilot — strictly governed by the model of the firm:

- **As a context map (meeting prep).** An advisor asks, "Prep me for my meeting with the Smith family." The AI traverses the ontology: household data, the last three meeting notes, open service tickets (an unresolved address change), and a flag that the IPS hasn't been re-signed in two years. A clean brief, in seconds, with citations back to the objects.
- **As a tool belt (post-meeting execution).** "Update their risk tolerance to Moderate and start a task to consolidate the external 401(k)." The AI updates the CRM properties through governed actions and drafts the rollover paperwork for the service team to review. It executes *administrative* work. It does not place trades.

5. Governance built into the fabric

Why general-purpose AI fails in wealth management

In highly regulated spaces, AI adoption cannot outpace the firm's ability to govern it. And when regulated firms evaluate AI platforms, they quickly discover a fundamental flaw in most of the market: **compliance is treated as a module bolted onto the architecture.** General-purpose orchestration frameworks (LangChain is the best-known) are designed to maximize developer speed and agent autonomy; they approach compliance through *post-processing* — trying to catch errors, redact information, or log actions **after** the AI has already made

decisions or touched sensitive data. Developer trace logs are built for debugging, not for evidence.

In an SEC-regulated environment governed by Reg S-P, the Marketing Rule, and Books-and-Records requirements, post-processing is insufficient. Regulators do not want a system that *tries* to be compliant. They expect a system that is **structurally incapable of the violation**.

For AI to be safely deployed at the core of an RIA, compliance must be **architectural**: take the most stringent SEC and FINRA obligations and translate them directly into low-level engineering primitives — below the AI agents themselves, into the ontology. What follows is the strategic thinking behind the core compliance design decisions I'm making at CloudLink, and equally, the questions any firm should put to any platform it evaluates.

SEC Rule 204-2 (Books and Records): the immutable audit primitive

The challenge. Examiners require an accessible, tamper-evident record of advice, communications, and advertisements spanning years. Standard application logs and AI trace logs are mutable and carry no cryptographic proof.

The design decision. An **immutable audit log as a core primitive** — not application logging. Every consequential event — an agent proposing a portfolio adjustment, a human approving a client email, a change to a governance policy — is written to an append-only, hash-chained ledger under cryptographic signatures. Tamper with a record and the chain breaks. When an examiner requests the books, the firm produces a mathematically verifiable history of what the AI did and who approved it. And because the platform operates on a digital twin, the firm keeps a **hybrid system-of-record posture**: the twin derives its intelligence from the existing compliant CRM and portfolio systems, while the authoritative records stay safely where they are.

OBLIGATION LENS		
engineering map, not legal opinion - mirrors counsel-tracked obligation-to-control rows for human review; confirm scope per firm.		
advice-line custody-avoidance la-204-2 marketing-rule-206-4-1 reg-sp-2024-ir reg-sp-safeguards		
ROLE	TARGET	NOTE
control	action:intake_documents	Document intake applies retention classes and surfaces unfiled records.
control	action:screen_communications	Communication screening supports 204-2 surveillance and retention evidence.
evidence	action:compliance_calendar	The compliance calendar points reviewers to recordkeeping evidence.
record	action:activity-reconciliation	Record-class workflow contributes to the books-and-records evidence surface.
record	action:beneficiary-annual-review	Record-class workflow contributes to the books-and-records evidence surface.
record	action:billing-run	Record-class workflow contributes to the books-and-records evidence surface.
record	action:capture-coverage-attestation	Record-class workflow contributes to the books-and-records evidence surface.
record	action:client-onboarding	Record-class workflow contributes to the books-and-records evidence surface.
record	action:compliance-calendar-adv	Record-class workflow contributes to the books-and-records evidence surface.
record	action:compliance-surveillance	Record-class workflow contributes to the books-and-records evidence surface.
record	action:crm-followup-task	Record-class workflow contributes to the books-and-records evidence surface.
record	action:crm-review-note	Record-class workflow contributes to the books-and-records evidence surface.
record	action:custodial-reconciliation	Record-class workflow contributes to the books-and-records evidence surface.
record	action:custodian-exception-resolution	Record-class workflow contributes to the books-and-records evidence surface.
record	action:data-hygiene-sweep	Record-class workflow contributes to the books-and-records evidence surface.
record	action:document-intake	Record-class workflow contributes to the books-and-records evidence surface.
record	action:esign-envelope-send	Record-class workflow contributes to the books-and-records evidence surface.
record	action:estate-event-response	Record-class workflow contributes to the books-and-records evidence surface.

Figure 8 — The obligation lens: each SEC obligation (204-2 shown) mapped to the specific actions that control it, evidence it, or contribute records to it. An engineering map for human review, not a legal opinion. Simulated firm data. *Source: CloudLink Capital.*

Reg S-P & GLBA: structural isolation and the non-escape gateway

The challenge. Safeguarding non-public personal information (NPI) is paramount, and the 2024 amendments to Reg S-P added tighter requirements: incident-response programs, service-provider oversight, and customer notification within 30 days. General-purpose AI stacks are notorious for data leakage and lack native boundaries.

The design decision. Address Reg S-P at three structural levels:

- **Isolation by architecture.** Reject pooled multi-tenant databases. Every firm operates in a structurally isolated environment — one database per firm — eliminating cross-tenant NPI bleed at the database level, not the policy level.
- **The LLM gateway.** AI agents never talk directly to language models. Every outbound inference request passes through a gateway that evaluates classification tags applied at ingestion and **strips NPI before the payload ever reaches an external model.**
- **Complete data lineage.** To meet a 30-day notification duty, a firm must know *what data, whose, and where* — instantly. A full lineage graph (origin → transformations → derivatives) turns breach-scope analysis from weeks into hours.

Rule 206(4)-1 (Marketing Rule) & fiduciary duty: human-bound execution

The challenge. The RIA is the fiduciary — not the software vendor. Machine output must never become advice rendered to a client without a human fiduciary in the loop, and client-facing communications must be fair, balanced, and substantiated.

The design decision. Draw a hard line in policy: **autonomous mode is structurally forbidden for anything classified as advice or client communications.** Agents generate proposals with their rationale. To authorize an action — sending an email, writing a CRM note — a human fiduciary reviews it, and approval isn't a mere button click: it mints a **single-use, time-restricted cryptographic grant bound to the exact hash of the payload.** The execution engine refuses to run without that proof. The AI does the heavy lifting; the human remains the author of record.

The Custody Rule: structural read-orientation

The challenge. Software that holds credentials capable of moving money can inadvertently pull the firm — and its vendor — into custody status, with the surprise-exam consequences that follow.

The design decision. Read-oriented by default, by architecture. No credentials capable of moving money. No payment initiation. No standing letters of authorization. Write-backs are limited to non-asset-moving operations — CRM notes, tasks, document envelopes — and every one is gated by the human-approval primitive. This is why the money-movement pattern in Section 4 stages and routes rather than executes: the right platform posture is to make the dangerous thing structurally unavailable, not merely discouraged.

A defensible posture

Translate regulatory obligations into engineering primitives and compliance stops being a bottleneck; it becomes a structural property of the platform. As the firm scales its use of AI, its regulatory posture scales with it — **defensible, examiner-ready, and evidenced** — because the controls an examiner tests are the same primitives the system runs on.

Nothing in this section is legal advice; see Important notices.

6. An asset that compounds

Owning an auditable, ontology-shaped data layer does more than make today's work efficient. It sets up **continuous improvement** — and eventually, model training — on terms the firm controls.

Use a standardized harness off the shelf from the big labs and you inherit two constraints: you are limited to the standard integrations, and your model selection narrows to a pre-determined menu. Meanwhile, open-weight models keep accelerating — as of this writing they are approaching parity with current-generation frontier models (think Opus 4.7/4.8, GPT 5.5) for a meaningful class of well-scoped tasks, at a fraction of the cost. The AI labs' revenue exploded in the first half of 2026 precisely because agents crossed the threshold of being genuinely useful for enterprise work; open weights are not far behind the frontier that crossed it.

"The future is composable models."

— Gavin Baker, All-In Podcast, June 2026

Our position agrees. The correct enterprise approach is **compositional model selection**: routing to the appropriately priced model **per task** — in ontology terms, per *action*. This is drastically more token-efficient, and it is not fully possible inside the labs' off-the-shelf platforms. Add the ontology's optimized context and the output quality of smaller models matches or approaches the frontier for those well-scoped tasks — at order-of-magnitude lower cost.

There is also one aspect of the frontier-model arrangement that runs exactly opposite to the enterprise's interest. Closed-weight labs improve their domain-specific products by accumulating data — including enterprise data — as they build verticals like legal, finance, and design assistants. Firms are waking up to the fact that **they do not have to participate in that arrangement**. A firm that owns a data layer mapping all of its context, actions, and history can adapt open-weight models on its own N=1 data — internally or through a service partner — and **own the resulting weights**. The result: a model whose tokens cost an order of magnitude less than frontier API calls, with the firm's business intelligence baked directly into the weights, that — combined with the ontology's context — performs at or near the frontier for the firm's own applications. And the freedom to plug frontier intelligence into the platform via API, inside the compliance envelope, never goes away for the tasks that genuinely demand it.

"The real opportunity is not in picking the best model but instead in building a learning loop on top of models where human capital and token capital compound."

— Satya Nadella, CEO of Microsoft, June 2026

That is precisely the architecture described in this paper. Human experts use AI to do their work; the traces, corrections, and approvals from that work feed back into the firm's own systems; the "token capital" gets smarter; the smarter it gets,

the more powerful the humans become. Every approval queue doubles as a labeling operation the firm owns. **The result is a continuous flywheel of institutional knowledge that the firm owns — not one it rents from a lab.**

7. Where to start

The foundation of this whole thesis is the **AI-native digital twin** of your existing data, software, and context. It is the ultimate lever for everything AI will do in your firm — and the best first step is unglamorous: **assess the current software and data landscape, and evaluate the API connections required to mirror it.**

Which systems hold the truth? Which expose it cleanly? Which will need partner-program admissions with real lead times? That inventory *is* the readiness assessment.

From there, sequence by trust earned, not by ambition:

1. **Quick wins** — client communications and meeting notes, review-meeting prep, paperwork. Largest reclaimed hours, contained scope, fast proof.
2. **Operational backbone** — onboarding and account opening, billing, RMDs. Deeper integration, higher stakes.
3. **Continuous and high-trust** — compliance surveillance, portfolio rebalancing. Attempted after the platform has earned it, workflow by workflow, on measured evidence.

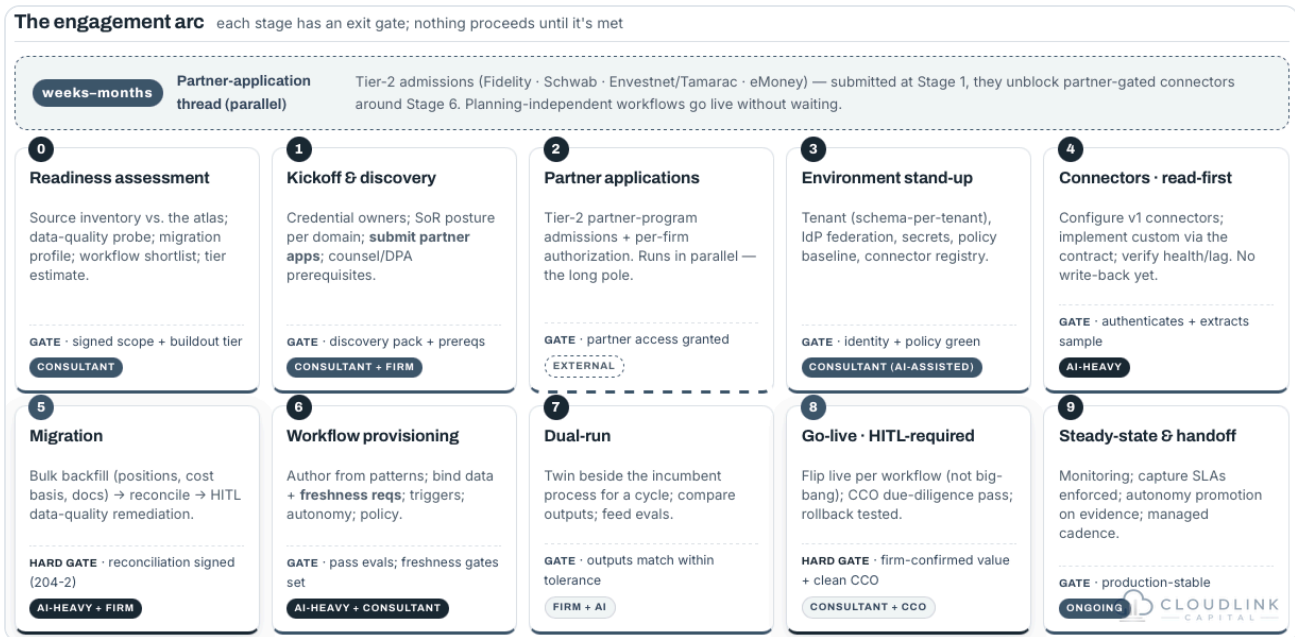


Figure 9 — An engagement arc for standing up a firm twin: each stage has an exit gate — reconciliation signed before go-live, outputs matched in dual-run before cutover — and nothing proceeds until the gate is met. *Source: CloudLink Capital.*

The twin runs beside the incumbent process until its outputs match; autonomy is promoted on evidence and revoked on regression; humans stay the authors of record throughout.

The reclaimed hours are the entry point. The owned platform — the harness, the ontology, the data — is the real asset. **Firms that own that layer compound the advantage year after year. Firms that rent everything are building on someone else's platform, on someone else's terms.**

SEE YOUR OWN NUMBERS

What could an AI-native operating model be worth to your firm?

The figures here are industry estimates. Your opportunity depends on your AUM, team, and how your advisors spend their time. The CloudLink ROI calculator models it conservatively from your own inputs — and shows the assumptions behind every number.

Run your ROI →

Sources

Industry figures are drawn from publicly reported research, including: **Cerulli Associates** (advisor metrics; RIA marketplace); **Charles Schwab** (RIA Benchmarking & Compensation Studies 2025); **Fidelity** (RIA benchmarking / productivity); **Kitces.com** (time-use, KPIs, compensation); **McKinsey** (*US Wealth Management in 2035*; advisor-shortage research); **DeVoe & Company** and **Echelon Partners** (RIA M&A, 2025); **Mercer Capital** (valuation/margins); **T3 / Inside Information** (software & AI adoption); and AI-adoption surveys reported by InvestmentNews, Financial Planning, and Publicis Sapient. Additional attributions: **Satya Nadella**, "A frontier without an ecosystem is not stable" (June 2026) and subsequent public remarks; **Gavin Baker** (Atreides Management), All-In Podcast, June 2026; **Clayton M. Christensen**, *The Innovator's Dilemma* (1997); **Palantir Technologies**, public materials on ontology-centered architecture (concept attribution); AI-lab growth reporting as covered by VentureBeat and Sacra (H1 2026). Statistics are estimates/ranges and will vary by firm and over time.

Important notices

CloudLink Capital is a technology and services provider, **not an investment adviser**, and does not provide investment, legal, tax, or compliance advice. Nothing in this paper — including the governance and regulatory discussion in Section 5 — is a recommendation or legal opinion. RIAs remain responsible for their own regulatory obligations and should consult qualified legal and compliance counsel. Product screenshots depict a **simulated advisory firm** (demonstration data; no

client information). Forward-looking and ROI statements are illustrative, not guarantees. Views on third-party products, companies, and market trajectories are the author's opinions as of the publication date.

About CloudLink Capital

CloudLink Capital builds governed, AI-native digital twins for Registered Investment Advisers — the architecture this paper describes is the architecture we build: an owned ontology and harness over the firm's existing systems, rented intelligence routed per task, and compliance translated into structural primitives. Learn more at cloudlinkcapital.com.

© 2026 CloudLink Capital. For discussion purposes.