Closing the Gap Between the Question and the Answer

Why RIAs and multifamily offices are putting a conversational layer over the systems they already own.

A wealth management firm stores data across a variety of systems. Household information sits across custodial feeds, portfolio accounting, CRM, planning software and a layer of spreadsheets. Reconciling data and automating routine workflows is challenging.

The current operating model for large wealth management firms is to manually reconcile data across all of these systems. That often means leveraging cross-functional teams from product managers to engineers and data scientists to validate routine business questions. AI is rapidly changing how data is queried and accessed throughout the enterprise. Chat is becoming the dominant channel to route questions.

The Adoption Dilemma

Integrating AI into existing data sources is challenging. Models can hallucinate or reason about data incorrectly, resulting in incorrect output. The solution is to create an ontology of data sources, key metrics, and unique variables (e.g. how to uniquely distinguish between individuals and households). Correctly mapping data across an enterprise and ensuring it stays up to date is challenging. However, not leveraging technology and AI to improve efficiency and streamline workflows results in a competitive disadvantage.

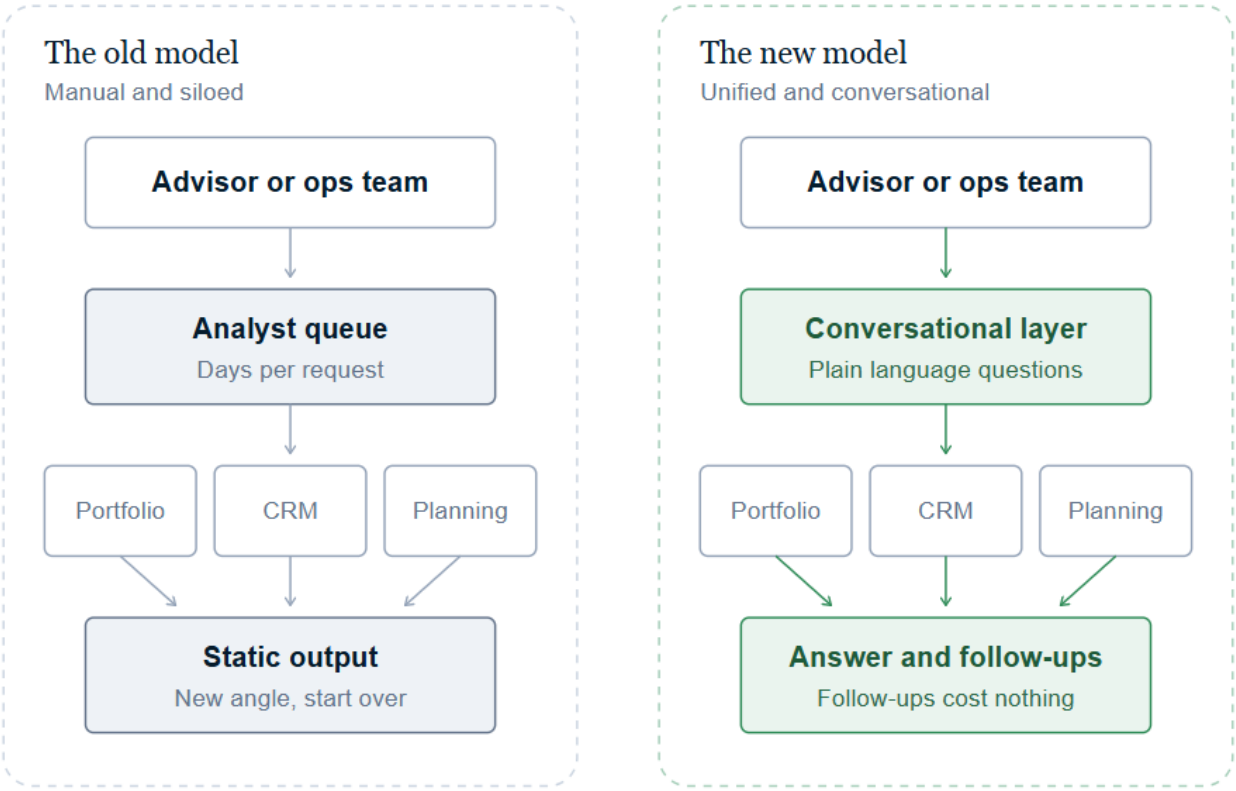
Overcoming this hurdle requires a shift in architecture. Rather than undertaking massive, multi-year data migrations to fix underlying silos, forward-thinking RIAs are deploying conversational layers that sit directly on top of their existing systems. By solving the ontology challenge at this access layer, firms are successfully transitioning from manual, heavily routed workflows to a unified, AI-driven model.

What We See Changing

	FROM - THE OLD MODEL	TO - THE NEW MODEL
Data Access	Joining data by hand and reconciling across multiple systems that store key data like household records, portfolio accounting, and CRM	One access layer over those same systems, joined at the household and the legal entities beneath it
Request path	The advisor asks operations, operations ask IT or an analyst, and the request joins a queue	The person with the question asks it directly, in plain language
Turnaround	Days, and longer at quarter end when the same people are producing client reporting	Minutes, and the same at quarter end as any other week
Output	A static file: a new angle on the same question restarts the pull	A conversation: the follow-up question costs nothing to ask
Cost behavior	Scales with households, advisors and every firm acquired	Scales with usage rather than headcount

Outside wealth management, firms further along this path have published the results: Klarna deprecated 1,200 SaaS applications and raised revenue

per employee by 152%, and LinkedIn cut issue resolution times by 29%. The mechanism is the same one described above; the numbers will not transfer directly to an RIA.



Connecting Data: context vs RAG vs agentic

Connecting internal data to an LLM accurately is where the efficiency gains are realized. The four levels below are a maturity ladder, not a sequence of steps to complete: each level costs more to build and maintain than the one beneath it, and most firms should stop climbing at the point where their questions are answered. Apply Occam's Razor - all else equal, the simplest level that works is the right one.

1. **Managing context per user** - LEVEL 1, LEAST MATURE
   No retrieval infrastructure at all. The material for one person and one task is placed directly in the model's context window. For an advisor working one household at a time, this alone handles a large share of daily questions. *Paste a household's IPS and current holdings into the model and have it draft talking points for the annual review.*

2. **Retrieval-Augmented Generation (RAG)** - LEVEL 2
   Once the data is too big and complex to hand over in advance, the system searches your databases at query time and grounds its answer in what it finds. The user stops selecting the context; the system selects it for them. *Ask which households hold more than ten percent in a single position and get the answer from portfolio accounting without an analyst pulling it.*

3. **Multi-hop RAG** - LEVEL 3
   Some questions cannot be answered by one lookup. Multi-hop RAG runs a defined sequence in which each step depends on the output of the last. The sequence is fixed and specified in advance. *Pull every account across a family's trusts and LLCs, compare the combined allocation against the household IPS, flag the drift, and draft the rebalancing memo.*

4. **Agentic workflows** - LEVEL 4, MOST MATURE
   The system determines its own sequence rather than following one you wrote. Described in the next section. *Investigate a billing break end to end: reconcile the fee calculation against the agreement, identify the cause, and prepare the correction for approval.*

Given the depth of most RIA and multifamily office questions, multi-hop RAG systems will often be sufficient. This approach balances complexity and an ability to automate routine workflows without worrying what an AI agent might do since there is always a human in the loop.

Level 4: The Path to AI-driven Systems

Where multi-hop RAG follows branching logic you defined, an **agentic workflow** determines its own path. It does not just answer a prompt; it works an open-ended problem, handling missing data and dead ends as it goes. An agent asked to review a portfolio will query current allocations, pull alternative asset data, run the risk scripts, resolve gaps iteratively, and draft a recommendation.

RIA example: the revenue lifecycle is the natural first candidate, because the rules are already written down. Fee calculation, collection and reporting

run on deterministic logic that an agent executes, including first-pass investigation of breaks. Humans approve the resolution and every client-facing communication. As the record of successful reconciliation builds, more of the investigation moves to the agent, while responsibility for client contact and for compliance with data protection rules stays with a person.

Strategic Takeaways

- **Start Simple:** Begin with prompt engineering and conventional retrieval before graduating to multi-hop or GraphRAG architecture. **RIA example:** in client onboarding, start by capturing key account data and having an agent locate contracts and accounts and flag what needs changing. Once that proves reliable, the harder task - gap-analyzing an incoming household structure against your own account schema and pricing - becomes a realistic human and agent workflow.

- **Match Tools to Tasks:** Avoid unnecessary complexity. Standard SQL and Level 2 retrieval satisfy the large majority of transactional needs. **RIA example:** most account, household and product retrievals do not require a knowledge graph. Once orchestration becomes genuinely multi-dimensional, a well-maintained RAG-based system and purpose-built client lifecycle tooling begins to make sense.

- **Embrace Autonomy Progressively:** Move from human-in-the-loop validation toward agentic workflows only as data governance and task predictability mature. **RIA example:** billing and collections run on rules an agent can execute, with a person approving break resolutions and client emails. Autonomy expands as the reconciliation record accumulates; final client-facing interaction and data protection compliance stay with a person.

Author: Brandon Goldney (September 2026)

Contributors: Russell Glorioso and Kreso Marusic