Data Quality in the Agentic AI Era

Alation

Introduction

No one operating a business that collects and uses data would argue that data quality is unimportant. For anyone working in the data management or data governance space, the reported risks of working with poor-quality data and the published statistics on time and money lost due to ineffective or insufficient data quality are tired mantras.



Organizations lose money.

Over 25% of organizations lose more than $5 million annually due to poor data quality, with 7% losing $25 million or more.

Forrester, 2023



Data scientists lose time.

Data professionals spend up to 80% of their time on preparation tasks, including cleaning and organizing data.

Pragmatic Institute, 2024



Data consumers lose trust.

Poor data quality can result in misguided actions, wasted resources, and missed opportunities.

Dataversity, 2025



Data stewards lose their minds.

“There are two things that will make your team look like amateurs: someone pinging you on Slack… or — worse — you not even knowing about it.”

Medium, 2025

As data volumes grow and reliability lags, organizations face a growing data trust gap — a disconnect between data producers and consumers that undermines confident decision-making. Chief Data Officers (CDOs), data stewards, and data engineers are well past the point of wondering if they should monitor data quality; the challenge these days is more about the how: How do we know what to monitor? How do we monitor quality at scale? How on earth do we get started?

This paper explores the challenges of building a robust data quality monitoring program at scale, including the limitations of current AI and ML approaches. It introduces agentic AI as a new path forward — and outlines how Alation's Data Quality Agent uses this approach to help organizations intelligently prioritize what to monitor, recommend and apply the right checks, and resolve issues more efficiently with proactive alerts.

Tackling Data Quality Monitoring At Scale

Teams in organizations at varying levels of data governance maturity report many of the same challenges in tackling data quality monitoring:

- Data is stored and accessed within silos, often accessible only by certain groups which may adhere to different rules for governance and data quality.
- Data stewards and engineers struggle to communicate the data quality rules required to make datasets fit-for-purpose for data consumers.
- Massive volumes of data spread out across data sources potentially require a huge amount of resources — time, expertise, and tools — to apply consistent data quality standards at scale.
- Alert fatigue and laborious root cause analysis require the right information and context to resolve data quality issues quickly.

Consider a global consumer goods company based in the U.S. Its data governance team uses a data quality tool to monitor parts of its ecosystem. But with teams working in silos, applying consistent rules across all systems is difficult. Access is another issue: only a few people can use the standalone data quality tool, leaving many data consumers in the dark about the accuracy or timeliness of the data they rely on. Even those with access to multiple tools struggle with the constant switching between data sources, catalogs, and quality platforms, making it hard to connect the dots. Despite having a mature data quality program, the company still falls short of meeting the business's needs.

Another example is a multinational manufacturer of industrial products. Its data team uses custom SQL queries to monitor data quality, but these don't scale to the size and complexity of their data estate. Their first priority is to improve how they manage known issues by applying broader, less-customized rules to data they already monitor. The bigger challenge lies ahead: identifying unknown issues and applying the right quality checks across their entire ecosystem, beyond what current tools and practices can support.

Knowledgeable and motivated as these teams are, the problem they are addressing isn't so much why data quality monitoring is important or convincing the business to take data quality seriously, but addressing the initiative at scale in the immediate term with the resources at hand. To continue doing what they are doing is simply not a viable option if they are to keep up with the business demand for accessible, trusted, properly-governed, high-quality data.

Not surprisingly, managing data quality is a top priority for organizations across industries. In an IDC survey in mid-2024, CDOs listed "Data quality management" as the second most important strategic investment priority for data in the next 12-18 months. (AI-Ready Data: Foundation for AI-Fueled Business, September 2024). And, as the examples above illustrate, the risks of working with poor-quality data and the practical challenges of establishing and maintaining data quality monitoring are compelling enough to warrant not just investment in data quality tools and expertise, but proactive, forward-thinking investment in using AI to tackle the problem.

These days, AI seems like the answer to almost everything. Generative AI has been heralded as a productivity game-changer (though measuring that gain continues to prove somewhat elusive), and retrieval-augmented generative AI (RAG AI) may reduce hallucinations. Still, its usefulness in data quality monitoring remains debatable.

Relative to the ML models that were trained on our data to recognize and leverage patterns in content and usage, GenAI seemed a leap forward, an astonishing, practical, and accessible tool for answering our questions and recommending solutions or actions. Yet its exact role in monitoring data quality is still evolving; AI involvement often manifests as question-and-answer interactions, tools with the intelligence to address specific requirements, like "create a data quality monitor that validates email address format," but one must know what to ask. Prompt engineering is amazing, but is it what teams need to tackle their immediate, real-world, resource-limited data quality monitoring initiatives at scale?

The Limits of ML and GenAI for Data Quality Monitoring

ML and GenAI hold real promise for improving data quality, but they can't solve every challenge. Without diving into a full-blown feature comparison of tools and vendors, here are three key areas where current AI solutions still fall short:

1 **Determining which data to monitor.**

A common instinct is to monitor all data — but that's neither practical nor necessary. Not every dataset is critical. If data isn't actively used in pipelines or doesn't feed into reports, monitoring it may only generate noise and lead to alert fatigue.

The smarter approach is to focus on the most valuable, frequently used data. But GenAI or ML cannot easily determine those critical assets on their own. While ML can profile and classify data, and GenAI can generate lists based on metadata and user prompts, neither is equipped — without significant customization — to intelligently prioritize which data assets should be monitored for quality.

2 **Applying rules at scale.**

Data teams often face the overwhelming task of managing thousands — or even hundreds of thousands — of tables across many data sources. Generating data quality monitoring rules with GenAI can speed things up, but rule creation is only half the battle.

Variations in column names, value ranges by region, and use-case-specific thresholds all complicate the process. ML-based anomaly detection helps surface issues, but it doesn't answer the tougher question: What quality rules should I apply across all my database tables? For that, human judgment, context, and domain knowledge are still essential.

3 **Gauging quality and providing the context to fix issues.**

Data quality tools on the market today excel at measuring and monitoring data health and issuing alerts. Still, the onus ultimately falls on a human to respond to the alert, determine the root cause, and then fix the problem. Lineage charts and pipeline diagrams with overlaid quality test results help, but addressing data quality issues can still be a disjointed and time-consuming experience.

In an ideal world, an AI could take up the diagnostic mantle and recommend remediation, or better yet, take autonomous action to remediate the issue itself, which is, as of yet, a future-state phenomenon.

To tackle these challenges, Alation has built an AI agent for data quality.

How the Alation Data Quality Agent Works

In a rapidly evolving arena of AI development, AI agents are remarkable entrants. Still new enough technology to be a bit subjective in its definition, an AI agent is software that can collect or access data, then use it to make decisions and take actions with minimal human intervention. Humans decide on the objective of the AI agent; the agent chooses how to achieve the objective. Alation Data Quality, as part of the Alation Agentic Data Intelligence Platform, maximizes the semi-autonomous abilities that AI agents provide.

Alation is a well-established and widely used platform for data cataloging and governance, and the engineering team spotted a unique opportunity to leverage existing cataloged metadata and an LLM service to build an AI agent that addresses the real-world gaps that data teams face as they start or scale their data quality monitoring initiatives.

Manifesting as a completely UI-based guided workflow, the Alation Data Quality Agent intelligently identifies critical data assets to monitor for quality, recommends appropriate data quality checks for each asset, and requests human approval of the checks before it starts executing them according to a regular hourly, daily, or weekly schedule.

Here's how it works:

1. Accessing cataloged metadata, Alation's Behavioral Analysis Engine prioritizes optimal assets for data quality monitoring based on usage that indicates a table's popularity. Where possible, it uses lineage information to identify tables that feed business intelligence reports which, therefore, qualify as optimal.

2. The agent uses common attributes of the assets (like data source, table and schema names, column ID, column name, and column data type) to construct a set of prompts that it sends to an LLM service.

3. The LLM then generates and returns recommendations for suitable check definitions for individual columns according to an open-source data quality rule library. For example, if the metadata analysis identifies a column as containing phone numbers, the LLM suggests a check to validate that all values in the column conform to a phone number pattern such as (xxx) xxx-xxxx.

4. The agent presents the check recommendations to an Admin or Steward in the catalog for their review, optional tweaking, and approval. It also presents options for scheduled check runs and arranges to send alert notifications of data quality check failures to a messaging tool, like Slack or MS Teams, so recipients can take action to resolve issues.

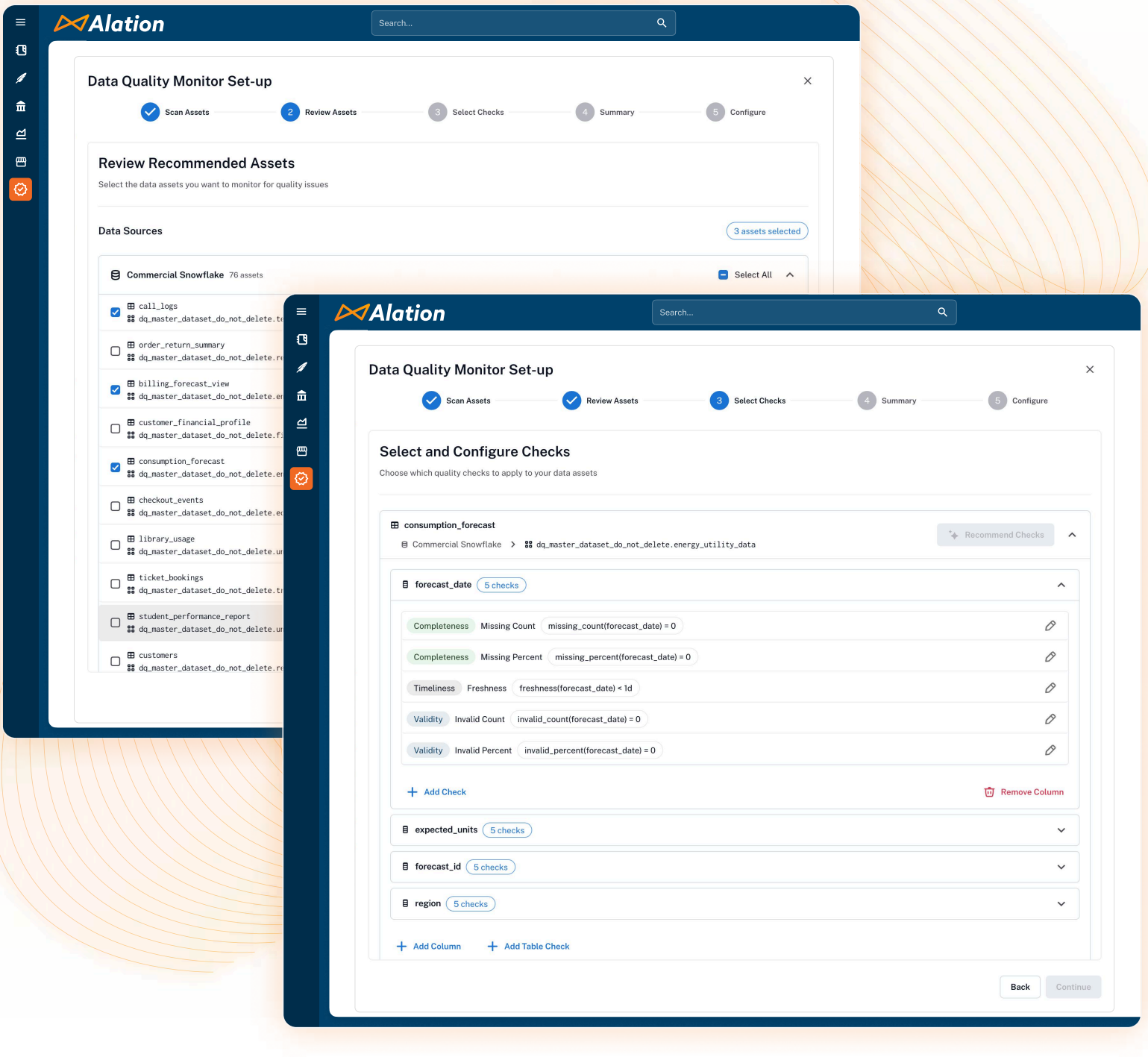


Figure 1: Review assets and AI-recommended data quality checks

This evolved experience, while not fully autonomous, significantly surpasses the capabilities of traditional GenAI tools. Unlike a GenAI model that simply suggests options based on general prompts, the Alation Data Quality Agent acts with contextual awareness. It analyzes metadata, understands data usage patterns, applies predefined logic, and proactively recommends tailored quality checks — then lets the user approve, refine, or schedule them.

It's the difference between asking GenAI, *"What vegetarian meals could I make tonight?"* and receiving a generic list of five popular recipes, versus asking an AI agent and getting, *"Based on what's in your fridge and what's expiring soon, you should cook this specific meal — and I'll preheat the oven for you."* The Alation Data Quality Agent doesn't just suggest ideas; it moves you closer to action with insight, precision, and practical support.

A Unified, Simplified Data Quality Solution

For data governance and analytics professionals, the idea of using an AI agent to launch or level up a data quality monitoring program is more than just intriguing — it's practical and increasingly necessary. That's no coincidence. Alation's research into customer needs uncovered a widespread desire for a unified, easy-to-deploy data quality solution that could scale quickly.

Teams across financial services, consumer goods, and multinational manufacturers echoed similar challenges. Regardless of industry, geography, or data maturity, one truth stood out: teams needed a smarter, simpler approach to data quality.

Data governance teams often lack the time, specialized skills, or resources to start — or sustain — a scalable, extensible data quality program. Yet, monitoring their most critical data assets is non-negotiable.

CDOs, CTOs, data stewards, and data engineers want actionable data quality insights embedded directly within the data catalog before considering separate tools or attempting to unify fragmented solutions spread across multiple platforms.

There's a growing demand for automated data quality checks that integrate with existing governance and compliance workflows, reducing manual effort and increasing consistency.

Everyone dreads alert fatigue. Teams need smart notifications that surface the right issues without overwhelming users with noise.

Working closely with early adopters, Alation confirmed that its Data Quality solution delivers on these expectations. Integrated into the Alation Agentic Data Intelligence Platform and powered by an AI agent, Alation Data Quality adds meaningful value from day one. As teams begin using it, the benefits become clear: data quality that's easy to adopt, seamlessly embedded, and built for scale.

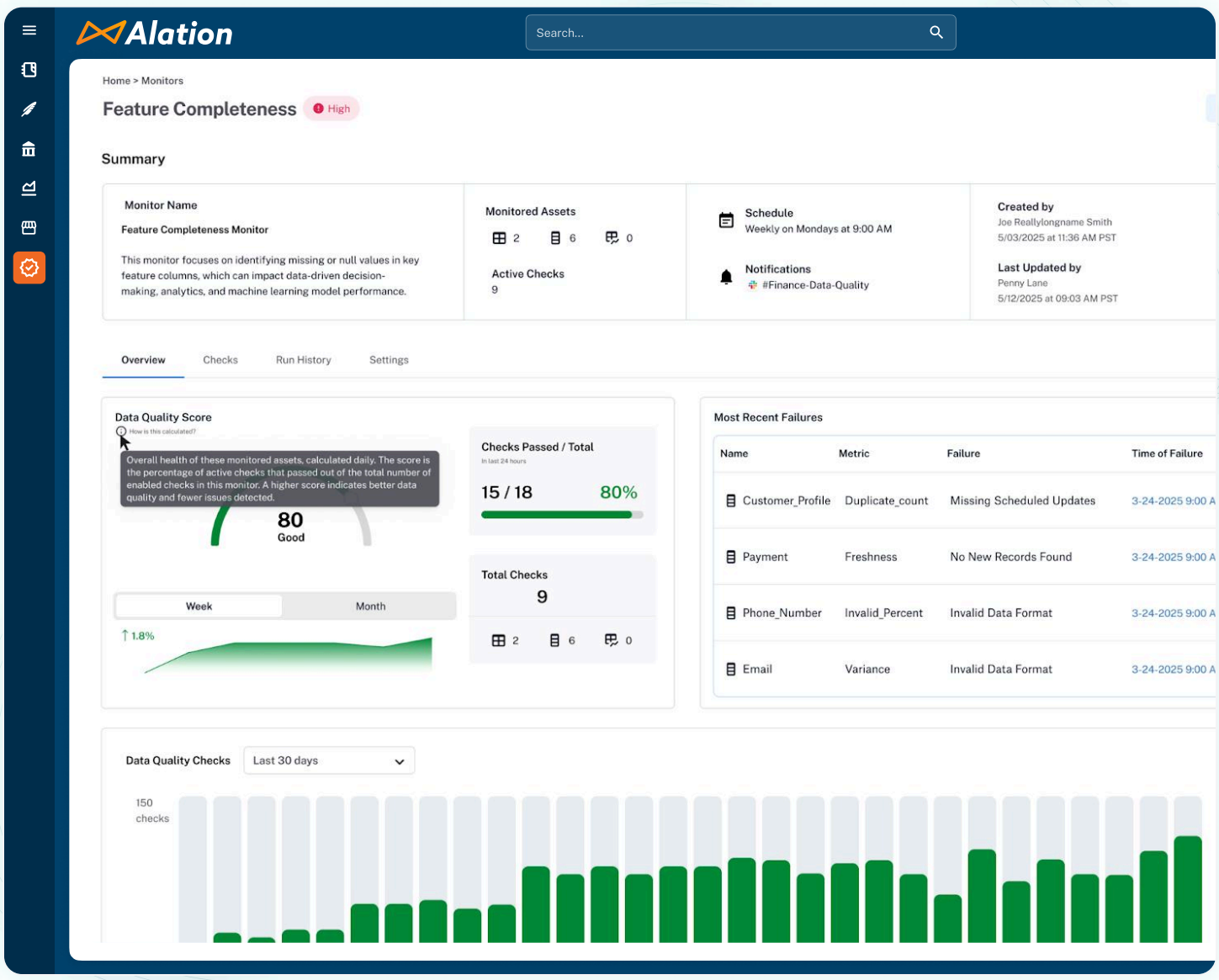


Figure 2: Review data quality monitor details to assess data health and take action to improve it

Imagine an organization with minimal data quality monitoring in place. A recent marketing campaign suddenly drives a spike in data usage, turning what was once a tolerable data issue into an urgent, business-critical problem. The need for rapid, scalable data quality monitoring and resolution becomes immediate.

In this scenario, the expected business outcomes from deploying Alation Data Quality would be both timely and transformative:

- **Intelligently prioritized initiative:** Alation's Behavioral Analysis Engine intelligently identifies the critical data assets across the entire data estate — the ones that consumers in the organization depend upon the most — that would result in the most impact to business operations when monitored for quality.
- **Data quality monitored at scale:** The agent recommends basic quality checks for hundreds of assets in multiple data sources. The humans in the loop would only need to approve the recommendations and set a schedule to get immediate-term visibility into data health.
- **Proactively resolved issues:** Thanks to the unified catalog experience that positions data quality within lineage charts, data stewards who get alerts on poor-quality data could quickly address root causes to resolve issues.
- **Increased trust in data:** Data consumers would have direct access to data quality health information directly in the catalog for a unified experience that gives them confidence in the data they use to make decisions and build models, dashboards, and reports.

Now consider a company with a more mature data culture — one that already uses a dedicated data quality tool to monitor key tables in platforms like Databricks and Redshift. Even in this more advanced scenario, challenges remain. Data consumers often struggle to connect data health insights from separate tools with the broader governance context, leading to a fragmented, inefficient experience that's hard to reconcile with enterprise-wide data policies.

By implementing this agent, this organization could see immediate and meaningful improvements across its data quality efforts:



Multiple approaches for efficient coverage: The team integrates their existing data quality tool with Alation via the Open Data Quality Framework, allowing test results to surface directly in the catalog. This gives users seamless visibility into data health without switching tools or losing context. Then, by layering in the Alation Data Quality AI Agent, they identify gaps in monitoring — especially on the most frequently used assets — and fill those gaps with AI-recommended checks. This dual approach preserves the value of their current investments while enhancing coverage with Alation's native AI-driven capabilities.



The AI agent got them there faster: Even with experience in data quality, the team uses the AI agent to significantly speed up their efforts. It generates smart check recommendations that they quickly review, tweak, and approve. For more complex scenarios, they create custom SQL checks directly within Alation, keeping all quality insights centralized in one platform.



Context invited collaboration: By displaying data health in the context of where consumers use it (Power BI, Alation Table pages, etc.), more users expect quality coverage on their tables and, in some cases, could use the AI agent to create their own no-code quality checks directly in the catalog. Data quality monitoring could finally breach the silos!

While this paper focuses on the use of agentic AI for data quality monitoring, it reflects a broader shift in how Alation is enabling intelligence across the data ecosystem. Trusted, high-quality data — paired with richly contextual metadata — is essential not just for governance, but for powering the next generation of agentic workflows and AI applications. As organizations move toward more autonomous, insight-driven decision-making, platforms like Alation are uniquely positioned to provide the foundation these workflows require. With the recent addition of Numbers Station, Alation continues to expand its ability to support agentic workflows that automate tasks, guide users, and adapt to business context — all grounded in trusted data and governed metadata.

Maximizing Efficiency and Inspiring Innovation

This bold new era of AI agents is ripe with possibilities for data quality monitoring and data product development. These latest initiatives, which fully exploit an evolved, semi-autonomous AI, serve not only to maximize efficiency but also to inspire innovation!

Alation has only explored the tip of the iceberg and is eagerly peering into the water to see how deep it goes. Imagine never again having to write scripts to weed out null values, laboriously normalize data entries, or guess the accuracy of the data that feeds a report on projected revenue. Hindsight will have us all wondering how we ever managed to do anything without agentic AI intelligently monitoring data and fixing quality issues.

CDOs, data stewards, and data engineers today definitely have an advantage over all those who came before them as they begin, or level up, their data quality initiatives. AI agents do not (yet!) solve every problem in a complex, extensive data ecosystem, but addressing even a fraction of the manual labor needed to reliably monitor data quality and facilitate data governance in an organization is a tremendous win.

By helping teams prioritize, monitor, and remediate data quality intelligently, agentic AI provides a scalable path to closing the data trust gap — and rebuilding confidence in every data-driven decision.

Learn more about Alation Data Quality, or contact us for a personalized demo.