Open in app ↗

Medium

Access to everything. Now 30% off. Upgrade now

Predict · Follow publication

You're reading for free via Saurav Kumar's Friend Link.

Member-only story

AI ✓ Data Engineering + Data Architecture + Analytics + Big Data +

How AI Agents Actually Consume Data (And Why Your Pipelines Aren't Ready)

We spent years optimizing data for dashboards. AI agents need something very different.

6 min read · Apr 16, 2026



Saurav Kumar Follow

Listen Share More

Non-member, click here to read

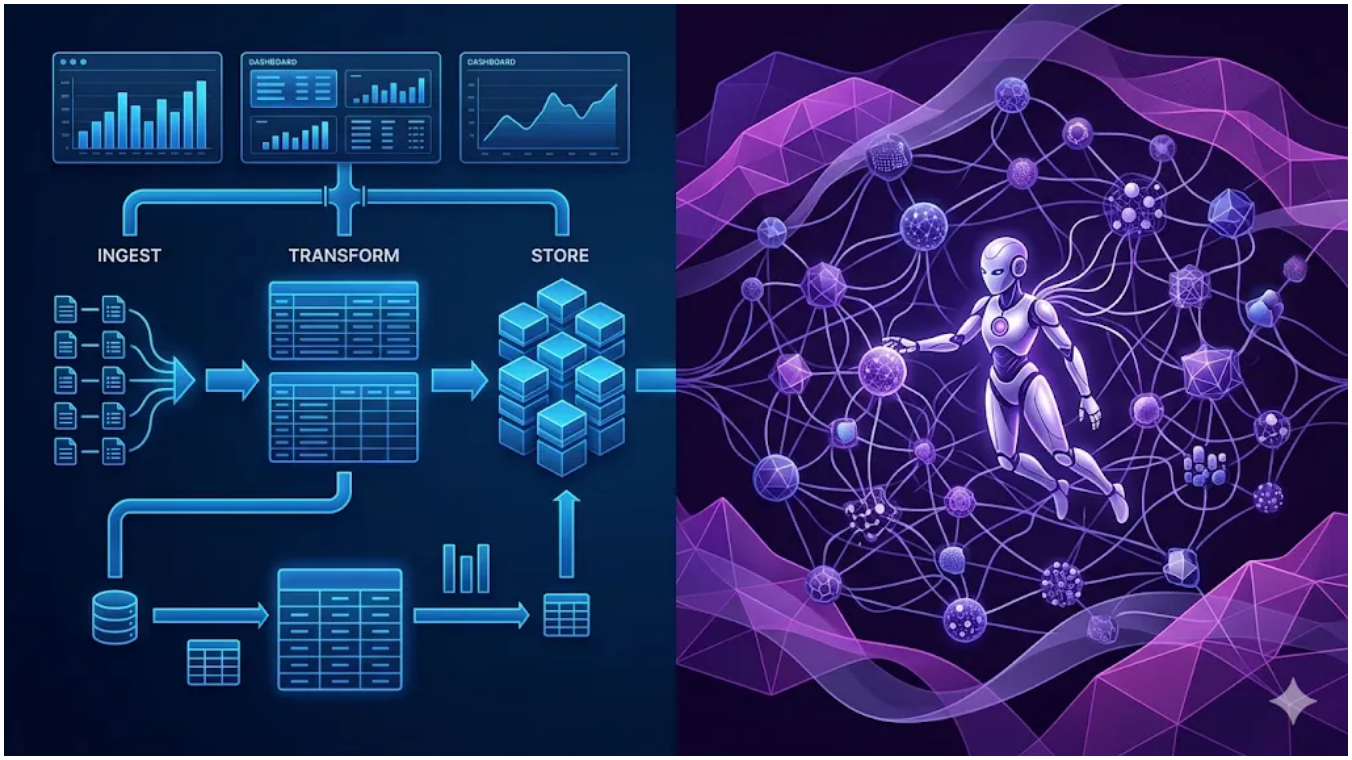


Image generated using AI

Most people think AI just “uses data.”

Like, it pulls from a neat database, runs some logic, and gives you an answer.

That picture feels clean. Almost comforting.

It’s also… not how things work anymore.

And if you’re building pipelines with that assumption sitting somewhere in the back of your mind, there’s a good chance you’re solving the wrong problem without even realizing it.

The mental model that’s quietly breaking

For years, the flow looked something like this:

Data → Pipeline → Dashboard → Human

You move data, clean it, shape it, and someone eventually looks at it. You answer questions. You build reports. You refine things over time.

It's structured. Predictable. Stable.

And honestly, it worked really well.

But AI agents don't sit at the end of that pipeline like a dashboard waiting for...

They don't wait for someone to ask the right question.

They behave differently. A bit messier. A bit more curious.

More like:

Data → Context → Retrieval → Reasoning → Action

That shift might look small on paper, but it changes almost everything.

Because now the system isn't just responding.

It's exploring.

AI doesn't really consume data. It consumes context

This is the part that took me a while to understand.

Traditional pipelines are built to answer known questions. What happened yesterday? How many users signed up? What's the revenue trend this quarter?

You define the question first. Then you prepare the data to answer it cleanly.

That's the whole game.

But AI doesn't start with a fixed question.

It figures things out as it goes.

Which means it needs more than clean, structured tables.

It needs raw signals. It needs relationships. It needs metadata that explains what something means. It needs history, not just snapshots.

And maybe more importantly, it needs the ability to move between these things freely.

Here's where things get a bit uncomfortable.

A lot of what we do in data engineering is about removing noise. We filter. We aggregate. We simplify. We shape data into something easy to read.

But in doing that, we often remove context.

And context is exactly what AI depends on.

Your clean data might actually be hiding things

This one is hard to admit, especially if you've spent months building well-designed models.

Because clean data feels like good data.

You name columns properly. You standardize formats. You build those polished gold layers that everyone trusts.

And for dashboards, that's perfect.

But for AI, it can feel restrictive in ways that aren't obvious at first.

When everything is flattened and predefined, you lose flexibility. You lose the ability to explore sideways. You lose the messy connections that often carry meaning.

A simple table might tell you what happened.

But it rarely tells you why.

It doesn't show how things relate beyond what you've already decided is important. It doesn't leave room for unexpected questions.

And that's usually where AI starts.

It doesn't just follow paths. It looks for them.

Queries are no longer the main interface

We're used to queries being the way we interact with data.

Even if you don't write SQL every day, the idea is familiar. You ask something specific, and the system gives you something specific back.

It's direct. It's controlled.

AI agents don't really work like that.

They don't always look for exact matches. They look for relevance. They try to find meaning across pieces of information that might not be obviously connected.

So instead of strict filters, you get semantic search. Embeddings. Similarity-based retrieval.

Underneath all that technical language is a simple idea.

The system is no longer matching exact values.

It's trying to understand intent.

And here's the problem.

Most pipelines aren't built for that kind of access.

They store data well. They organize it well. But they don't make it easy to search by meaning.

If your data isn't chunked properly, enriched with metadata, or represented in a way that supports semantic lookup, it becomes very hard for an AI agent to use it effectively.

It's there.

But it's not really accessible.

Pipelines optimized for storage don't translate well

A lot of modern data systems are optimized for things that made perfect sense until recently.

Batch processing. Storage efficiency. Structured outputs. Scheduled updates.

These are not bad things. They're actually important.

But they were designed with a certain kind of user in mind.

A human.

Someone who logs in, runs a report, looks at a dashboard, and makes a decision.

AI agents operate differently.

They need quick access to small pieces of data. Not entire tables.

They need connections between sources. Not isolated datasets.

They need freshness. Not data that's a few hours behind.

If your pipeline updates every six hours, a human might not notice.

But an AI agent will build its understanding based on whatever is available at that moment.

If that understanding is outdated, everything that follows can be slightly off.

Not dramatically wrong. Just… off enough to matter.

There's a missing layer most people don't talk about

Meaning.

We talk a lot about data quality, data modeling, and data governance.

But meaning doesn't automatically come from any of those.

You can have perfectly clean data that still doesn't make sense in a broader context.

A record might say user_id equals 123 and event equals click.

That's accurate.

But what does that click represent?

Was it intentional? Was it part of a larger sequence? Does it usually lead to something else?

Humans are surprisingly good at filling in those gaps. We use intuition. We use experience. We connect dots without thinking too much about it.

AI doesn't do that unless you design for it.

And most pipelines don't.

They move data efficiently. They transform it correctly. But they don't carry meaning forward.

That layer is missing.

So what actually needs to change?

Not everything.

You don't need to throw away your current architecture and start from scratch.

But you do need to shift how you think about what you're building.

It's no longer just about preparing data for reporting.

It's about making data usable in a more flexible, exploratory way.

That means keeping more raw data than feels comfortable sometimes. It means adding metadata even when it feels excessive. It means thinking about how data will

be discovered, not just stored.

It also means reducing unnecessary delays where possible.

And maybe most importantly, it means designing systems that allow movement. Movement between datasets, between levels of detail, between related pieces of information.

Because that's how AI works.

It moves.

The part that's easy to ignore

Most modern data stacks look impressive.

You have ingestion layers, transformation layers, orchestration tools, and governance frameworks. Everything is documented. Everything is versioned.

On paper, it all looks solid.

But underneath, many of these systems are still built around a simple assumption.

A human will come in and ask a question.

AI changes that assumption.

Now the system is doing the asking.

It's exploring. It's forming hypotheses. It's pulling in pieces of information from different places and trying to make sense of them.

If your pipeline doesn't support that behavior, it starts to feel rigid.

Even if everything inside it is technically correct.

And that's the tricky part.

You can have a system that is perfectly designed for its original purpose and still not be ready for what's coming next.

Final thought

This isn't really about whether your data is clean.

Clean data still matters. Probably always will.

But it's not the full picture anymore.

The better question is something slightly different.

If an AI agent tried to use this data, would it actually understand anything useful?

Would it be able to move through it, connect ideas, and form something meaningful?

Or would it just see a collection of well-organized pieces with no real sense of how they fit together?

For a lot of pipelines today, the honest answer is somewhere in the middle.

And that's okay.

Because this shift is still happening.

But it does mean one thing.

The way we think about data needs to expand.

Not completely changed. Just expand enough to include a new kind of user.

One that doesn't just read data.

But tries to understand it.

I write about data engineering, architecture, and lessons learned from building real-world data systems.

If you found this helpful, follow for more practical insights on designing scalable data platforms.

AI ✓ Data Engineering + Data Architecture + Analytics +

Big Data +



Follow

Written by Saurav Kumar

1.5K followers · 145 following

Data Engineer | Building clean pipelines, not just code. Writing about modern data tools, Azure, and the real-life side of engineering.



Follow

Published in Predict

114K followers · Last published 15 hours ago

where the future is written