atlan

The Data Catalog Primer for Enterprise AI

Data catalogs are going through a paradigm shift – again. And this time, it could make or break your AI.

This eBook breaks down where data catalogs came from, why they're changing, and where they're going in the age of AI.

Table of Contents

Chapter 4: The rise of the context layer

The next evolution of data catalogs and how they will power AI

- From catalog to context infrastructure
- The new look of AI-ready context infrastructure
- The business impacts of context infrastructure

AI broke the modern data stack – and it's time to move on.

It's no longer just about having the best tools to handle massive amounts of data. Now, data teams need the best tools to make sense of that data. Because as more users – including AI agents – demand access to data, it will need to be accurate, reliable, and valuable in order to deliver value at scale.

The good news? The modern data stack as we know it today is built on a solid foundation – fast, easy to scale up in seconds, and little overhead required.

The bad news? It's becoming obsolete – and amid the shift, many companies are grappling with how to adapt.

AI has upended the way we work with data, making context essential. **If there's no context, there's no AI.**

But as the rest of the data stack has rapidly evolved, catalog-based metadata management has remained stuck in the past. While companies ingest more and more data, catalogs are struggling to keep up. They're becoming more chaotic and less trustworthy.

And now, all that is being amplified by AI.

Companies are rushing to deploy AI agents and systems, only to discover that their AI is only as good as the context it can access. Without rich, reliable metadata, AI hallucinates, misinterprets business logic, and acts on stale assumptions that erode accuracy and reliability. And the impact is clear: **95% of AI pilots fail in production.**

What should metadata look like in today's AI-driven world? How can data catalogs evolve from passive documentation into active context infrastructure that makes AI trustworthy at scale? Why does metadata management need a fundamental shift from its old-school roots to today's demands?

We've spoken with countless data leaders and practitioners to understand how metadata is changing, and bringing data catalogs with it. This helped us develop a vision for the future of data catalogs – one built around context, agility, trust, and collaboration, designed to power the next generation of AI systems.

This primer is meant to give you a look at where we've been, and the forces that are shaping where we're going – so you can make the right catalog and infrastructure decisions to power the future of your data and AI initiatives.

CHAPTER 1

The evolution of data management

Did you know that data catalogs and metadata have been around since ancient times?

Descriptive tags were attached to each scroll in the Library of Alexandria, listing the scroll's title, subject, and author. Callimachus of Cyrene also created The Pinakes, a 120-volume catalog of the entire collection. It was organized by subject with information like author, length, and excerpt for each scroll. This was the first major data catalog!

In 1996, when the internet as we know it was in its infancy, Bill Gates stated that "content is king." At that point – and for the next decade – data was all about aggregation and storage. Where is all the data that we need? How can we ingest it? Where can we store it?

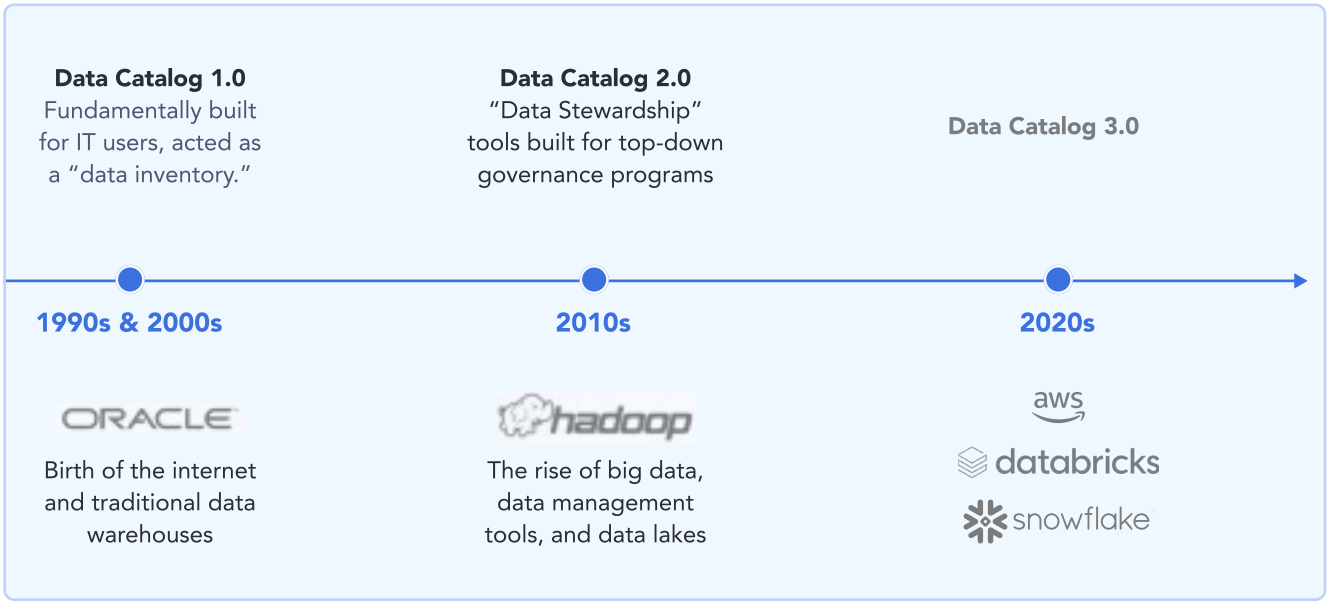
Metadata management has come a long way since then – and today, context is king.

Now, it's all about actually using our data, and empowering others to – even if they're not the traditional "data users." How can we keep data safe and organized? How can we document context for every piece of data? How can we use the right data in the right way? And most urgently, how can we make our data AI-ready?

These questions steer the roadmap for CDOs and CDAOs, because metadata management is the solution, but it's easier said than done.

Data practitioners have been struggling with their metadata for decades.

To understand how we got to our current cataloging crisis, we need to understand how it started. This history can broadly be broken down into three stages.

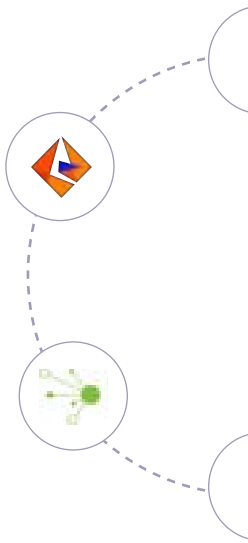


1990s & 2000s: The birth of the internet and traditional data warehouses
Data Catalog 1.0: Fundamentally built for IT users, acted as a "data inventory"

2010s: The rise of big data, data management tools, and data lakes
Data Catalog 2.0: "Data stewardship" tools built for top-down governance programs

2020s: The modern data stack goes mainstream
Data Catalog 3.0: Modern metadata for the modern data stack

2025+: The AI-ready data stack
The next era: Context infrastructure for AI



Data Catalog 1.0: Metadata management by and for IT teams

The birth of the internet and the rise of big data led companies to create an "inventory of data."

1990s – 2000s

In the 1990s, we set aside floppy disks and embraced this newfangled tool called the internet. Guess what that meant? More data.

Data suddenly became accessible to everyone, everywhere (assuming they were on an IT team). Where once people could only access data on a floppy disk or the specific computer where it was stored, now data could be shared across the internet.

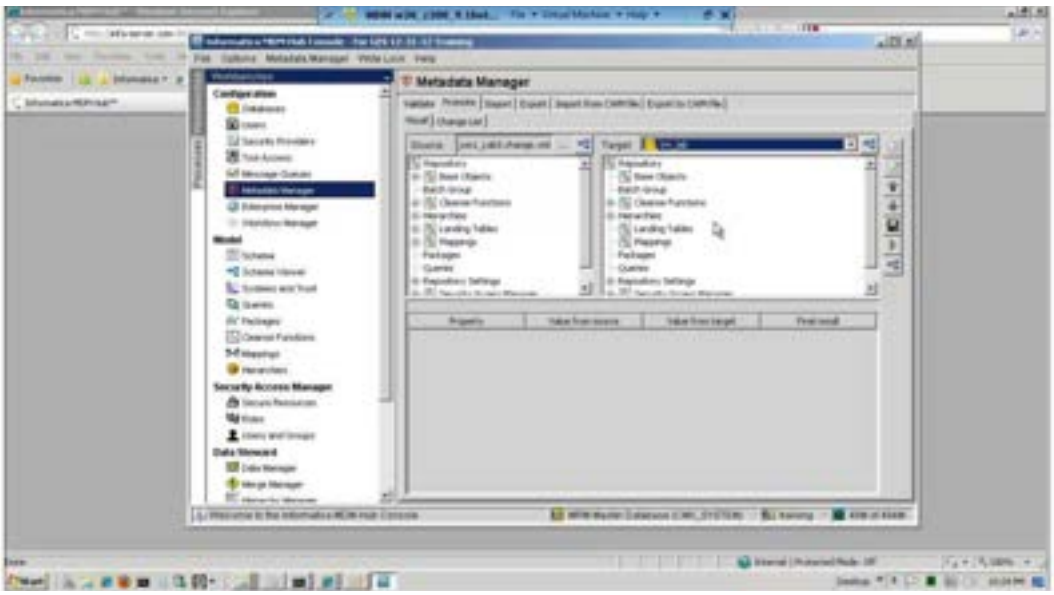
Back in 1997, Michael Lesk estimated that the internet was growing ten-fold each year, and there were already up to 12,000 petabytes (1 PB = 1,000 TB) of information on the internet. As the internet grew, the volume of data started increasing exponentially and we quickly entered the world of big data.

For data teams, the promise of big data quickly turned into a headache. They had plenty of data but no context.

Companies had no idea how to interpret the new deluge of data, much less use it. To stay on top of it, IT teams were tasked with creating data inventories that listed all of a company's stored data, along with its metadata (i.e. key attributes like source, date, and definitions).

Products like Informatica and Talend took an early lead in metadata management, but they didn't prove to be the perfect solutions. Instead, onboarding and maintaining these first-generation data catalogs became a constant struggle for IT teams.

> "Data warehouse teams often spend an enormous amount of time talking about, worrying about, and feeling guilty about metadata. Since most developers have a natural aversion to the development and orderly filing of documentation, metadata often gets cut from the project plan despite everyone's acknowledgement that it is important."
>
> Ralph Kimball, Author and Expert on Data Warehousing



Informatica's Metadata Manager in 2012

Data Catalog 2.0: Data inventories powered by data stewards

The birth of the internet and the rise of big data led companies to create an "inventory of data."

2010s – 2020

Starting around 2010, the data ecosystem became a bit of a mixed bag

The good: Countless people were dipping their toes into the world of data. It became more mainstream and spread beyond the cloistered IT teams.

The bad: with cloud data warehouses, data lakes, and big data technologies like Hadoop, data infrastructure was extremely complex. And so, despite technological advances, companies struggled to find and understand their data.

This is when the idea of "data stewardship" took hold. Data stewards were the people dedicated to and responsible for taking care of an organization's data. They would handle metadata, maintain governance practices, manually document data, and so on.

At the same time, the idea of metadata shifted.

As companies started setting up massive Hadoop implementations, they realized that a simple IT inventory of data wasn't enough anymore. To get value, they needed to blend data inventory with business context.

The second-generation data catalog was rooted in these new ideas of data stewardship and context-driven metadata. Rather than just inventorying data, teams sought to finally create a single source of truth.

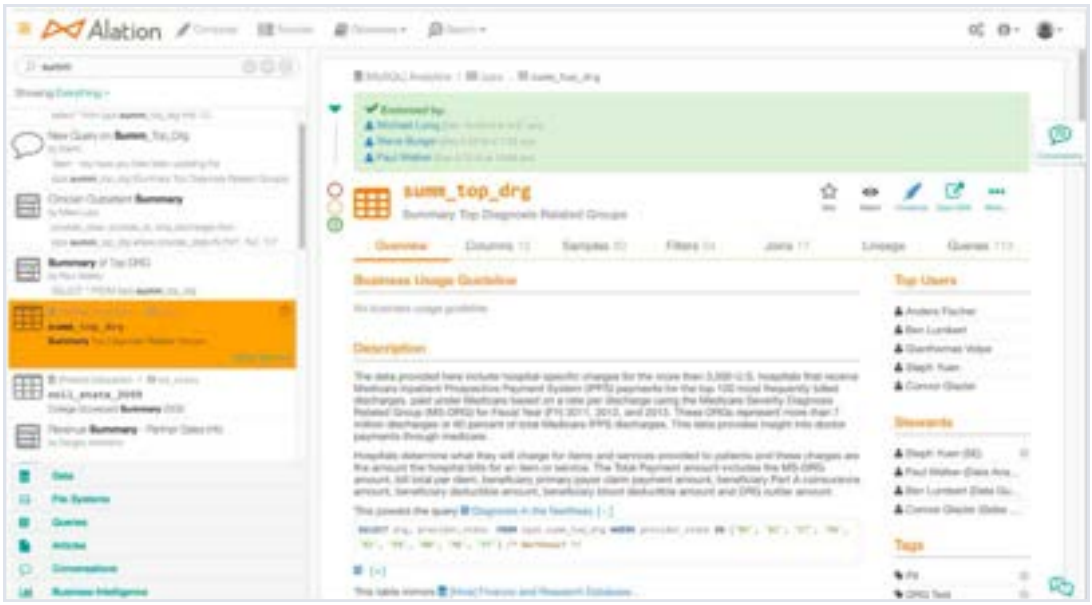
At least, that was the plan…

In this era, data catalogs like Collibra and Alation were built on monolithic architectures and deployed on-premises. Each data system had its own installation, and companies couldn't roll out software changes by simply pushing a cloud update.

As a result, these data catalogs proved difficult to set up and maintain. They involved rigid data governance committees, formal data steward roles, complex technology setups, and lengthy implementation cycles.

All in all, the installation process could take up to 18 months. (We've heard of projects that were even more drawn out, like one that took five years!)

In the end, technical debt grew and metadata management steadily fell behind the rest of the modern data stack.



Alation's Data Catalog in 2019

CHAPTER 2

The problem with traditional data catalogs

Traditional began data catalogs falling short in the modern data stack.

While data infrastructure evolved, metadata management didn't. Traditional data catalogs were built for on-prem data warehouses and the Hadoop era. But as the modern data stack quickly evolved, they fell behind.

Second-generation data catalogs – which many companies still rely on – are:

- Built for on-premises systems, rather than the cloud
- Designed for top-down IT teams, rather than diverse, collaborative humans of data
- Opaque "one-size-fits-all" pricing, rather than flexible pay-as-you-go models
- Stuck with high support and maintenance overheads, rather than quick-to-start and self-maintaining
- Locked into specific vendors and platforms, rather than adopting open standards
- Fundamentally passive, rather than active metadata platforms

To fill this gap, most early adopters of the modern data stack built internal tools for metadata management. Companies like LinkedIn (DataHub), Facebook (Nemo), Lyft (Amundsen), Netflix (Metacat), Uber (Databook), and Airbnb (Dataportal) all created their own solutions.

However, this isn't feasible for all companies, not to mention that it's inefficient to stand up dozens of similar tools.

Industry realities demanded a modern metadata solution – one that is just as fast, flexible, and scalable as the rest of the modern data stack. And crucially, one that's purpose-built to power AI at scale.

The new world of modern metadata management

The rise of third-generation data catalogs began in 2020 – but it didn't start out of the blue.

Instead, just like the birth of the internet, big data, and data stewardship triggered the creation of first- and second-generation data catalogs, third-generation catalogs started as a response to major changes in the data ecosystem.

The modern data stack went mainstream, with a full range of unprecedentedly fast, flexible, cloud-native tools.

Data teams became more diverse than ever, leading to chaos and collaboration overhead.

Teams started reimagining data governance, from top-down, centralized rules to bottom-up, decentralized initiatives.

Passive metadata systems started being scrapped in favor of active metadata platforms.

As metadata became big data, the metadata lake emerged to power use cases like data discovery, lineage, observability, meshes, and more.

And now, AI is demanding context at scale. Without metadata, AI systems hallucinate, misinterpret, and fail to deliver value in production.

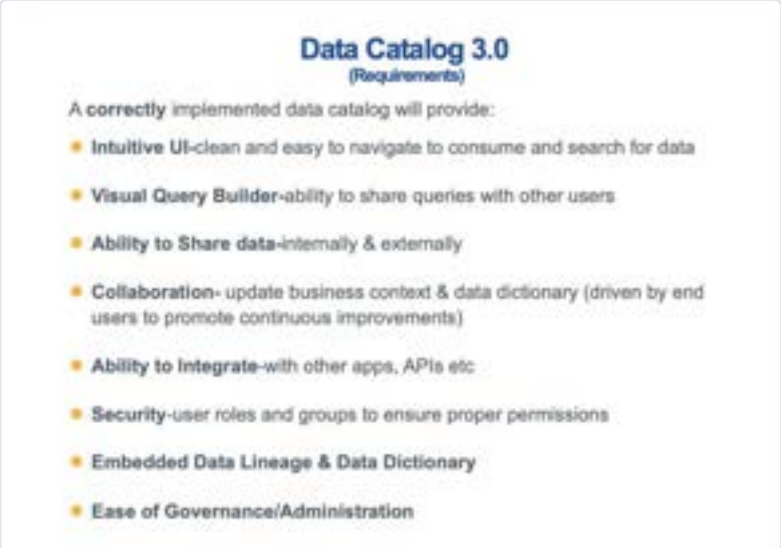
The backstory of "Data Catalog 3.0"

Is this the first time you've heard about "Data Catalog 3.0" or "third-generation data catalogs"?

One of the first instances of this idea was in December 2020, when Shirshanka Das wrote about the three generations of metadata architectures. A month later, we wrote about the three generations of data catalogs: Data Catalog 1.0, 2.0, and 3.0.

The terminology exploded. We even received RFPs that specifically asked for a Data Catalog 3.0!

And in 2025, Gartner re-released its Magic Quadrant for Metadata Management Solutions after a five-year hiatus – signaling a fundamental shift in the industry. As Gartner put it, **"Metadata is now the indispensable foundation for modern and agentic AI systems."**



Data Catalog 3.0
(Requirements)

A **correctly** implemented data catalog will provide:

- **Intuitive UI**-clean and easy to navigate to consume and search for data
- **Visual Query Builder**-ability to share queries with other users
- **Ability to Share data**-internally & externally
- **Collaboration**- update business context & data dictionary (driven by end users to promote continuous improvements)
- **Ability to Integrate**-with other apps, APIs etc
- **Security**-user roles and groups to ensure proper permissions
- **Embedded Data Lineage & Data Dictionary**
- **Ease of Governance/Administration**

Snippet of an anonymized RFP for a Data Catalog 3.0

Key characteristics of the modern data stack

Cloud-first — Elastic, scalable services integrate natively with public clouds and managed infrastructure.

Warehouse/data lake–centric — Architectures are designed around a cloud data warehouse or lakehouse as the analytical source of truth.

Composable, single-purpose tools — Each tool solves a focused problem and snaps into the pipeline without heavy rewrites

SaaS or open-core delivery — Managed services (and/or open-core options) reduce operations overhead and time to value.

Low entry barrier — Usage-based pricing, low/no-code setup, and fast onboarding accelerate adoption across teams.

Community-driven — Vibrant practitioner communities, accessible documentation, and extensibility fuel best practices and faster iteration.

Metadata-first and interoperable — Metadata is the "glue" across ingestion, transformation, BI, and ops; open APIs and active metadata connect the stack end-to-end.

AI-ready and automation-friendly — Built to support advanced analytics/ML and orchestration at scale, enabling agentic and human-in-the-loop use cases.

The five trends driving third-gen data catalogs

Why do third-generation data catalogs exist? **It comes down to five key trends.**

1. The creation of the modern data stack

Starting around 2016, the modern data stack went mainstream. Organizations began adopting flexible collections of tools and capabilities to store, manage, and use their data. These tools are unified by three key ideas:

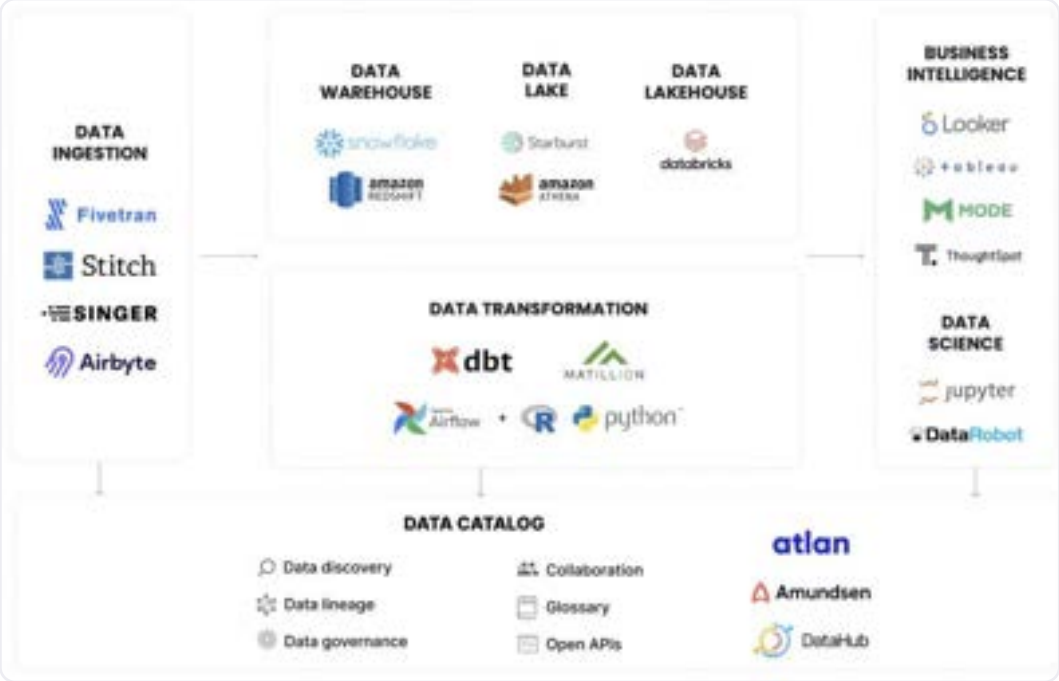
- Self-service for a diverse range of users
- Agile data management
- Cloud-first and cloud-native

We embraced the modern data stack for its ability to set up easily, pay as you go, and plug and play with existing platforms. With tools like Fivetran and Snowflake, users became able to set up a data warehouse in less than 30 minutes.

But as the data stack evolved, traditional data catalogs couldn't keep up. Even second-generation catalogs require significant setup time and at least five calls to get a demo. This necessitated a truly modern metadata solution that was just as fast, flexible, and scalable as the rest of the modern data stack.

Learn more about the modern data stack

Read more



2. The diverse "humans of data"

A few years ago, only the IT team would get their hands dirty with data.

As the modern data stack evolved, data teams became more diverse than ever before. They now include data engineers, analysts, analytics engineers, data scientists, product managers, business analysts, citizen data scientists, and more. Each of these people has their own favorite – and equally diverse – data tools, from SQL, Looker, and Jupyter to Python, Tableau, dbt, and R.

This diversity is both a strength and a struggle.

All of these users also rely on different tools, skill sets, tech stacks, workflows, and ways of approaching a problem. Essentially, they each have a unique data DNA.

More diverse perspectives mean more opportunities for creative solutions and out-of-the-box thinking. However, it also means more room for chaos within collaboration.

In this world, self-service is no longer optional. Data tools need to be intuitive for a wide range of users with a wide range of technical expertise. If someone wants to bring data into their work, they should be able to easily find the data they need without having to ask an analyst or file a request that may take days or weeks to fulfill.

Learn more about the humans of data

Read more

3. The new vision for data governance

Data governance was traditionally seen as a bureaucratic, restrictive process – a set of rules pushed down "from the top" that slows down work. That perspective persists, and the reality is, that's often how it actually works. Companies surround their data with complex security processes and restrictions, all dictated by a distant data governance team.

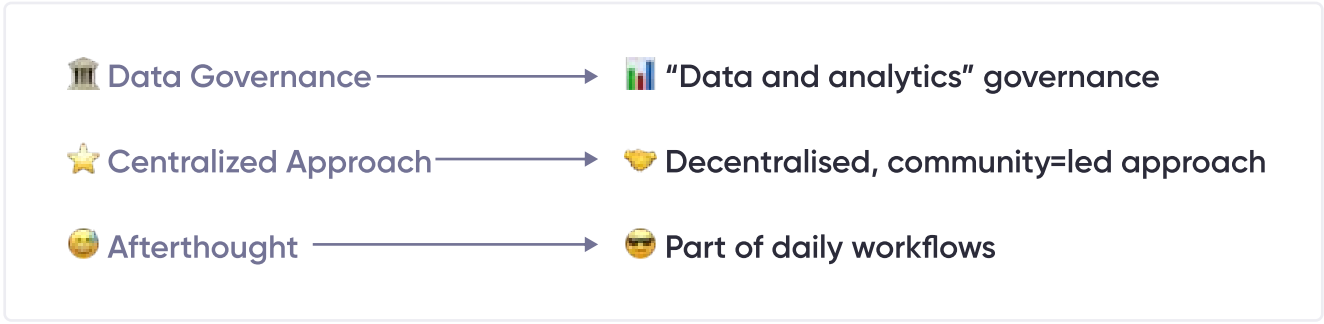
As the modern data stack made it easier to ingest and transform data, this idea of data governance became a major barrier. For the first time, the need for governance was felt bottom-up by practitioners, instead of being enforced top-down due to regulation.

Data governance is undergoing a paradigm shift.

It's becoming something that the humans of data embrace rather than fear. At its heart, governance is now less about control, and more about helping data teams work better together.

As a result, data governance is being reimagined as a set of collaborative best practices by and for amazing data teams – ones focused on building and empowering better data teams, not controlling them.

Modern, community-led governance needs a new type of data catalog, one built around bottom-up collaboration rather than outdated, top-down, steward-based data management processes.



> "Governance is a product area whose time has come.... This problem has only been made more painful by the modern data stack to-date, since it has become increasingly easy to ingest, model, and analyze more data. Without good governance, more data = more chaos = less trust."
>
> Tristan Handy, CEO & Founder, dbt

4. The rise of the metadata lake

In 2005, more data was being collected than ever before, with more ways to use it than a single project or team could dream of. Data had limitless potential, but setting up a data system for limitless use cases was a nonstarter. That led to the birth of the data lake.

Metadata became big data, and technical advances (i.e. elasticity) in compute engines like Snowflake and Redshift made it possible to derive intelligence from metadata in a way that was unimaginable even a few years prior.

As metadata increased, and the intelligence we could derive from it increased, so too did its use cases.

Even today, the most data-driven organizations have only scratched the surface of what is possible with metadata. However, it holds the key to unlocking future-ready data systems — powering not just the use cases of today like data cataloging, lineage, and observability, but also emerging use cases like auto-tuning data pipelines, auto-scaling compute engines, and more.

The metadata lake is what makes this possible.

The metadata lake is a unified repository that can store all kinds of metadata, in both raw and further processed forms, in a way that can be shared with other tools in the data stack to drive a wide variety of use cases.

Key characteristics of the metadata lake

Open APIs and interfaces: The metadata lake needs to be easily accessible, not just as a data store but via open APIs. In any metadata solution, the fundamental metastore should be open and usable, allowing teams to draw on it as a "single source of truth" for a wide variety of future use cases and applications.

Built on a knowledge graph: The metadata lake is most powerful when the connections between data assets come alive. For example, if one column is tagged as "confidential," metadata and lineage can be used to tag all derived columns as "confidential." The knowledge graph is the best way for these interconnections to be stored.

Power humans & machines: The metadata lake can improve both humans' daily work (such as discovering data and understanding its context) and machines or tools' daily workflows (such as auto-tuning data pipelines). This means that metadata lakes need to be fundamentally flexible and adaptable.

5. The birth of active metadata

In August 2021, Gartner scrapped its Magic Quadrant for Metadata Management and replaced it with the Market Guide for Active Metadata Management. This marked the end of the traditional approach to metadata management and kicked off a new focus on active metadata.

In 2025, Gartner brought back the Magic Quadrant for Metadata Management Solutions – a clear signal of category maturation and recognition that metadata management has evolved from passive documentation into the foundational infrastructure for enterprise AI.

Read the report here

Gartner

Organizations leveraging traditional catalogs had to rely on human effort to curate and document data. This was unsustainable. Active metadata platforms emerged in response, offering an always-on, intelligence-driven, and action-oriented alternative. These systems are:

Always on: Rather than waiting gfor humans to manually enter metadata, they continuously collect metadata from logs, query history, usage stats, etc.

Intelligence-driven: They constantly process metadata to connect the dots and create intelligence, such as automatically creating lineage by parsing through query logs.

Action-oriented: Instead of being passive observers, these systems drive recommendations, generate alerts, and operationalize intelligence in real time.

> **"Metadata management is shifting from augmented data catalogs to metadata-anywhere orchestration platforms — the driving layer for AI readiness and agentic AI."**
>
> Gartner, 2025 Magic Quadrant for Metadata Management Solutions

Gartner

CHAPTER 3

The evolution of Data Catalog 3.0

What are third-generation data catalogs?

Third-generation data catalogs don't look and feel like their old-school predecessors.

The shift from slow, on-prem Data Catalog 2.0 to the era of Data Catalog 3.0 ushered in a generation that felt more like the collaborative, self-service tools in modern workplaces – think GitHub, Figma, Slack, Notion, and Superhuman.

The 4 pillars of third-generation data catalogs

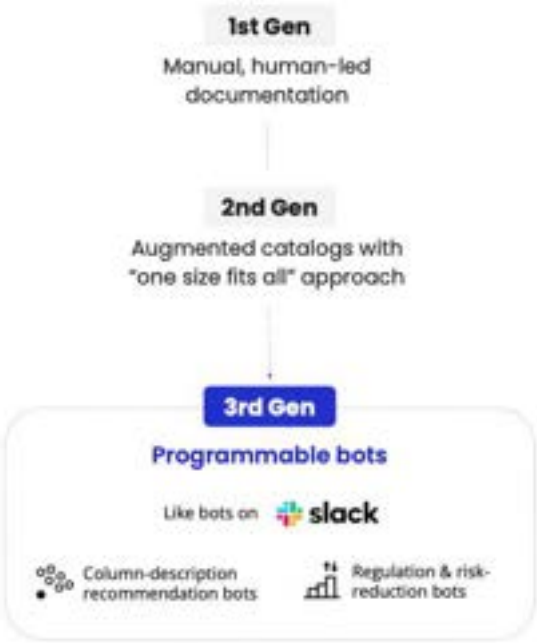
The fundamental principles behind the third generation of data catalogs.

1. Programmable bots

Augmented data catalogs, which use machine learning to automate a set of manual tasks, have become increasingly popular – but they're still not a silver bullet.

That's because every company is unique. Every industry is unique. Every individual team's data is unique. So no matter how good it is, no ML algorithm can magically create context, identify anomalies, and anticipate needs for every industry, company, and use case.

That's why third-generation tools rely instead on programmable bots, a framework that lets teams create their own machine learning or data science algorithms. These could include customized versions of "out-of-the-box" bots, or bots programmed from scratch by a company's data team.



Examples of programmable bots in action:

Security & compliance bots: Increasing regulation and the global nature of business operations means companies will have to follow more rules. Custom security bots can be used to identify and tag sensitive columns based on the regulations that apply to each company.

Example: In India, there's a government ID card called the "PAN card," which should be classified as PII data. A generic PII bot won't catch it, but a custom PAN classification bot can be built to correctly find and identify PAN card data.

Classification bots: Companies with specific naming conventions for their data sets can create bots to automatically organize, classify, and tag their data ecosystem based on preset rules.

Example: Companies can create their own bots based on CIA (Confidentiality, Integrity, Availability) ratings for information security.

Data observability bots: Companies can take out-of-the-box observability algorithms, and customize them to their data ecosystems and use cases.

Example: A financial firm customize a data freshness bot to flag when transaction data hasn't been updated within 15 minutes during trading hours, but allow hourly updates during off-hours, reflecting their specific business operations rather than using a generic rule.

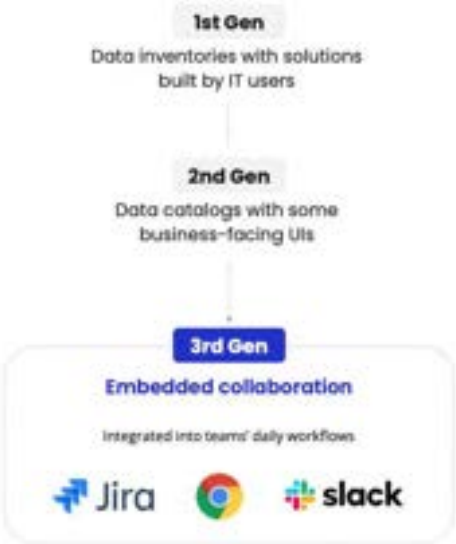
2. Embedded collaboration

With more diverse data teams, data tools need to be designed to integrate seamlessly with their daily workflows. This is where the idea of embedded collaboration comes alive.

Embedded collaboration allows users to work where they already are, with minimal friction.

- What if you could request access to a data asset when you get a link, just like with Google Docs?
- What if the owner could approve or reject your request without leaving Slack?
- What if you could trigger a support request on Jira without leaving a data asset?

Embedded collaboration unifies these micro-workflows that waste time, cause frustration, and lead to tool fatigue for data teams, and instead make these tasks delightful.



Examples of embedded collaboration in action:

Reverse metadata: Make context available in daily tools, embedded directly into BI dashboards, query editors, and the applications where people work.

Contextual discussions: Bring context into daily workflows through Slack notifications, team channels, and communication tools where decisions are already being made.

3. End-to-end visibility

Tools from the Data Catalog 2.0 era made significant strides in improving data discovery. However, they didn't give organizations a "single source of truth" for their data. We're still dealing with frantic calls to engineers when an important dashboard breaks, the "Why is there a v2_final_final.csv and v3_final.csv?!" questions, and the gut-wrenching "This data doesn't look right…" emails from the boss.

That's because today, context is spread across different places, and often exists in silos. To get the full picture, users need to ask multiple people and check information across multiple tools. And AI agents won't know to do this – they'll simply rely on the first answer they can find.

With complete visibility into every data asset, third-generation data catalogs help teams finally achieve the holy grail: **a single source of truth about every data asset in the organization.**

This end-to-end visibility includes opening up information like:

- Who owns a dataset, auto-generated from query history
- Where it comes from, from automated lineage
- Whether it's trustworthy, from quality scores and how recently it was updated
- Which columns are used most, and how people use them
- And most importantly, a preview of the data itself!

All of this is available in one seamless experience – no more constant toggling between all the tools in your modern data stack.

Examples of end-to-end visibility in action:

KNOWLEDGE & CONTEXT	TRUST & VISIBILITY
• Owners & experts • Readmes & documentation • Business terms & glossary • Tags & classifications • Data dictionary • Custom metadata (e.g. freshness)	• Column-level lineage • Usage ranking • Custom data quality metrics • Query sharing • Data preview • Data query

Think of it like this: Just as code has a profile on GitHub and customers have a profile on Salesforce, every data asset should have a complete profile that tells you everything you need to know at a glance.

4. Open by default

We've talked about how the modern data stack is built on open standards and tools. But why does that matter?

We're also fast approaching a world where metadata itself will be big data. Integrating this "big metadata" with the rest of the tools in the data stack will help teams understand and trust their data more.

Metadata will also be the key to unlocking new superpowers in the modern data stack, such as auto-tuning pipelines based on demand, automatically creating column-level lineage from SQL code in query logs, or training AI systems to be autonomous and accurate.

For instance, query logs are just one kind of metadata available today. By parsing through the SQL code from query logs in Snowflake, it's possible to automatically create column-level lineage, assign a popularity score to every data asset, and even deduce the potential owners and experts for each asset.

Rather than being closed off and isolated, data catalogs need to be open by default to power this innovation.

By connecting to all other parts of the modern data stack, data catalogs will go from passively storing metadata to actively improving data teams' daily work.

Examples of "open by default" in action

Custom plugins for data asset integrations: MongoDB for new data asset syncs

---

Embedded collaboration plugins: Jira for issue management

Programmable bots: Classifying CIA ratings (Confidentiality, Integrity, Availability)

Metadata Plugin Marketplace: Design plugins on Figma or code sharing on npm

CHAPTER 4

The rise of the context layer

Why AI demands a new evolution in metadata management

Third-generation data catalogs solved critical problems: they made metadata fast, collaborative, comprehensive, and open. They brought metadata management into the modern data stack era with programmable bots, embedded collaboration, end-to-end visibility, and open APIs.

But the emergence of AI as a fundamental force in how organizations work with data has revealed a gap that even the best third-generation catalogs weren't built to fill.

AI systems don't just need to find data — they need to understand it.

When a human analyst searches for "revenue," they can look at a few results, ask a colleague for clarification, and figure out which dataset is the right one for their analysis. They can work around incomplete metadata, patch together context from multiple sources, and make judgment calls about data quality.

AI agents can't do any of that.

When AI lacks context, it fails in predictable and often costly ways. It hallucinates answers that sound plausible but are wrong. It misinterprets business logic, using Finance's definition of "customer" when Product's definition was needed. It violates governance rules by accessing data it shouldn't touch. It acts on stale assumptions, making decisions based on outdated information.

"We deployed things on top of our agent, and it couldn't answer one question," said Joe DosSantos, VP of Enterprise Data and Analytics at Workday. "We started to realize that we were missing this translation layer. We had no way to interpret the human language against the structure of the data."

This is the AI context gap — and it's why we're seeing Data Catalog 3.0 evolve into something more fundamental: **the AI-era context layer.** It's not a new generation, but a new shift – you can think of it as Data Catalog 3.0.1.

> **"Data governance has outgrown its compliance roots: In today's AI-fueled and data-saturated enterprise, it's the control plane for trust, agility, and scale. The next frontier isn't about managing metadata – it's about activating it."**
>
> The Forrester Wave: Data Governance Solutions, Q3 2025
>
> Read the report here
>
> FORRESTER®

From catalog to context infrastructure

We're seeing a shift from traditional metadata management to context infrastructure — treating metadata not as documentation or even as an asset to be managed, but as the foundational layer that makes AI systems work.

This evolution is built on three core architectural pillars:

Active metadata management

Traditional catalogs were passive — they waited for humans to document data and update metadata manually. Third-generation catalogs improved on this with automation and intelligence, but now we're moving even further ahead.

Active metadata platforms are always-on systems that continuously collect, process, and operationalize metadata in real time. They pull metadata from logs, query history, usage patterns, and AI interactions automatically. They connect the dots to create intelligence — building lineage by parsing SQL, identifying experts through usage patterns, detecting quality issues before they break downstream systems.

Most importantly, they don't just observe — they act. Active metadata platforms drive recommendations, generate alerts, enforce governance policies, and deliver context to AI systems the moment it's needed.

The context layer

While traditional catalogs focused on helping humans discover data, the context layer is designed to deliver understanding to AI systems at the point of action.

It provides:

- **Semantic understanding** — What does "revenue" mean in this business context? How is "active user" defined across different teams?

- **Operational intelligence** — Is this data fresh? Who owns it? What breaks if it changes?
- **Governance at runtime** — What can this AI agent access? What actions are permitted? What policies apply?
- **Lineage and trust signals** — Where did this data come from? How has it been transformed? Can we trust it for this use case?

The context layer goes beyond storage by orchestrating metadata across the entire data estate. This ensures that every AI system, human user, and downstream tool has access to the same, continuously updated understanding of what data means and how to use it correctly.

The metadata lakehouse

To serve as infrastructure for AI, metadata itself needs to be treated as data — queryable, versionable, and analytics-ready at scale.

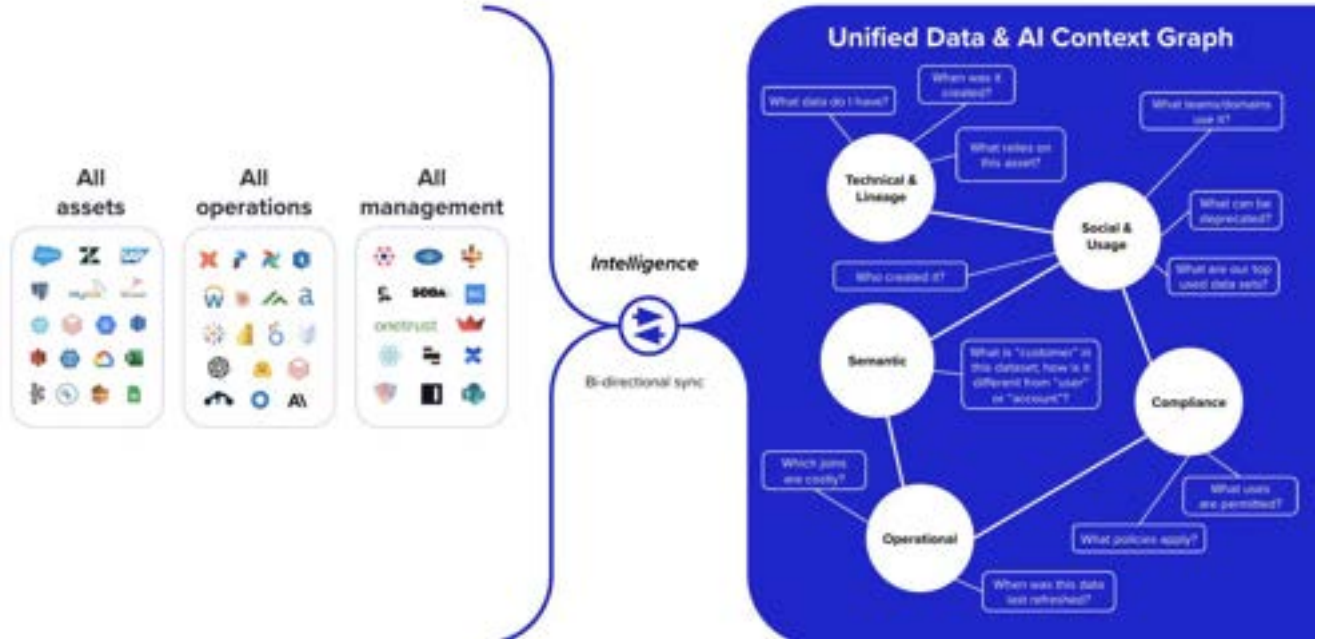
This is where the metadata lakehouse comes in. Built on open table formats like Apache Iceberg, the metadata lakehouse provides:

- **ACID transactions** for metadata, ensuring consistency across concurrent reads and writes
- **Schema evolution** so metadata structures can adapt as organizations change
- **Time travel and versioning** for audit trails and understanding how context evolved
- **Open, analytics-ready storage** that allows teams to bring their own compute and query metadata like any other dataset
- **Knowledge graph** foundation that captures relationships between assets and enables intelligent propagation of context

This turns metadata from a lookup system into a context store that can power everything from AI agents querying semantic definitions to compliance dashboards analyzing governance coverage across billions of assets.

What makes context infrastructure different from traditional catalogs?

Context infrastructure goes beyond documentation, discovery, and configurable workflows. It provides the semantic understanding, governance rules, and operational metadata that AI systems need to act autonomously and correctly. It's what takes AI from being able to understand language to being able to understand your business – so it can give grounded answers, make better decisions, and deliver real value.



Gartner predicts that leveraging metadata analytics will allow companies to deliver new data assets up to 70% faster in the coming years. But to get there, they'll need to uplevel their tooling.

The new look of AI-ready context infrastructure

To power AI at scale, context infrastructure requires these foundational capabilities:

1. Context extraction — Connectors that continuously mine metadata and knowledge from catalogs, BI tools, docs, CRMs, and code to bootstrap shared context at scale.

2. Semantic unification — Normalization of definitions, relationships, metrics, and domains into a common semantic layer using open standards so context travels across tools.

3. Context products — Curated, reusable bundles of datasets, definitions, queries, and rules validated with test questions to ensure minimum viable context.

4. Context store — An open, analytics-ready substrate (like an Iceberg-native Metadata Lakehouse) that persists context as queryable tables, treating metadata as data.

5. Retrieval and delivery APIs — Low-latency services that deliver context to humans and AI agents in real time, grounding conversational analytics and agentic workflows.

6. Embedded governance — First-class policy, permissions, lineage, and audit capabilities so context is safe to activate in production – enforced and observable, not just documented.

7. Intelligent automation — Programmable bots and rules that auto-classify, tag, propagate policies, and trigger actions based on events and workflow changes.

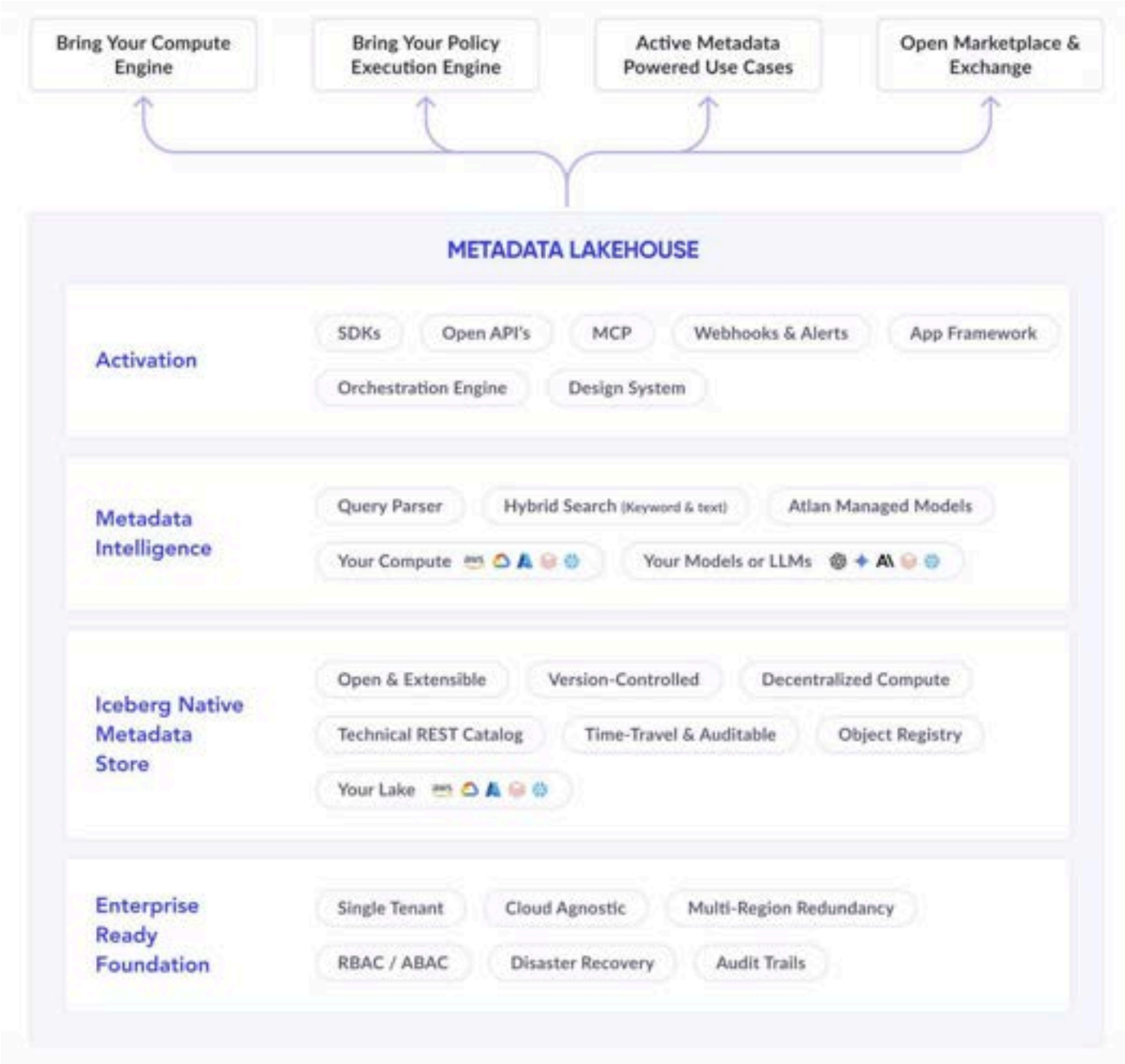
8. Feedback loops — Human-in-the-loop correction, telemetry, and evaluations to keep context accurate and continuously improve it over time.

9. Performance at inference speed — High-throughput, low-latency architecture designed for agents to fetch and apply context in milliseconds at scale.

10. Versioning and time awareness — Built-in history and change tracking for explainability, audits, and trend analysis on context enrichment and adoption.

The companies that will succeed with AI aren't necessarily those with the most sophisticated models. They're the ones that have built the context infrastructure that makes those models trustworthy, accurate, and usable at scale.

This shift is the foundation for enterprise AI that delivers value — not headaches — in production.



Key characteristics of the metadata lakehouse

Iceberg-Native Open Storage — Built on Apache Iceberg to provide ACID transactions, schema evolution, time travel, and versioning for metadata, enabling it to function like a database while remaining open and interoperable.

Unified Context Repository — Consolidates all types of metadata (technical, business, operational, social, and AI-specific) in a single high-performance repository, eliminating silos and providing a holistic view of the enterprise data estate.

Open APIs and Compute Flexibility — Enables organizations to "bring your own compute" and query metadata using their preferred engines (Snowflake, Databricks, Spark), avoiding vendor lock-in through standardized APIs and open architecture.

Knowledge Graph Foundation — Uses a graph structure to capture relationships between data assets and enable intelligent context propagation, such as automatically tagging derived columns when source columns are classified.

Real-Time Event Streaming — Continuously collects and synchronizes metadata across the data estate through event-driven updates, ensuring context is always fresh and available when AI systems or users need it.

Scalability for Billions of Assets — Designed to handle enterprise-scale metadata by decoupling storage from compute, allowing independent scaling while maintaining fast performance for both analytical queries and operational lookups.

AI-Native Design — Architected specifically to serve as the context layer for AI systems, capturing AI-specific metadata (prompts, model lineage, usage patterns) and delivering semantic understanding to agents in real time.

Embedded Governance — Integrates governance directly into the architecture rather than treating it as a separate layer, with policy enforcement at query time, lineage-based tag propagation, built-in audit trails, and automatic access controls.

The business impacts of context infrastructure

When metadata becomes context infrastructure, it unlocks capabilities that weren't possible with traditional catalogs:

Agentic AI that acts autonomously: AI agents can query data, understand business context, and take actions without constant human oversight. Being grounded in rich metadata tells them not just what data exists, but what it means and how to use it correctly.

5x improvement in AI accuracy: In Atlan AI Labs experiments with customers, embedding metadata context and context supply chains in talk-to-data applications increased accuracy by over 5x compared to standard retrieval approaches.

70% faster time to value: Gartner estimates that by 2027, companies that broadly leverage metadata analytics will deliver new data assets up to 70% faster – a critical advantage when AI initiatives are measured in weeks, not quarters.

Trustworthy AI at scale: Context infrastructure provides the governance, lineage, and semantic understanding that makes AI outputs explainable, auditable, and trustworthy — the requirements for moving AI from pilots to production.

The journey from Data Catalog 1.0 to 3.0 to metadata infrastructure isn’t just about better technology – it’s about a fundamental shift in how we think about metadata, from passive documentation to active infrastructure that powers the next generation of AI systems.

MEET ATLAN

The context layer for modern data teams

Cloud-native metadata orchestration platform

In 2025, Atlan was named a Leader in Gartner's Magic Quadrant for Metadata Management Solutions – the only vendor to score above average in all five category use cases, with top rankings in AI Readiness, Data Engineering, and Data Governance.

> **"Atlan's vision of being the metadata control plane to capture, unify, and understand enterprises' data estates is central to supporting all consumption and AI use cases and providing the necessary context and data for agentic solutions."**

Why leading enterprises choose Atlan

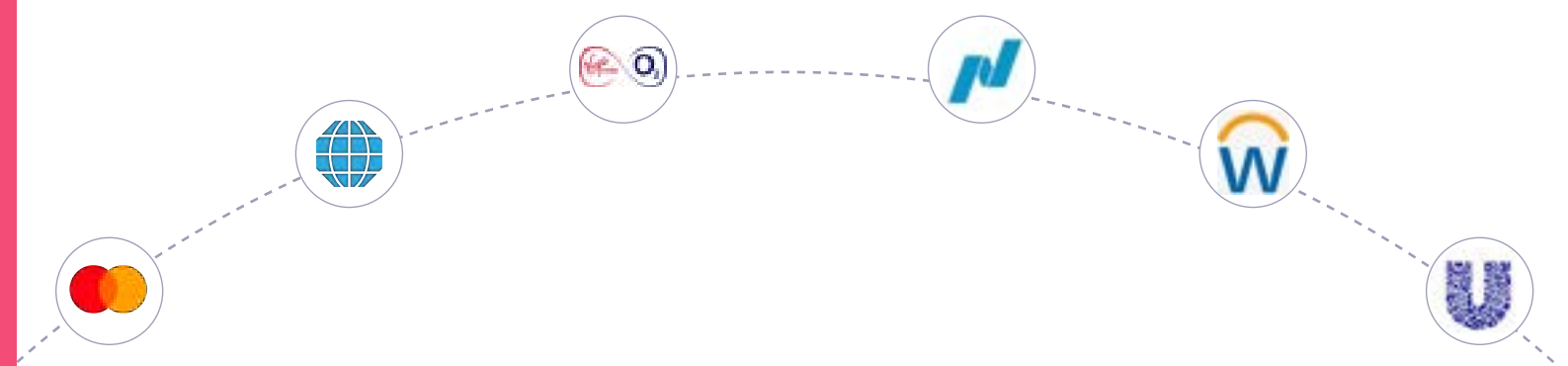
Built for AI from day one: Atlan's Metadata Lakehouse provides an Iceberg-native store, real-time event streaming, and a knowledge graph for semantic understanding – creating the unified context layer that makes AI trustworthy at scale.

Fastest time to value: Setup and onboarding takes weeks, not months – with no armies of consultants, no rigid governance committees required.

Open and extensible: Open APIs, plug-and-play integrations, and a thriving ecosystem of connectors mean Atlan works with your existing tools and adapts as your stack evolves.

Context everywhere: Embedded collaboration brings metadata into Slack, Jira, Chrome, and BI tools, so context lives where your teams work – not in a separate catalog they have to remember to check.

Trusted by data leaders across industries



Recognition from industry leaders

Leader in the Gartner 2025 Magic Quadrant for Metadata Management Solutions

Backed by marquee investors:
Insight Partners, Sequoia Capital, and the top founders & CEOs pioneering the modern data stack, including leaders from Snowflake, Looker, Stitch, and DataRobot.

What sets Atlan apart

Democratization for businesses: Intuitive interfaces and embedded collaboration that make metadata accessible to everyone, not just data engineers.

Governance for IT teams: Real-time policy enforcement, automated lineage, and compliance controls that scale with your AI initiatives.

AI-ready architecture: Purpose-built for the age of agentic AI, with metadata orchestration that provides the context layer AI systems need to be accurate and trustworthy.

Ready to build context infrastructure for your AI future?

Talk to sales →

Take a Product Tour →

atlan