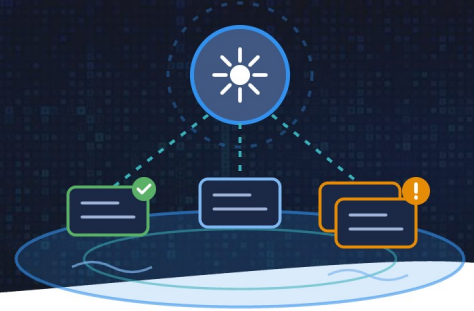


Why Your Data Lake Isn't Ready for AI Agents Yet

A practical brief for data and AI leaders



Where most teams are today

Most enterprises have spent the last few years moving data into a lake or warehouse. The idea was to put everything in one place and use it as the foundation for analytics and, eventually, AI.

That part mostly happened. What didn't change much is the data itself. Duplicate customers, terms that mean different things to different teams, and records nobody owns all came along for the ride. Analysts still go back to source systems when they need a number they can defend, and it's still common for finance and sales to show up to the same meeting with different revenue figures.

With dashboards, people learned to live with it. They knew which records to ignore and fixed the rest in their heads. **But AI agents can't do that.**

Why this is getting more urgent

ON A DASHBOARD

When a dashboard has a duplicate customer, someone sees a wrong number and usually catches it.

WITH AN AGENT

When an agent has a duplicate customer, it might send an invoice to an old account or answer a support question using outdated contract terms. The mistake isn't sitting in a report waiting to be noticed. It has already happened.

And agents are coming quickly, while Gartner has been pointing at data as the bottleneck for a while.

3 in 4

companies expect to use agentic AI at least moderately within two years
Deloitte, 2026

21%

say they have a mature way to govern those agents
Deloitte, 2026

60%

of AI projects without AI-ready data will be abandoned through 2026
Gartner prediction

63%

of organizations lack the right data practices for AI, or aren't sure
Gartner

What we're seeing from teams that get agents into production

The teams that make it past the pilot stage tend to have a few things in common.

They start small.

Usually that means one use case and one or two types of data, like customers or products, cleaned all the way through. Deloitte's research points the same way: the most successful companies begin with lower risk use cases and expand from there.

Governance happens during the build.

Owners, definitions, and access rules are decided before an agent goes live, so the first mistake isn't what forces the conversation.

They fix data where it's created.

Cleanup logic doesn't get rebuilt in every report and model built on the lake, because issues are resolved before data reaches the lake.

The business is involved early.

The people who create and use the data help decide what "clean" means, which makes adoption much easier after launch.

In the conversations we've had with data leaders this year, the model is rarely what holds things up. A pilot runs on a clean sample of data someone prepared by hand. Then it moves to production, starts reading from the lake everyone actually uses, and stalls.

Five things to fix before agents touch the lake

You don't need a new platform to start on any of these. Mostly, they come down to deciding where problems get fixed and who's responsible for fixing them.

01 Clean up your most important data before it reaches the lake.

Customers, products, suppliers, and locations cause most of the trouble. If duplicates and conflicts get resolved in the systems where the data is created, the lake receives one version of each record to work with.

- Start with whichever type of data your first AI use case leans on most.
- Count how many duplicate records it has now, so you have a baseline to show progress against.

02 Settle on one definition for each critical term.

If sales, finance, and support each have their own idea of an "active customer," an agent has no way of knowing which one you meant.

- Make a list of the 10 terms that cause the most debate in meetings.
- Write down one definition for each and assign someone to own it.

03 Give every area of data an owner, like customer or product data.

When customer data belongs to nobody, each team patches it locally and the fixes don't stick.

- The owner should sit in the business. IT can support, but the business knows what the data should look like.
- Give each owner a simple quality target they can track.

04 Limit what agents can access.

An agent should only see the data its job requires. Personal information, financial data, and confidential information each need their own rules.

- Before an agent goes live, label the data it will use by how sensitive it is.
- Keep a log of what each agent accessed and why.

05 Treat cleanup as ongoing work.

A one-time cleanup starts to decay almost immediately. New records come in, people change jobs, and systems overwrite each other when they sync.

- Add checks where data enters, in addition to where it's reported.
- Watch for the same issues showing up quarter after quarter. That usually means the problem starts at the source.

WHERE TO START THIS WEEK

Pick the AI use case your team is most excited about. Write down the three types of data it depends on most, like customers, products, or suppliers, and find out who owns each one.

If nobody can tell you, you've found your starting point.