COLLATE

WHITE PAPER

Metadata ROI in the Age of AI

Executive Summary

The ROI case for metadata management has always been sound. Centralized discovery, lineage, data quality automation, and governance tooling eliminate real, quantifiable costs, from time wasted finding data to duplicate work, infrastructure bloat, and slow onboarding. The framework in this paper sizes those gains conservatively, using a $60/hour fully-loaded rate and a 46-week year. At 50 users and 5,000 tables, the productivity savings alone exceed $2 million annually. And that figure excludes the harder-to-quantify costs of bad decisions, regulatory exposure, customer churn, and lost trust, all of which industry research from Forrester, MIT, and Gartner confirms are substantial. For example, Forrester found that 25% of survey respondents estimate their companies to lose more than $5 million annually due to poor data quality, with 7% reporting their companies lose $25 million or more.

$2M

savings exceeded annually with metadata management

based on $60/hour fully-loaded rate and a 46-week year, at 50 users and 5,000 tables

AI raises the stakes categorically. Every major data platform now ships its own semantic layer. Snowflake, Databricks, dbt, Tableau, and others each maintain their own definitions of core concepts like “revenue” and “customer.” For human analysts, this fragmentation creates friction. For AI agents, it creates a structural risk: agents don’t detect semantic conflicts, they confidently return wrong answers. BCG’s September 2025 research found that only 5% of organizations are achieving AI value at scale, and 68% cite lack of access to high-quality data as a major blocker. McKinsey’s State of AI 2025 confirms that workflow redesign, not model selection, is the single strongest factor separating organizations that see EBIT impact from those that don’t. The organizations that will lead in the AI era are not those with the most data or the most sophisticated models. They are the ones that govern what their data means.

1. The metadata ROI case is proven

Time savings, duplicate work reduction, and infrastructure efficiency already justify the investment, and have for years. Research from Forrester, MIT, and Gartner confirms it.

2. AI changes the equation

Beyond productivity gains, metadata now serves a second and more critical function: preventing AI failure at enterprise scale caused by semantic inconsistency.

3. Owning meaning creates competitive advantage

As every tool ships its own semantic layer, the risk is no longer data availability, but rather, fragmented definitions. The organization that governs shared meaning wins.

The Traditional Metadata ROI and Why It Still Holds

For most organizations, the ROI for metadata management was always about productivity, i.e., helping people find data faster, reducing duplicated work, and cutting documentation overhead. These gains are real, quantifiable, and well-documented. The table below provides a framework for sizing them at your scale.

The table below assumes:

- $60/hour fully-loaded hourly rate for employees
- 40 hours per work week
- 46 weeks/year (1840 hours/year), subtracting paid-time-off and holidays
- Conservative estimates for the Quantified Improvement column

Opportunity	How Collate Helps	Quantified Improvement	Annual Impact
Time Saved in Data Discovery	Centralized search and discovery + lineage drastically reduce time spent finding data	Discovery time drops from 4 hours → 1 hour per week per user 3 hours saved x 46 weeks = 138 hours per user per year	138 hours/user → $8.28K 25 users → $207K 50 users → $414K 250 users → $2.07M
Reduced Duplicate Work	Visibility into existing assets prevents rebuilding datasets, models, dashboards	20% duplicated work → cut in half 20%/year → 368 hours 50% → 184 hours	184 hours/user → $11.04K 25 users → $276K 50 users → $552K 250 users → $2.76M
Infrastructure Cost Savings	Fewer duplicate / unused tables reduce wasted storage and compute costs	15% storage reduction 100GB per table at $0.4/GB → $40/table	1000 tables → $6K 5000 tables → $30K 25K tables → $150K (Additional savings from compute costs, not included in this model)
Productivity in Documentation & Governance	Native collaboration, tasks, automation replace Jira / Slack / manual tagging	50% reduction on 10 hours of manual effort per table → 5 hours saved per table	1000 tables → $300K 5000 tables → $1.5M 25K tables → $7.5M
Savings from Faster User Onboarding	Captured tribal knowledge, ownership, history, and lineage	Onboarding time drops from 4 weeks → 1 week	$7.2K per hire 3 hires → $21.6K 10 hires → $72K

Opportunity	How Collate Helps	Quantified Improvement	Annual Impact
Data Quality & Observability Savings	Faster detection + 50% faster resolution via profiling, tests, and lineage	40% of time on DQ → cut in half 40%/year → 736 hours 50% → 368 hours	368 hours/user → $22.08K 3 users → $66.24K 10 users → $220.8K (Additional savings from reduced revenue risk and compliance risk, not included in this model)

Even More Considerations for ROI

Note that the table above only captures the cost savings that any enterprise can reasonably calculate to understand the ROI of their metadata management investment. This means that the estimates above are very conservative because they exclude several significant costs and risks that are harder to quantify. If you consider the cost and risk dimensions below, you can see how a formal metadata management practice becomes even more justifiable.

- Bad business decisions made on wrong data
- Regulatory and compliance exposure
- Customer churn driven by data quality failures
- Loss of trust in data teams
- Opportunity cost of delayed initiatives

Industry research provides some quantification here. Forrester[1] found that 25% of global data and analytics employees estimate their companies to lose more than $5 million annually due to poor data quality, with 7% reporting their companies lose $25 million or more. In an older article that remains valid today in light of challenges of using data for AI, MIT[2] estimated the cost of bad data to be 15% to 25% of revenue for most companies. In terms of regulatory costs, Gartner[3] asserted that businesses can reduce regulatory expenses by 20% by leveraging effective governance technologies. These are examples of costs that don't appear on any infrastructure bill but show up in outcomes.

Bottom line: a metadata management platform pays for itself many times over on productivity alone, and further justifies itself when considering downstream risk and undesired outcomes.

1 Millions Lost in 2023 Due to Poor Data Quality, Potential for Billions to be Lost With AI Without Intervention, Forrester, July 30, 2024.
2 Seizing Opportunity in Data Quality, MIT Sloan Management Review, November 17, 2017.
3 Global AI Regulations Fuel Billion-Dollar Market for AI Governance Platforms, Gartner, February 17, 2026.

The Next Challenge in the Age of AI

The productivity ROI case for metadata management has always been sound, but AI introduces a second, more urgent category of value. BCG[4] recently published a report that estimates 1.7X higher revenue growth for companies at the forefront of AI innovation as compared to AI laggards. In the same report, BCG reported that agentic AI accounted for 17% of total AI value in 2025, and is expected to reach 29% by 2028. Accenture[5] reports 2.5x higher revenue growth for reinvention-ready companies, i.e., those with fully modernized, AI-led processes, compared to peers with the lowest operations maturity.

AI is proving to be extremely valuable, especially when it is tasked with taking action on data (and not just managing data). But it also introduces risk that most organizations have not yet fully priced. That's because enterprises are creating chaos due to semantic fragmentation, as every major tool they use now ships its own semantic layer, optimized locally but inconsistent globally. As a result, core concepts like "customer" or "revenue" mean different things across systems. Imagine AI taking action, at AI speeds, on misunderstood meaning.

4 The Widening AI Value Gap, BCG, September 2025.
5 How Reinvention-Ready Companies Are Driving Growth and Relevance with Gen AI, Accenture, September 2024.

This fragmentation understandably impacts humans, and it's far more dangerous for AI. A human who gets a suspicious number will pause and investigate. On the other hand, agents don't detect semantic conflicts, they confidently return wrong answers. And in enterprise settings today where AI agents are being deployed across finance, operations, and customer-facing workflows, that confident wrongness is a structural risk, not just an inconvenience.

The fix isn't another semantic layer, but a shared semantic metadata foundation that aligns business meaning, technical logic, and governance across tools.

> AI doesn't fail because of bad models. It fails because enterprises don't agree on what their data means.

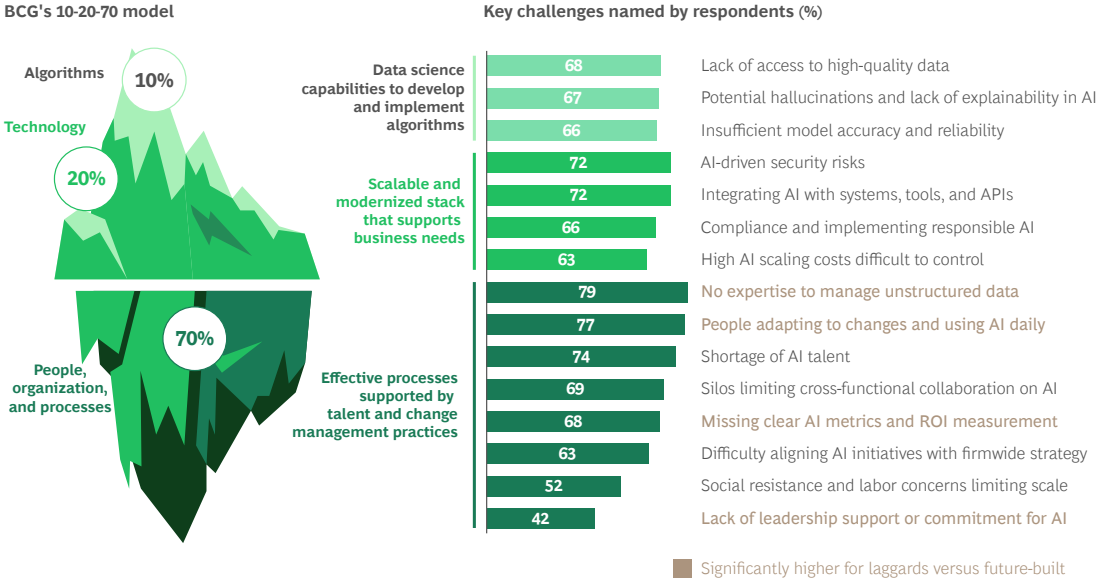
Semantic Fragmentation	Semantic Intelligence Foundation
Each tool manages its own definitions independently • "Revenue" = gross sales in Snowflake • "Revenue" = net after tax in dbt • "Revenue" = ARR only in the BI layer • "Revenue" = paying accounts in the AI platform **AI Output:** **Wrong answers. Impossible to debug. Agents select definitions arbitrarily.**	Collate semantic metadata graph serves as shared foundation • "Revenue" has multiple defined variants, each bound to business context • "Customer" definitions linked to lineage, policy, and intended use • All tools read from the same governed semantic graph **AI Output:** **Reliable, explainable answers. Auditable and policy-aware. Uses the correct definition based on business intent.**

The Foundation Matters More Than the Model

This isn't a theoretical concern. Independent research from multiple sources confirms that the primary blockers to AI success are not model quality. They are the data and governance foundations beneath the models.

In the September 2025 BCG report mentioned earlier, 1,250 executives found that 68% of organizations cite lack of access to high-quality data as a major AI challenge. Only 5% of organizations are achieving AI value at scale. BCG's 10-20-70 model is a good high-level guide to what organizations need to focus on, as it asserts that 70% of AI transformation value is determined by people and processes, versus 20% by technology and 10% by algorithms.

Most AI Roadblocks Involve People, Organization, and Processes



Source: BCG Build for the Future 2025 Global Study (n = 1,250).

IBM's 2025 CDO study[6], spanning 1,700 senior data and analytics leaders across 27 geographies, found that only 26% of CDOs are confident their data capabilities can support new AI-enabled revenue streams.

6 The 2025 CDO Study: The AI multiplier effect, IBM, November 2025.

Deloitte's quarterly tracking[7] of 2,773 enterprise leaders found that improved data management has been the one constant priority across every survey wave, unchanged even for organizations that consider themselves data-mature.

McKinsey's State of AI 2025[8] found that while 88% of organizations now use AI in at least one function, only 39% report any measurable impact on earnings. And among 25 factors tested, workflow redesign had the single strongest effect on an organization's ability to see EBIT impact from AI, not which models they chose.

The pattern is consistent across every major research body tracking this space: organizations are not failing at AI because their models are insufficient. They are failing because data is fragmented, semantics are inconsistent, governance is not centralized, and every new tool added to the stack increases entropy rather than reducing it. AI is not creating this problem, it is exposing it faster than previous waves of tooling could.

> "To realize AI's full potential, a strong data foundation isn't optional, it's mission critical. If your C-suite still considers data engineering as a support role, you're already five years behind."
>
> **— Chris Child, VP of Product, Data Engineering, Snowflake**

7 Now decides next: Generating a new future, Deloitte, January 2025.
8 The state of AI in 2025, McKinsey, November 2025.

A New Category of ROI, From Productivity Gains to Semantic Risk Reduction

The traditional ROI case for metadata (i.e., faster discovery, less duplicate work, lower infrastructure cost) remains valid and quantifiable, but it is now a floor, not a ceiling. In the AI era, metadata delivers a second category of value that is harder to model in a spreadsheet but strategically more significant.

Semantic Risk Reduction

Every disconnected semantic layer is a liability. When AI agents query data whose meaning varies by tool, they return answers that are confidently wrong and impossible to audit. A shared semantic metadata graph eliminates this class of failure by ensuring consistent interpretation across all systems.

Integration Complexity Reduction

Organizations with federated data architectures (multiple cloud platforms, regional data ecosystems, acquired company stacks) face exponential integration costs when each system maintains its own definitions. A unified metadata foundation dramatically reduces the mapping and reconciliation effort required to connect them.

Self-Service That Actually Works

Self-service analytics fails not because tools are inadequate, but because users can't trust the data they find. When data assets carry governed definitions, ownership, quality scores, and lineage, business users can act on data without routing questions through data platform teams.

AI Governance & Audit Readiness

Regulators and internal audit functions increasingly require organizations to explain how AI-driven decisions were made. This requires not just data lineage but semantic lineage, the ability to trace which definition of a term was used, under which governance policy, in which context.

How Collate Addresses Both Dimensions of ROI

Collate is the Semantic Intelligence Platform built on OpenMetadata. It delivers both dimensions of metadata ROI, the quantifiable productivity gains from the traditional catalog case, and the structural risk reduction that AI-era enterprises require, through a single, unified platform.

The Semantic Metadata Graph

At the core of Collate is the OpenMetadata semantic metadata graph, a governed, shared model of meaning that connects business definitions, technical logic, and governance policy across all tools and platforms. Unlike tool-specific semantic layers that optimize locally and fragment globally, the semantic metadata graph serves as the single authoritative source for what data means, who owns it, how it flows, and under which policies it may be used.

When Databricks, Snowflake, dbt, Tableau, and your AI platform all read definitions from the same graph, the fragmentation problem disappears. "Revenue" means what your organization has decided it means, consistently, everywhere, with full lineage and governance context.

AI Capabilities Built on Semantic Foundation

Collate AI capabilities, i.e., AskCollate, AI Studio, and AI SDK, are what distinguish it from legacy technologies that provide limited productivity gains. These are not AI features bolted on top of a metadata store; they are AI agents that reason over a governed semantic graph, which is what makes their output trustworthy rather than merely fast.

- **AskCollate:** Conversational AI that answers questions about data using governed metadata as context, not just search results, but semantically grounded responses that reflect actual definitions, lineage, and ownership.
- **AI Studio:** A no-code agent builder for creating scheduled governance workflows (PII classification, data quality testing, lineage enrichment) that run continuously without manual intervention.
- **AI SDK:** A Python, Java, and TypeScript framework for building custom AI agents that leverage the Collate semantic metadata graph programmatically, enabling teams to extend governance into any workflow or application.

120+ Native Connectors

Collate connects to 120+ different data sources including databases, cloud warehouses, data lakes, BI tools, pipeline orchestrators, SaaS applications, and storage systems. This ensures the semantic metadata graph reflects the full data estate rather than a curated subset. Coverage matters because fragmentation hides in the systems you haven't cataloged yet.

Conclusion: Owning Meaning Is the New Competitive Foundation

The organizations that will lead in the AI era are not necessarily those with the most data or the most sophisticated models. They are the organizations that have established governance over what their data means consistently across every tool, team, and use case.

Metadata has always delivered measurable ROI. The productivity case is proven and quantifiable. Time savings, duplicate work reduction, and infrastructure efficiency justify the investment many times over before any AI use case is considered. But AI raises the stakes categorically. Tool proliferation without semantic governance no longer just creates friction, it creates AI systems that return confident wrong answers, fail audit scrutiny, and erode the trust that makes AI-driven decisions actionable.

The fix is not another semantic layer. Every platform already ships one. The fix is a shared semantic metadata foundation that all tools read from. One that is governed, versioned, lineage-aware, and owned. That is what Collate provides.

In closing, keep these three principles top-of-mind as you continue on your data journeys:

- **Semantic consistency is a prerequisite for scale.** Separate data ecosystems, acquired stacks, or multi-cloud architectures only work reliably if core concepts like “customer,” “revenue,” and “quality” mean the same thing everywhere. Without that, you scale inconsistency, not data.
- **Self-service only works with shared meaning.** Removing the data platform team as a bottleneck requires data products that are discoverable, owned, and trusted by definition, not by tribal knowledge.
- **AI needs governed context to be safe.** LLMs and agents can’t detect semantic ambiguity. A unified semantic layer is what prevents hallucinations, audit risk, and loss of trust.

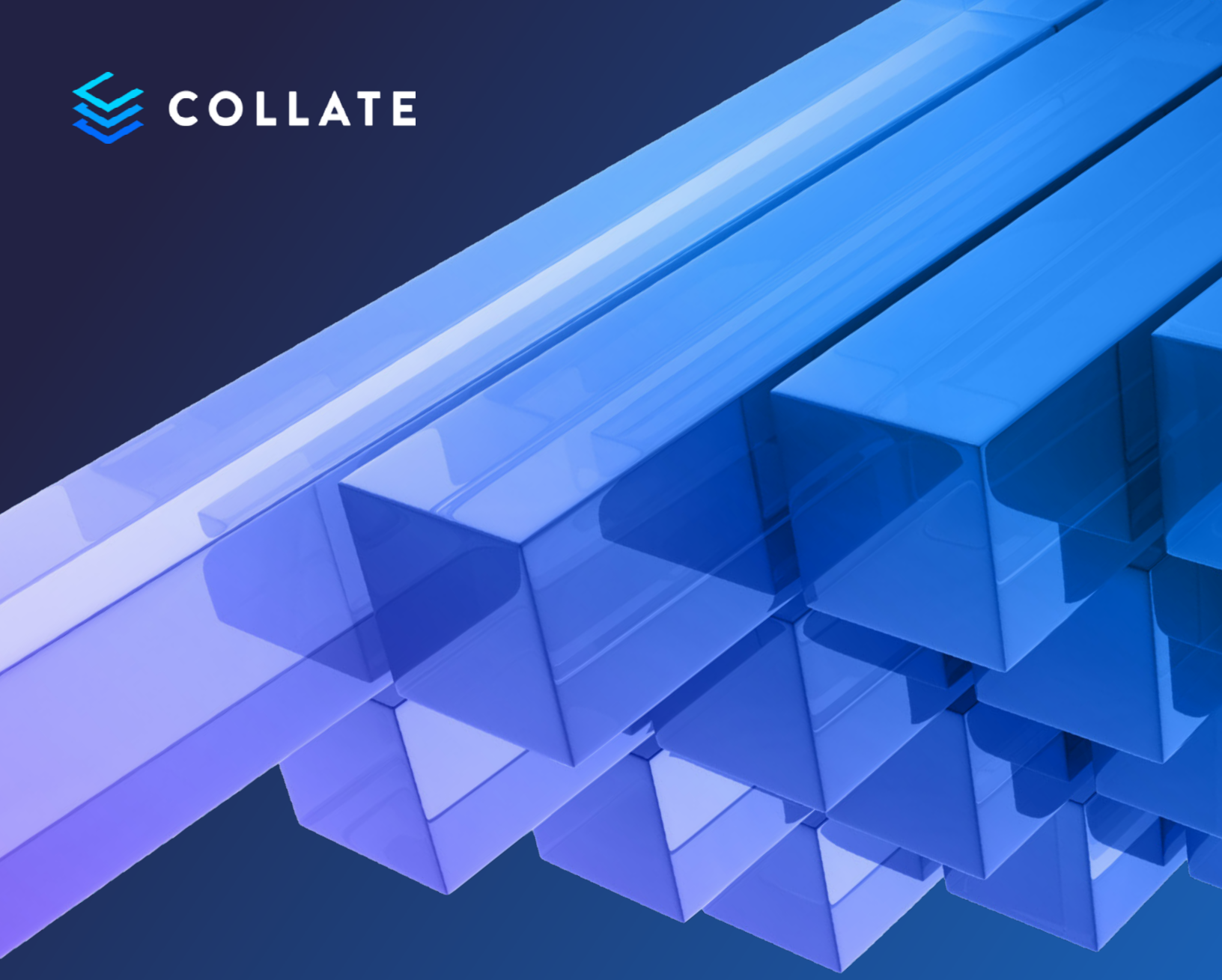
In the AI era, competitive advantage shifts from owning data to owning meaning.

Ready to see Collate in action?

Schedule a personalized demo to see how Collate can help your organization establish semantic consistency, discover and govern your data, and build the AI-ready foundation your data teams need. Contact us to get started.

About Collate

Collate is the Semantic Intelligence Platform and the company behind the OpenMetadata project. It turns metadata into shared meaning so people and AI can work from the same understanding of data. Collate applies that semantic foundation across discovery, lineage, quality, observability, and governance to enable trusted analytics, explainable AI, and automated governance at enterprise scale. Global 2000 companies and innovative startups rely on Collate to accelerate insights and build AI-ready data foundations. Headquartered in Silicon Valley, Collate is backed by world-class investors including Venrock, Unusual Ventures, and Karman Ventures. Learn more at getcollate.io.



Interested in learning more about Collate? Email sales@getcollate.io to book a demo with a product expert.

getcollate.io