Alpha

Seven Steps to AI-Ready Data in Investment Management

AI success starts with the right data foundation

August 2026

Seven Steps to AI-Ready Data in Investment Management

The investment management industry is quickly adopting artificial intelligence, from large language models that summarize earnings calls to AI agents that execute multi-step tasks across enterprise systems. But the firms that get the most value from AI won't be those with the most sophisticated models. They will be the ones whose data is clean, connected, well-governed, and described in terms that machines can understand.

Data readiness is the foundation on which every successful AI effort is built. Without it, even the best algorithms will produce unreliable outputs, compliance risks, and wasted spend. This article lays out seven steps investment firms should take to get their data AI-ready, and an operating model for putting them into practice.

1. Establish Data Governance and Quality

Many investment firms have made real progress in centralizing their data on modern platforms like Snowflake, Databricks, or Microsoft Fabric. Others are still in the early stages, with data spread across trading systems, order management platforms, CRMs, research tools, legacy databases, and inevitably, spreadsheets. Most firms fall somewhere in between.

But having a modern data platform doesn't mean the data in it is well-governed or high quality. A Snowflake warehouse full of inconsistent entity names, stale pricing, duplicate records, and undocumented transformations will produce AI outputs that are just as unreliable as anything built on legacy infrastructure. The platform is the container. Governance and quality are what make the contents useful.

A strong governance framework, organized around formal data products, defines clear data ownership, sets quality metrics (completeness, accuracy, timeliness, consistency), and creates processes for ongoing monitoring and remediation. Every dataset that feeds an AI tool needs a named owner who's accountable for its accuracy and freshness. This isn't a one-time project. It's an operating discipline that has to run continuously.

Metadata deserves special attention here. Data without metadata, without clear records of what it is, when it was created, who owns it, how fresh it is, and what classification level it carries, is data that AI tools can't properly interpret or trust. Investing in active metadata isn't overhead. It's what turns raw data into something an AI can reason about.

Every material data asset used by AI should have metadata that answers the basic questions: What is it? Where did it come from? How current is it? Who is responsible for it?

2. Build a Unified Data Layer

Regardless of how far along a firm is in its platform journey, AI readiness requires that data be queryable from a consistent, unified layer. This doesn't necessarily mean every byte lives in one database. It means that an AI tool, or a human analyst, can access positions, trades, market data, risk metrics, performance attribution, and reference data through a consistent interface without having to know which underlying system each piece came from.

A modern data architecture typically involves cloud-native storage, proper ETL/ELT pipelines, and well-defined interfaces that abstract away the source systems. The goal is a single source of truth for each data domain, with clear ownership and well-defined update cadences. Without this, AI tools will either miss critical data or, worse, pull conflicting versions of the same data from different sources and produce inconsistent results.

For firms still early in this journey, the unified layer doesn't have to be perfect on day one. Start with the data domains most relevant to your first AI use case, build the pipelines and governance around those, and expand from there.

3. Unlock Unstructured Data

Most of the data that investment firms have spent decades collecting and governing is structured: positions, trades, prices, NAVs, risk metrics. But a huge share of the firm's intellectual capital lives in unstructured formats: research notes, earnings call transcripts, investment memos, client communications, due diligence reports, and meeting notes. This is where large language models can create some of their most distinctive value, and where many firms remain least mature.

Even among firms that have invested heavily in data strategy over the past decade, very few addressed unstructured data as part of that strategy. The focus was almost entirely on structured data: getting positions, trades, and pricing into a warehouse, building reporting and analytics on top of it. Unstructured data was treated as outside the scope of data management. That gap now matters, because it's exactly the kind of data that LLMs are built to work with.

Making unstructured data AI-ready means getting it into a searchable, indexed format where AI tools can access it alongside structured data. That could mean ingesting documents into a vector database for retrieval-augmented generation, building a document management layer with proper tagging and classification, or simply ensuring that research and communications are stored in formats and locations that AI tools can reach.

4. Implement a Semantic Layer

Ask three teams at any investment firm for AUM by region, and you'll often get three different numbers. Not because the data is wrong, but because each team defines "region" differently, pulls from a different source, or applies a different set of filters. Now imagine an AI tool trying to answer the same question. Without a shared set of definitions and business rules, it has no way to know which version is right.

A semantic layer solves this by providing a single, authoritative set of definitions for the firm's data. It specifies what each metric means, how it's calculated, where the data comes from, and how tables and fields relate to each other. When every team and every AI tool references the same semantic layer, they work from the exact same definitions, logic, and metrics. That consistency is the foundation of trust in AI-generated outputs.

Start with a Semantic Model

The semantic layer is the capability. The semantic model is the artifact that powers it. A semantic model is a structured, machine-readable description of your actual data: what your tables mean, how they relate, what your metrics are, and what business rules apply. It's tied to your warehouse and built for a specific purpose, enabling AI tools and analysts to understand what the data means, not just where it lives.

Think about what happens when you point an AI tool at your data warehouse today. It sees names like instrument, position, and party.

The LLM has to guess at relationships. Does party_id in the position table refer to the issuer, the counterparty, or the custodian? Is rate the coupon rate, the benchmark rate, or the FX rate? Does amount mean the trade amount, the settled amount, or the notional?

A semantic model, whether defined in YAML, JSON, or a platform-native format like Snowflake Semantic Views or Unity Catalog Business Semantics, fixes this by giving the AI a structured map of your data. It spells out what each table stands for in business terms, what grain it's at, what columns mean (not just their data types but their business context), how tables link to each other and what those links signify, and the business rules and derived concepts your firm uses. For instance, your semantic model might say that "investment grade" means BBB- or above, that duration buckets are short (0–3 years), intermediate (3–7), and long (7+), and that your sector groupings follow GICS but include a custom class for digital assets.

This does not typically require new database infrastructure. It sits alongside your existing data warehouse and serves as the instruction manual that makes AI tools far more effective at writing accurate queries, reading results, and giving useful answers.

Add Depth with Ontologies

A semantic model and an ontology serve different purposes. A semantic model is specific to your firm's data. It provides a single, authoritative definition for every field, table, metric, and relationship in your warehouse. It's the translation layer that tells your AI tools what your data means. An ontology is a domain representation that defines how concepts relate to each other. Industry ontologies like FIBO are generic, defining how the investment management domain works conceptually: an Issuer can have multiple Securities; Securities can be Equities, Fixed Income, or Derivatives; a Bond has a Maturity Date, a Coupon Rate, and a Credit Rating assigned by a Rating Agency; an Issuer can have Subsidiaries. It tells your AI tools how the investment world works.

Both make AI smarter, but in different ways. A semantic model gives your AI tools precision and consistency when working with your data. An ontology gives them the ability to reason across domains. Consider an AI tool trying to answer: "If this issuer defaults, what's our total exposure including subsidiaries, derivatives, and securities lending?" A semantic model can ensure the tool pulls the right fields from the right tables. But an ontology knows that an issuer can have subsidiaries, that a default can trigger collateral calls on related derivatives, and that securities lent against that issuer's bonds may need to be recalled. That's cross-domain reasoning that a semantic model alone can't do.

The Financial Industry Business Ontology (FIBO), developed by the EDM Council and governed by the Object Management Group, is the most comprehensive, open-source ontology for financial services. FIBO defines over 2,400 classes covering business entities, securities, derivatives, corporate actions, and more. Firms can adopt the modules that matter to their business, using FIBO's entity hierarchy definitions to standardize how they model legal entities, or its instrument classifications to bring consistency to how they group securities, without taking on the entire ontology. Even firms that don't deploy FIBO in its native OWL format can use it as a blueprint to align their internal data models and AI vocabularies to industry standards.

Most firms should start with a semantic model, because that's where the immediate, practical gains are. But as AI use cases mature and firms need their tools to reason across asset classes, counterparty relationships, and regulatory frameworks, ontologies become the layer that makes that reasoning possible. The semantic model tells AI what your data means. The ontology tells AI how the domain works. Firms that build both will get the most from AI.

The Emerging Context Layer

The industry is increasingly talking about a "context layer" that sits between a firm's data infrastructure and its AI tools. The context layer combines the semantic layer described above with data lineage, governance policies, access controls, and sensitivity rules into a

single governed surface of active metadata that AI agents can query at runtime. In other words, it's not a replacement for the semantic layer. It's the semantic layer plus the lineage and access control steps that follow in this article, working together as one integrated capability.

AI tools don't just need to know what the data means. They need to know where it came from, how fresh it is, what business rules apply, and what relationships connect it to other data across the firm. When all of that context is available at runtime, AI tools can reason about your data the way a knowledgeable analyst would, giving precise, well-grounded answers instead of plausible-sounding guesses. Get the context right, and AI becomes a tool people can rely on.

5. Track Data Lineage

If an AI tool tells a portfolio manager that the fund has a concentration risk in a particular sector, the PM's first question will be "based on what?" If the tool can't answer that clearly, the insight gets ignored, and rightly so. Trust is the prerequisite for adoption, and trust requires lineage.

Data lineage means tracking where each piece of data came from, how it was transformed, and when it was last updated. Provenance is the record of a data point's origin and history. Together, they let anyone trace an AI-generated output back through the model to the underlying source data and verify that the data was accurate and current when the output was produced.

This matters for internal trust first and regulatory compliance second. Portfolio managers and risk officers will act on AI-generated insights when they can see the evidence chain behind them. They won't when the outputs feel like a black box. Building lineage tracking across your data pipelines, from source ingestion through transformation to the point of consumption, is a must for responsible AI deployment.

Regulators are also increasingly expecting firms to explain the basis for investment decisions and risk assessments. If an AI model flags a risk or suggests a rebalancing action, the firm needs to be able to show that the data behind it was sound. Lineage gives you that audit trail.

6. Enforce Access Controls

Not all data should be available to all models or all users. Sensitive client data, material nonpublic information (MNPI), and proprietary research all need tiered access controls. This matters for compliance, particularly information barriers and insider trading rules, and for how AI tools are deployed inside the firm.

As firms roll out AI assistants, copilots, and analytical tools, they need to make sure these systems respect the same information barriers that apply to human employees. An AI tool used by the public equities team shouldn't have access to the private credit team's pipeline data, and vice versa. Building access controls into the data layer, rather than relying on application-level restrictions alone, creates a stronger and more auditable compliance posture.

Firms should also think about data classification frameworks that tag information by sensitivity level, retention rules, and allowed use cases. This metadata becomes increasingly valuable as AI applications spread across the organization. It's especially critical when firms begin unlocking unstructured data, where the risk of surfacing sensitive material is highest.

7. Ensure Time-Series Integrity

Investment data is tied to time, and AI models that work with historical data need that data to reflect what was known at each point in time. This requires point-in-time data modeling: the ability to reconstruct the exact state of your data as it existed on any historical date, without folding in information that only became available later.

Look-ahead bias, where a model accidentally uses future information to make "past" decisions, is one of the most common and costly mistakes in quantitative analysis. AI systems are especially prone to it when trained on poorly built datasets. To prevent this, firms should put strict data management practices in place to ensure that backtests, model training, and scenario analyses rest on "As-of" data that faithfully mirrors the information available at each historical point.

Data Products: The Operating Model for AI Readiness

The seven steps above describe what firms need to achieve. Data products are how they achieve it in practice.

A data product is a trusted, reusable data asset, owned by the business, built for a clear purpose, and designed to be easily discovered and consumed across the organization. It has a named business owner, a defined schema, documented quality standards, a delivery SLA, and an interface (an API, a table, a view) through which consumers access it. Think of it as the difference between "here's a database, good luck" and "here's a curated, documented, quality-controlled portfolio holdings data product that updates daily and is guaranteed to reconcile to the NAV."

For investment firms, data products generally fall into two categories. Foundational data products serve as shared building blocks that other products consume. Examples might include security master, pricing, credit ratings, and classifications. An entity master product, for instance, is a data product in its own right, with its own owner, schema, and quality standards. It defines the canonical view of every issuer, counterparty, and broker the firm deals with. A pricing product does the same for market data. These foundational products exist primarily to be incorporated by other data products.

Consumer-facing data products draw on certified foundational products to provide a fit-for-purpose, consumer-ready view for end users and AI tools. Examples might include positions, transactions, NAV, performance attribution, and risk metrics. A positions product, for example, may combine holdings with relevant security master, issuer, and classification data through a governed view or materialized table. Consumers access a single product, while the foundational products remain authoritative for the data they provide.

The power of the data product model is that it bundles nearly all of the readiness steps into a single, manageable unit. Each data product has a named owner from the business domain closest to the data, solving the governance problem. It lives in the unified data layer with a defined schema, solving the consolidation problem. It carries metadata and semantic definitions that describe what the data means, solving the labeling and description problem. It has documented lineage showing where the data came from and how it was transformed. And it has access controls baked in, so sensitive data products are only available to authorized consumers.

This is a fundamentally different way of thinking about data management. Instead of governing data as a sprawling, firm-wide problem, you break it into discrete products that the business teams closest to the data own and publish. The operations team owns and publishes the positions and trades products because they understand the nuances, edge cases, and quality issues better than anyone else. The risk team owns and publishes the risk metrics product. The performance team owns and publishes the returns and attribution products. Each team is accountable for the quality, freshness, and documentation of its products, and each product is built to be trusted, discoverable, and consumable by both humans and AI tools.

Firm-wide standards ensure that all data products share a common approach to metadata, semantic definitions, identifier usage, quality metrics, and access controls. Individual teams operate within those standards while retaining ownership of the data they know best. This balance between central standards and distributed ownership is what makes the model scalable. A central data team can't possibly maintain deep knowledge of every dataset across the firm. Domain teams can.

Data products also create a natural path for incremental progress. Rather than trying to make all data AI-ready at once, firms can identify the three or four data products most critical to their first AI use case, build those to production quality, and expand from there. Each new data product extends the firm's AI-ready data estate. Over time, the library of well-defined, well-governed data products grows, and the barrier to deploying new AI applications drops with each addition.

For more on Data Products, see our Data Products Primer

In Summary

The firms that will get the most from AI aren't necessarily those with the most data or the best models. They are the ones whose data is clean, connected, well-governed, and described in terms that both humans and machines can understand. Organizing that data into well-defined products with clear ownership, quality standards, and semantic meaning is the most reliable path to getting there. That's where the competitive advantage sits.

Authors

Jason Inzer | Director

Jason is Head of Data Strategy within Alpha's North American Data Practice. With over 25 years of experience in financial services, he specializes in enterprise data strategy, modern data architecture, data governance, and the data product operating model, helping investment management firms build the data foundations needed for successful AI adoption. Previously, Jason spent 15 years at Deutsche Bank, where he served in leadership roles including Head of Research Analytics for DWS and Global Head of Data for DB Equity Research.

Gaurav Jangid | Senior Partner

Gaurav is Senior Partner and Head of Alpha's Data and Technology Practice in North America. He leads large-scale data transformation programs for asset and wealth managers, specializing in data modernization, cloud adoption, and platform implementation. Gaurav has worked across strategy, implementation, and delivery for firms modernizing their end-to-end data management capabilities, including data strategy, modern data platform selection, data governance, and master data management.

Confidence in the complex