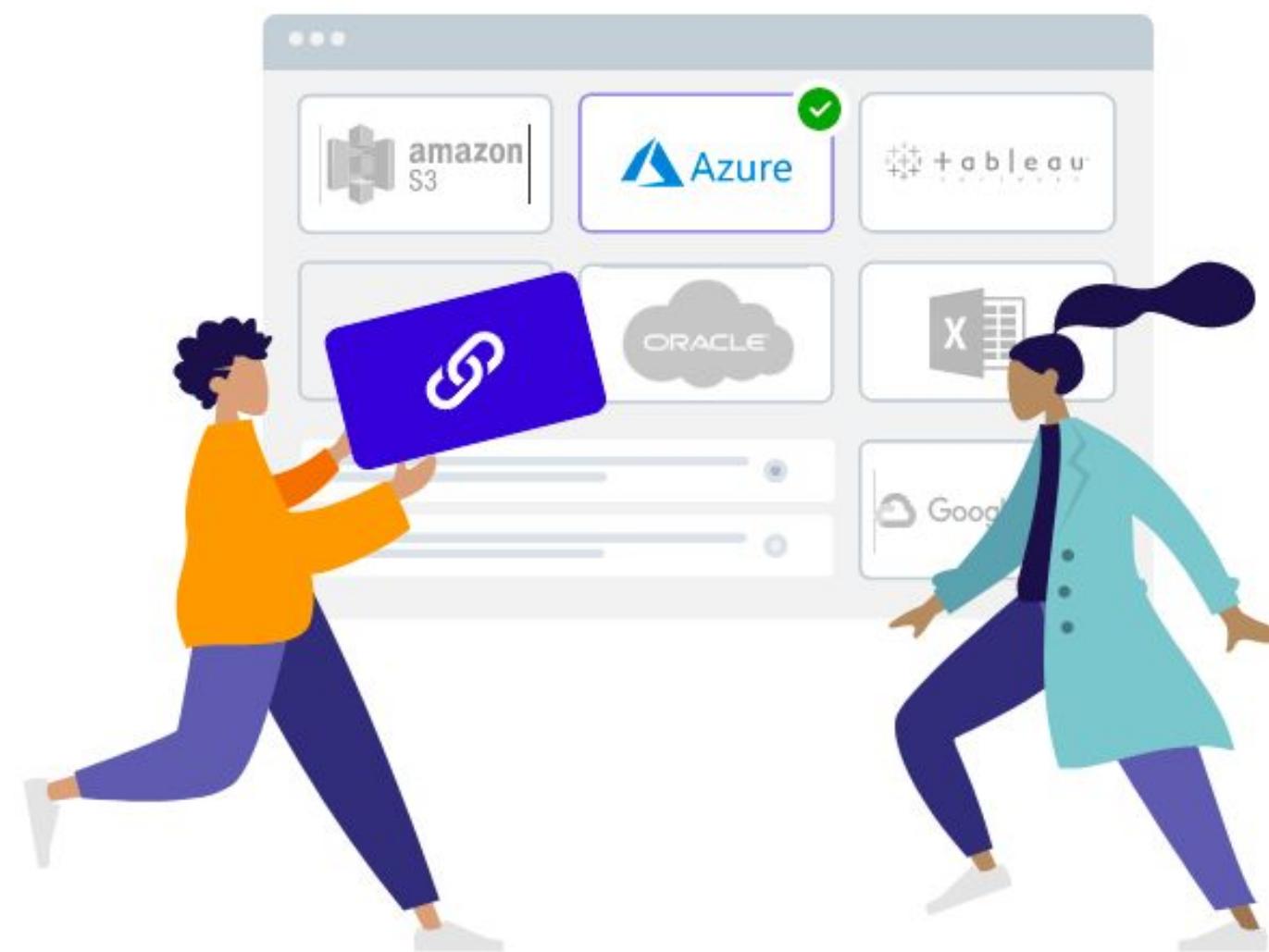


The Third-Generation Data Catalog Primer

Rise of the Active Metadata Platform



Data catalogs are going through a paradigm shift.

This ebook breaks down where they came from, why they're changing, and where they're going.

The data world has recently converged around the best set of tools for dealing with massive amounts of data, aka the “**modern data stack**”.

The good? The modern data stack is super fast, easy to scale up in seconds, and requires little overhead. The bad? It’s still a noob in terms of bringing governance, trust, and context to data.

That's where metadata comes in. 

However, as the rest of the data stack has accelerated, metadata management is stuck in the past. **While companies ingest more and more data, data catalogs are struggling to keep up. The result is chaos and mistrust.**

So what should modern metadata look like in today's modern data stack? How can data catalogs evolve into a powerful vehicle for data democratization and governance? Why does metadata management need a fundamental shift from its old-school roots to today's modern demands?

We spoke to over 350 data leaders to understand their struggles with metadata. This helped us construct a vision for the future of data catalogs — one built around agility, trust, and collaboration. **We call this “Data Catalog 3.0” or the “Third-Generation Data Catalog”.**

Table of Contents

Chapter 1: The evolution of metadata management	1
<i>A brief history of how metadata management has changed since 1990</i>	
The evolution of metadata management	2
Data Catalog 1.0: Metadata management by and for IT teams	3
Data Catalog 2.0: Data inventories powered by data stewards	5
Chapter 2: The problem with traditional data catalogs	7
<i>Why data teams today are searching for a better way to manage their data</i>	
The new world of modern metadata management	8
The five trends driving third-gen data catalogs	9
Modern metadata for the modern data stack	17
Chapter 3: The era of Data Catalog 3.0	18
<i>The objectives, principles, and examples of third-generation data catalogs in today's modern data stack</i>	
What are third-generation data catalogs?	19
The four pillars of third-generation data catalogs	21



Chapter 1

The evolution of data management

- **Data Catalog 1.0:** The era of metadata management by and for IT teams
- **Data Catalog 2.0:** The era of data inventories powered by data stewards

The evolution of metadata management

Metadata management has come a long way since the birth of the internet in the 1990s.

For over a 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?

Today, though, it's all about actually using our data. 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?

These questions keep CDOs (Chief Data Officers) and CDAOs (Chief Data and Analytics Officers) up at night. We all know that metadata management is the solution, but that's easier said than done.

Data practitioners have been struggling with their metadata for decades. 😞

Before we dive into today's metadata management world, let's take a quick stroll through its history to understand how we got to our current cataloging crisis. This history can broadly be broken down into three stages.

¹ Patricia Kennedy, "Manifestations of metadata", Australian Library Journal (2008); Dataversity, "A Brief History of Metadata" (2021)

Data Catalog 1.0

Fundamentally built for IT users, acted as a "data inventory"

1990s & 2000s

ORACLE

The birth of the internet and traditional data warehouses



Hadoop
The rise of big data, data management tools, and data lakes

Data Catalog 2.0

"Data stewardship" tools built for top-down governance programs

2010s



Data Catalog 3.0

2020s

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!¹



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



Informatica



talend

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. 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.²

As the internet grew, the volume of data started increasing exponentially and we quickly entered the world of big data.

“

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 acknowledgment that it is important](#).³

”



Ralph Kimball

AUTHOR AND EXPERT ON DATA WAREHOUSING

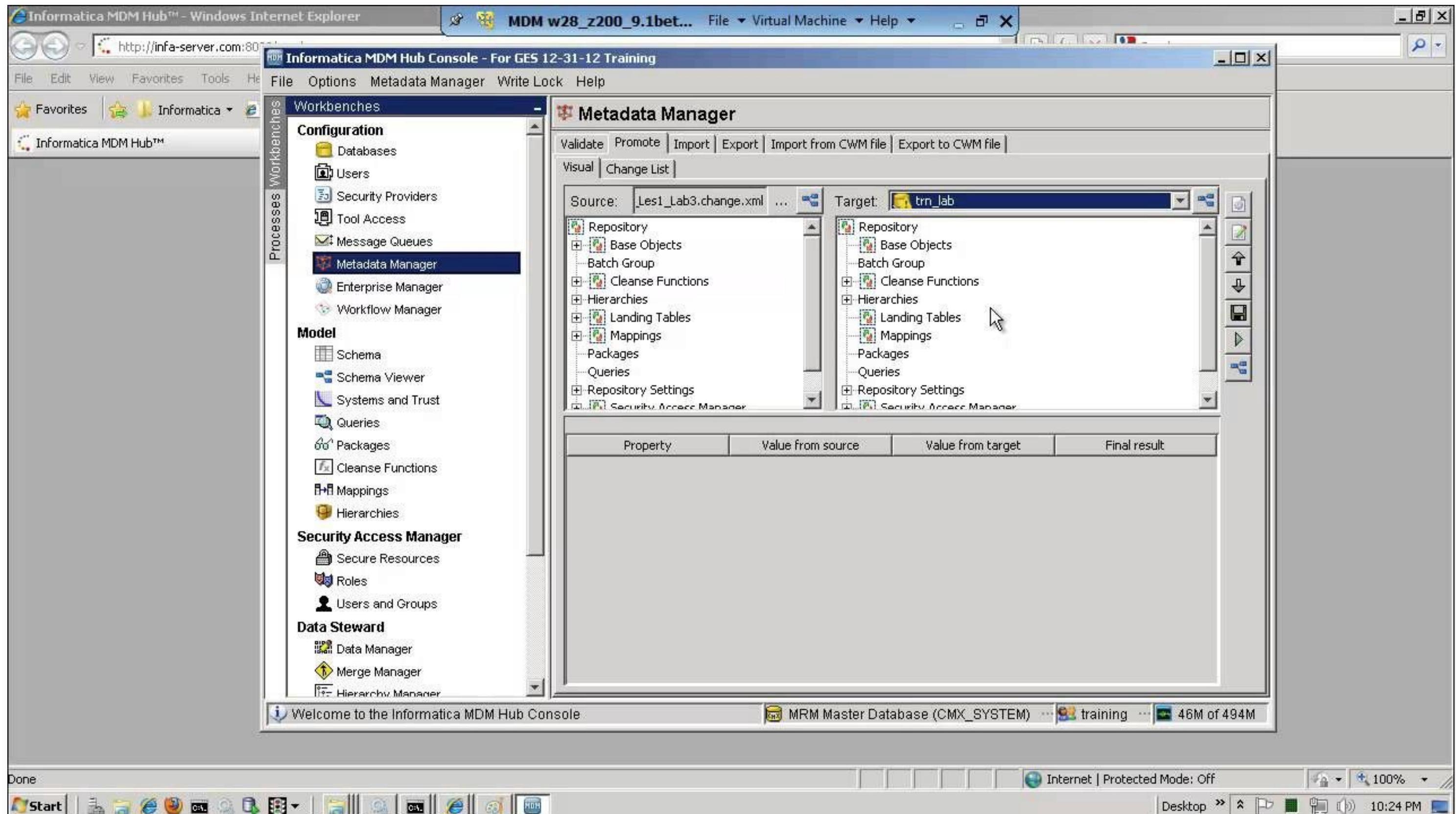
[For companies, the promise of big data quickly turned into a headache. They had plenty of data but no context.](#)

Companies had no clue how to interpret or use the new deluge of data. To keep on top of their new data, IT teams were put in charge of creating data inventories. These 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 solution. Instead, [setting up and keeping on top of these first-generation data catalogs became a constant struggle for IT folks](#).

² Lesk, "How Much Information Is There In the World?" (1997)

³ Ralph Kimball and Margy Ross, *The Data Warehouse Toolkit: Second Edition* (2002)



Informatica's Metadata Manager in 2012

Data Catalog 2.0

Data inventories powered by data stewards

New concepts of data stewardship and context-driven metadata led to the creation of data catalogs.

2010s – 2020



Collibra



Alation

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

The good: countless people were dipping their toes into the world of data. Data 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 root. This referred to a dedicated suite of people who were 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](#). Instead, 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, they 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-premise. Each data system would have its own installation, and companies couldn’t roll out software changes by pushing a simple cloud update.

[These data catalogs proved difficult to set up and maintain. They involved rigid data governance committees, formal data stewards, complex technology setup, and lengthy implementation cycles.](#)

[All in all, this process could take up to 18 months.](#) (We’ve heard of projects that took even longer, like one that took five years!)

The result — technical debt grew, and metadata management steadily fell behind the rest of the modern data stack. 😢

The screenshot shows the Alation Data Catalog interface. The top navigation bar includes 'Compose', 'Sources', 'Glossaries', 'Search', and user profile icons. A search bar at the top left contains the query 'summ'. The main content area displays a search result for 'summ'.

Search Results:

- [MySQL] Analytics / ippss . summ_top_drg** (Endorsed by Michael Long, Steve Burger, Paul Walker)
- Clinician Outpatient Summary** (by Mike Lupo)
- Summary of Top DRG** (by Paul Walker)
- summ_top_drg** (Summary Top Diagnosis Related Groups) - This is the selected item, showing its details:
 - Overview**: Columns 12, Samples 90, Filters 34, Joins 17, Lineage, Queries 113.
 - Business Usage Guideline**: No business usage guideline.
 - Description**: The data provided here include hospital-specific charges for the more than 3,000 U.S. hospitals that receive Medicare Inpatient Prospective Payment System (IPPS) payments for the top 100 most frequently billed discharges, paid under Medicare based on a rate per discharge using the Medicare Severity Diagnosis Related Group (MS-DRG) for Fiscal Year (FY) 2011, 2012, and 2013. These DRGs represent more than 7 million discharges or 60 percent of total Medicare IPPS discharges. This data provides insight into doctor payments through medicare.
 - Powered By**: Diagnosis in the Northeast [-]
 - Query Example**:

```
SELECT drg, provider_state FROM ippss.summ_top_drg WHERE provider_state IN ('NY', 'NJ', 'CT', 'MA', 'RI', 'PA', 'NH', 'ME', 'VT') /* Northeast */
```
 - Mirrored Table**: Finance and Research Database.
- [Presto] Education / ed_stats coll_stats_2009** (College Scorecard Summary 2009)
- Revenue Summary** - Partner Sales Info (by Sergey Astretsov)

Sidebar Navigation:

- Data
- File Systems
- Queries
- Articles
- Conversations
- Business Intelligence

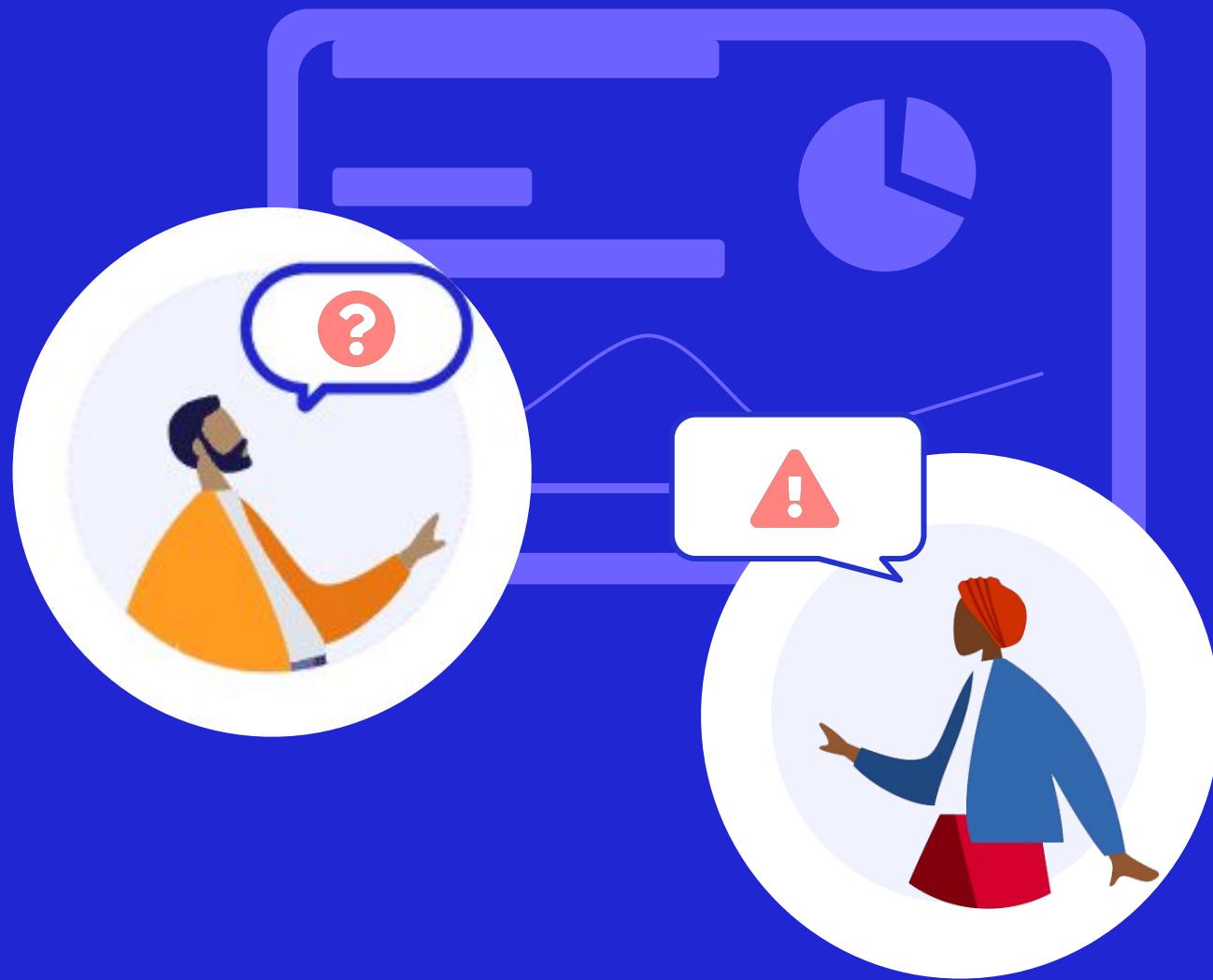
Right-hand Side Panel:

- 73 Conversations**
- Top Users** (Anders Fischer, Ben Lambert, Gianthomas Volpe, Steph Yuen, Connor Glazier)
- Stewards** (Steph Yuen (SE), Paul Walker (Data Ana...), Ben Lambert (Data Gu...), Connor Glazier (Sales ...))
- Tags** (PII, DRG Test)

Alation's data catalog in 2019

Chapter 2

The problem with traditional data catalogs



- The five trends driving Data Catalog 3.0
- Modern data cataloging for the modern data stack

The new world of modern metadata management

In the past couple of years, we've entered the world of third-generation data catalogs.

Third-generation data catalogs didn't come about out of the blue. Instead, they're a response to major changes in the data ecosystem — just like the birth of the internet, big data, and concepts like data stewardship triggered the creation of first and second-generation data catalogs.



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



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



Data governance is being reimaged from top-down, centralized rules to bottom-up, decentralized initiatives.



As metadata itself becomes big data, **the metadata lake has the potential to power infinite use cases** like data discovery, lineage, observability, meshes, and more.



Passive metadata systems are being scrapped in favor of **active metadata platforms**.

¹ Shirshanka Das, "DataHub: popular metadata architectures explained", LinkedIn Engineering (2020)

² Prukalpa Sankar, "Data Catalog 3.0", Humans of Data (2021)

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

The backstory behind "Data Catalog 3.0"

Is this the first time you've heard about "Data Catalog 3.0" or "third-generation data catalogs"? Some people have already seen these terms in the data wild, but they're fairly new.

One of the first instances of this idea was in December 2020, when LinkedIn wrote about the three generations of metadata architectures.¹ Shortly after, we wrote about the three generations of data catalogs — calling them Data Catalog 1.0, 2.0, and 3.0 — in January 2021.²

Since then, we've seen this terminology explode. ⚡ We've even seen RFPs that specifically ask for a Data Catalog 3.0!

The five trends driving third-gen data catalogs

Before we dive into third-generation data catalogs, let's take a closer look at why they exist.

1. The creation of the modern data stack

Starting around 2016, the modern data stack went mainstream. This refers to a flexible collection of tools and capabilities that help businesses today 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

Today's modern data stack is easy to set up, pay as you go, and plug and play — people won't put up with anything else these days! [Tools like Fivetran and Snowflake let users set up a data warehouse in less than 30 minutes.](#)

In this new world, traditional data catalogs can't keep up. Even trying out second-generation catalogs involves significant setup time and at least five calls to get a demo. This has created a need for a truly modern metadata solution that is just as fast, flexible, and scalable as the rest of the modern data stack.

Key characteristics of the modern data stack



Super fast set-up

No lengthy sales process, dozens of demo calls, or long implementation cycles. Just log in, pay with a credit card, and get started.



Pay as you go

No upfront payments and million dollar licenses. Power is in the hands of the consumers, who only pay for what they actually use.



Plug and play

With constant evolution, the best tools don't enforce old-school “lock in”. They're built around open standards, APIs, and easy integrations.



Elastic compute

Compute happens on the cloud with elasticity and auto-scalability. When users aren't consuming data, why process it?



No monoliths

On-premise, monolithic systems turned into agile architectures that are easier and quicker to modify from anywhere.



Always available

No waiting to ingest or process data before it can be used. Data is always available in a data warehouse or lake.

[Learn more about the modern data stack →](#)

DATA INGESTION

Fivetran

Stitch

SINGER

Airbyte

DATA WAREHOUSE

snowflake

amazon
REDSHIFT

DATA LAKE

Starburst

amazon
ATHENA

DATA LAKEHOUSE

databricks

BUSINESS INTELLIGENCE

Looker

+ tableau

M MODE

ThoughtSpot

DATA SCIENCE

jupyter

DataRobot

DATA TRANSFORMATION

dbt

MATILLION

Apache Airflow + R python™

DATA ACCESS & GOVERNANCE

📁 Data Discovery

🔍 Data Cataloging

🔍 Data Observability

⚙️ Visual Query Workbench

⌚ Metrics Repository

🔗 Data Lineage & RCA

atlan

acceldata

transform

DataHub

MONTE CARLO

Amundsen

Key elements and tools in the modern data stack ✨

2. The diverse “humans of data”

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

However, **today's data teams are more diverse than ever**

before. They 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, equally diverse data tools — everything from SQL, Looker, and Jupyter to Python, Tableau, dbt, and R.

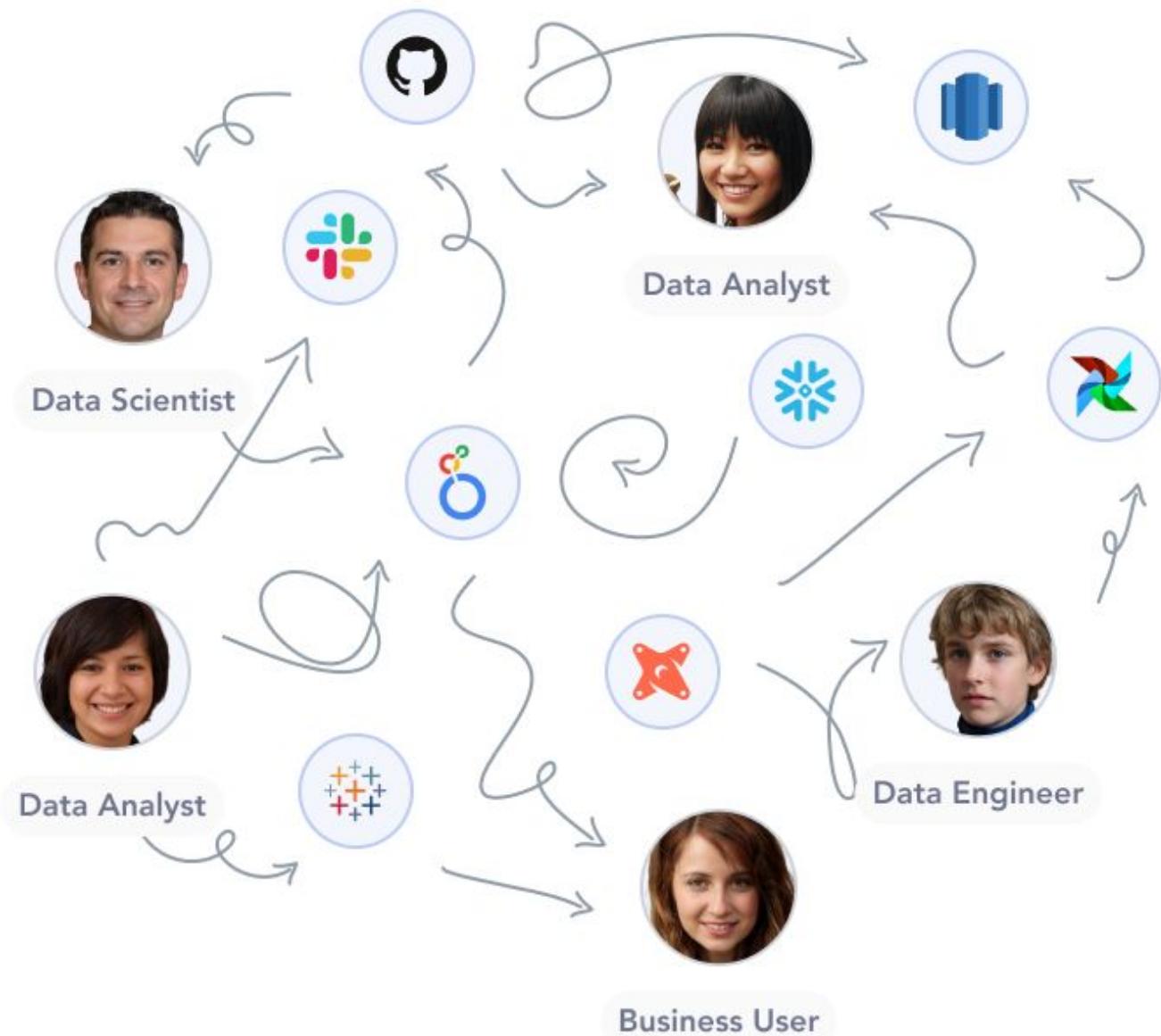
This diversity is both a strength and a struggle.

All of these people have different tools, skill sets, tech stacks, work styles, 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 usually means more chaos within collaboration.** 🤬

This diversity also means that self-service is no longer optional. Modern data tools need to be intuitive for a wide range of users with a wide range of skill sets. 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.

Modern metadata is emerging as the solution to provide critical context as we bring an increasingly diverse set of people and tools into our data ecosystem.



[Learn more about the humans of data →](#)

3. The new vision for data governance

Data governance is seen as a bureaucratic, restrictive process — a set of rules dropped down from on high to slow down your work. 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.

However, as the modern data stack has made it easier to ingest and transform data, **this idea of data governance has become one of the biggest barriers in daily data work**. For the first time, the need for governance is being felt bottom-up by practitioners, instead of being enforced top-down due to regulation.

Data governance is currently in the middle of a paradigm shift. Today, governance is becoming something that the humans of data embrace rather than fear. At its heart, it's 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 that are about empowering and creating better data teams, not controlling them.

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

[Learn more about modern data governance →](#)

 Data governance

→  "Data and analytics" governance

 Centralized approach

→  Decentralized, community-led approach

 Afterthought

→  Part of daily workflows

“

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.³**

”



Tristan Handy

CEO & FOUNDER, DBT

³ Handy, ["The Modern Data Stack: Past, Present, and Future"](#), dbt (2020)

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 how can you set up a data system for limitless use cases? That led to the birth of the data lake.

Today, metadata is at the same place. **Metadata is itself becoming big data**, and technical advances (i.e. elasticity) in compute engines like Snowflake and Redshift make it possible to derive intelligence from metadata in a way that was unimaginable even a few years ago. 🤯

As metadata increases, and the intelligence we can derive from it increases, so too do its use cases.

Today, even the most data-driven organizations have only scratched the surface of what is possible with metadata. However, it holds the key to unlocking a truly intelligent modern data stack — powering not just the use cases of today like data cataloging, lineage, and observability, but also those of tomorrow 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.



Powered by 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.

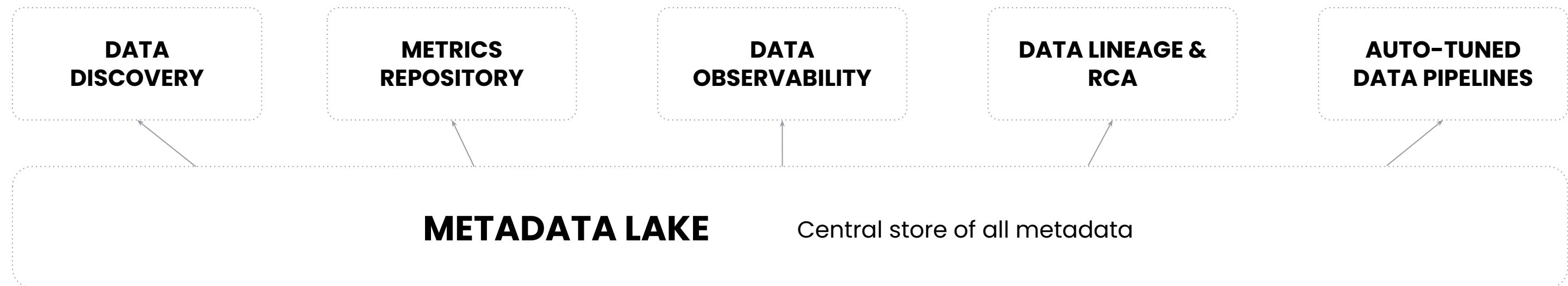


Power humans & machines

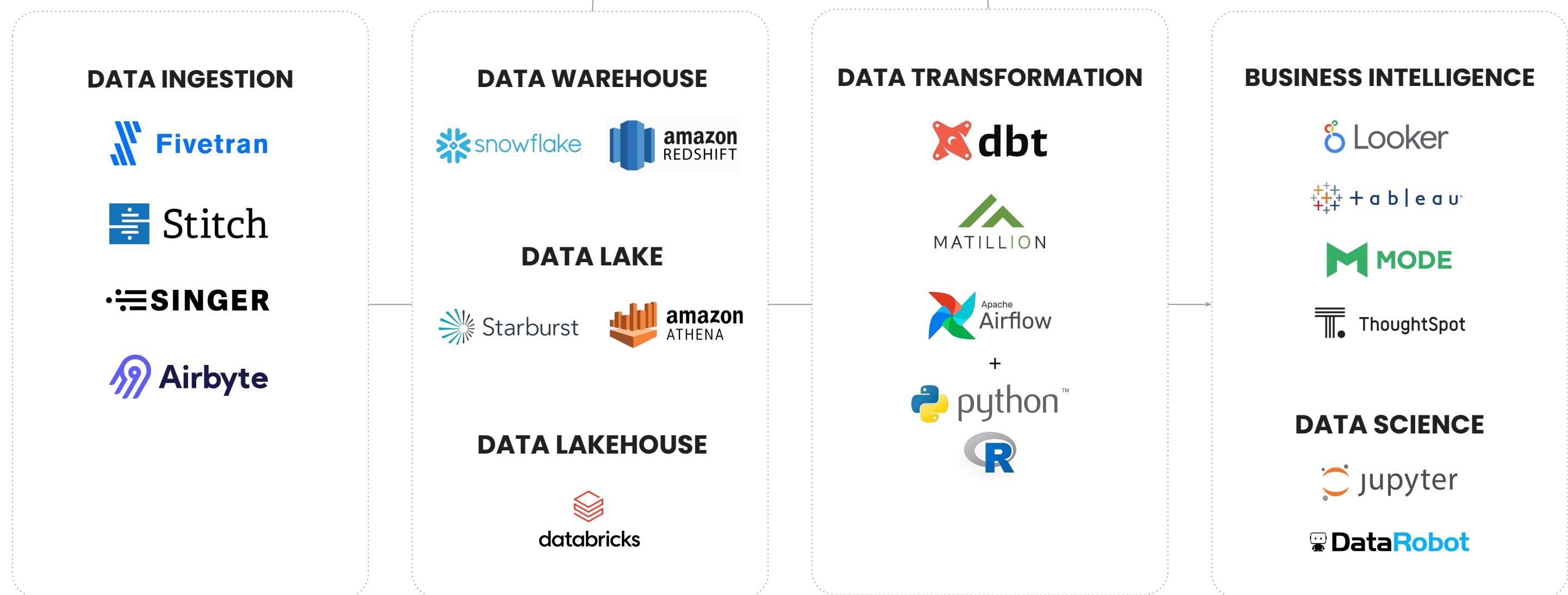
The metadata lake can improve both humans’ daily work (such as discovering data and understanding its context) and machines or tools’ daily workflows (such as auto-tuning data pipelines). This means that metadata lakes need to be fundamentally flexible and adaptable.

[Learn more about metadata lakes →](#)

A FEW OF THE MANY USE CASES:



MODERN DATA STACK:



Architecture of a metadata lake

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 way of thinking about metadata.

Traditional data catalogs are passive. They are fundamentally static systems that don't drive any action and rely on human effort to curate and document data.

However, **an active metadata platform is an always-on, intelligence-driven, action-oriented system.**

 **Always-on:** Rather than waiting for humans to manually enter metadata, it continuously collects metadata from logs, query history, usage stats, etc.

 **Intelligence-driven:** It constantly processes 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.

[Learn more about active metadata →](#)

The Gartner Magic Quadrant for Metadata Management was just scrapped. Here's everything you need to know.

This marks a paradigm shift in how we should approach metadata. Here's the why and how.

An announcement of this massive change from Gartner

“

The stand-alone metadata management platform will be refocused from augmented data catalogs to a metadata “anywhere” orchestration platform.

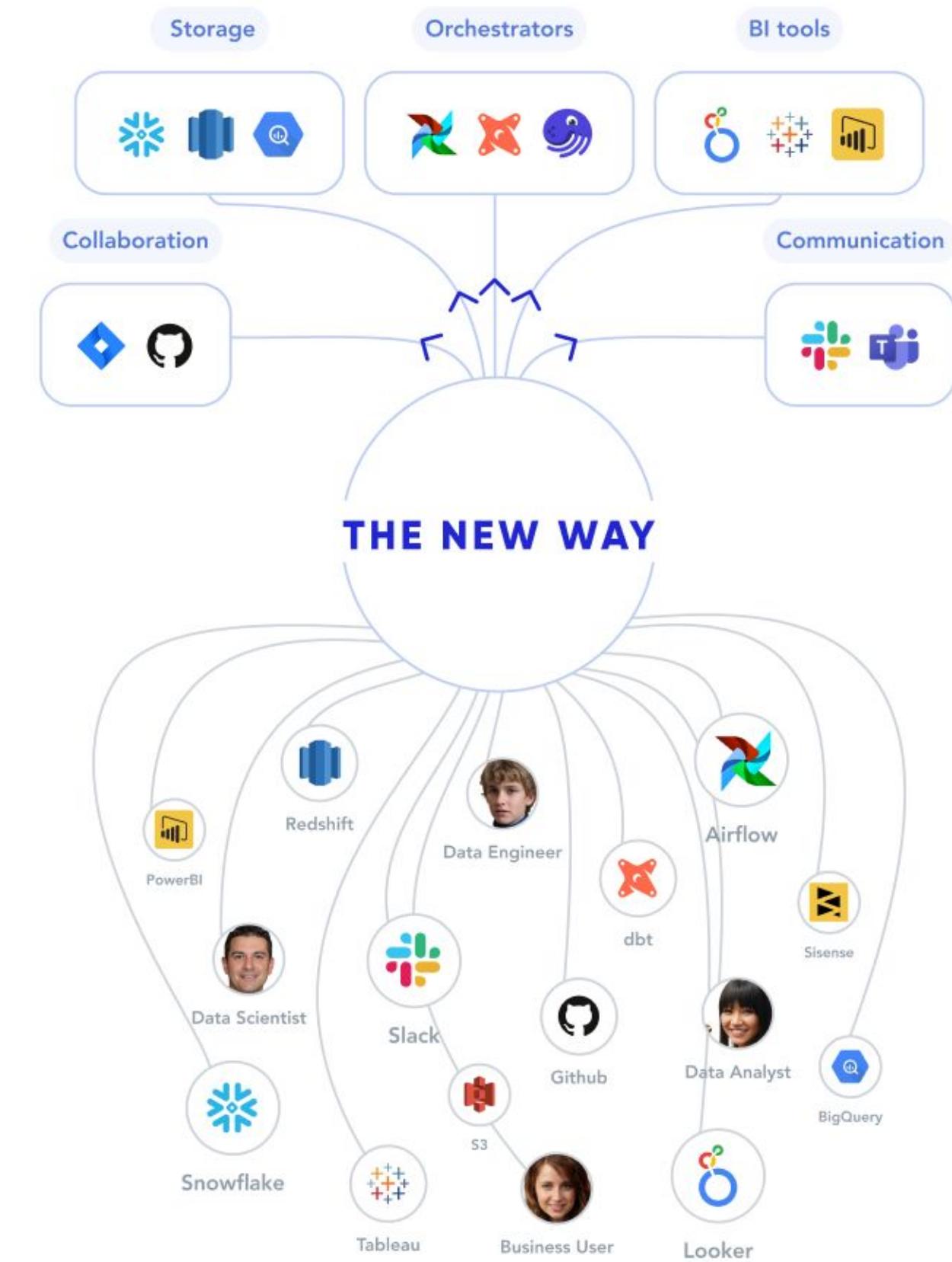
”



Gartner

MARKET GUIDE FOR ACTIVE METADATA MANAGEMENT (2021)

“another siloed tool!” 😠



Our vision for active metadata

Modern metadata for the modern data stack

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

While data infrastructure has evolved, metadata management hasn't. Traditional data catalogs were built for on-prem data warehouses and the Hadoop era. As the modern data stack is innovating at an unprecedented rate, these data catalogs are falling behind.

Second-generation data catalogs are...

- Built for on-premise systems, rather than the cloud world
- 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.⁴ Check out some examples below.

However, not all companies can do this, and it's inefficient to create dozens of similar tools.

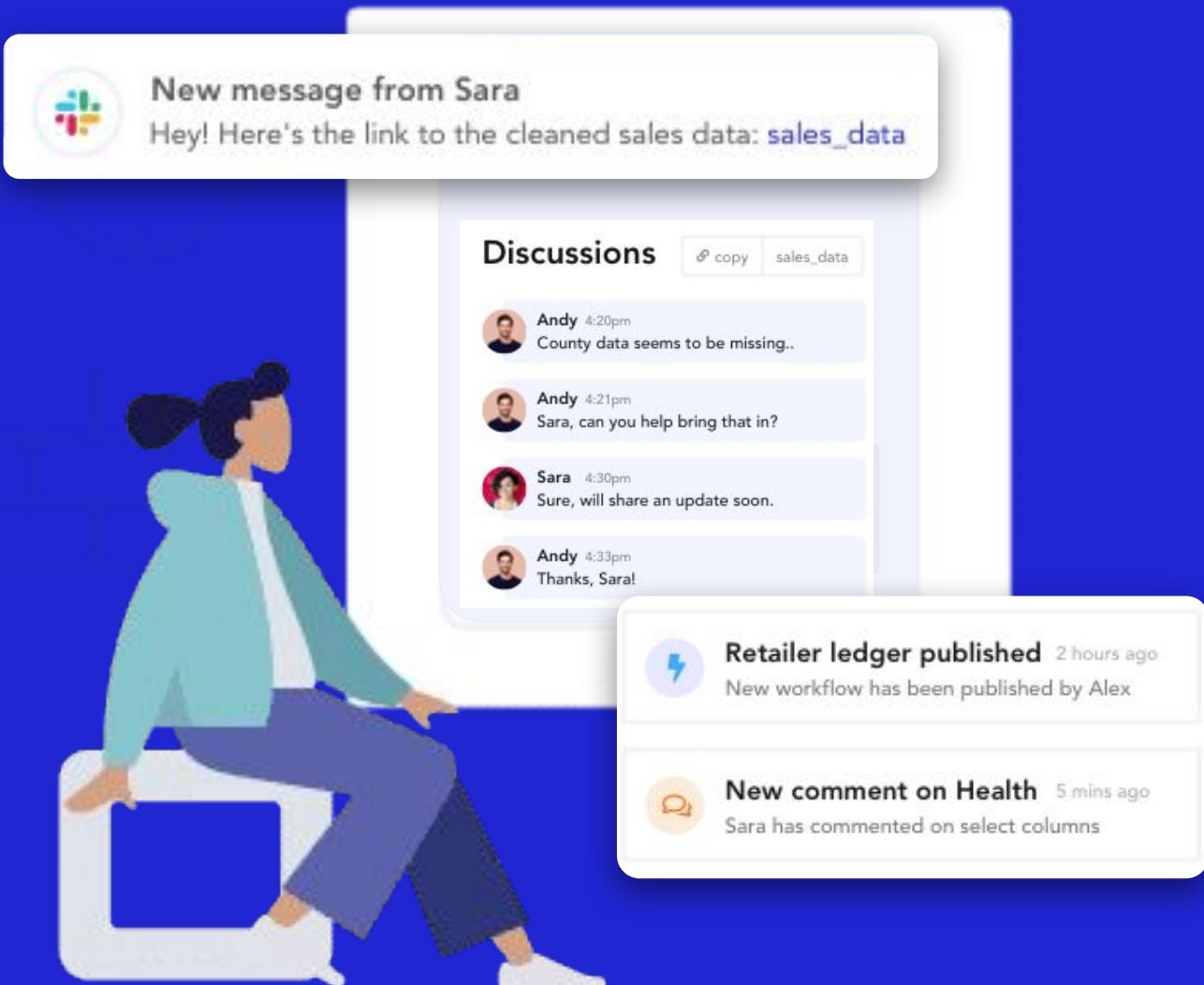
It's time for a modern metadata solution, one that is just as fast, flexible, and scalable as the rest of the modern data stack.



⁴ LinkedIn, [DataHub](#), GitHub (2021); Mark Grover, [“Amundsen”](#), Lyft Engineering (2019); Chris Williams, [“Democratizing Data at Airbnb”](#), Airbnb Engineering (2017); Haran Talmon, [“Nemo”](#), Engineering at Meta (2020); Majumdar and Li, [“Metacat”](#), The Netflix Tech Blog (2018); Li, Onuk, and Tindal, [“Databook”](#), Uber Engineering (2018)

Chapter 3

The era of Data Catalog 3.0



- The objectives of a third-generation data catalog
- The four pillars of a third-generation data catalog

What are third-generation data catalogs?

Third-generation data catalogs will not look and feel like their old-school predecessors.

Today we're at an inflection point in metadata management — a shift from the slow, on-premise Data Catalog 2.0 to the start of a new era, Data Catalog 3.0.

Instead of emulating their old-school predecessors, third-generation data catalogs will feel more like the collaborative, self-service tools in today's modern workplace.

Think GitHub, Figma, Slack, Notion, and Superhuman.

Third-generation data catalogs will be built on four key pillars:



One size doesn't fit all in augmented data management.

Every company, industry, and context is different, and a single machine learning algorithm won't solve all your data management problems. Third-gen tools understand this and build programmability into AI/ML algorithms.



Context should be embedded into teams' daily workflows.

Third-generation data catalogs will not look and feel like their predecessors from the 2.0 generation. Instead, they will be built on the premise of embedded collaboration, borrowing principles from the modern tools that teams already use and love.

Data Catalog 1.0

Fundamentally built for IT users, acted as a "data inventory"

1990s & 2000s

ORACLE

The birth of the internet and traditional data warehouses

Data Catalog 2.0

"Data stewardship" tools built for top-down governance programs

2010s

hadoop

The rise of big data, data management tools, and data lakes

Data Catalog 3.0

Modern metadata for the modern data stack

2020s

databricks aws

snowflake

The modern data stack goes mainstream

1990s & 2000s



Data Catalog 1.0

Metadata management by and for IT teams

The birth of the internet, big data, and traditional data warehouses led companies to create simple inventories for their data — run by and for IT users.

Setting up and keeping on top of these first-generation data catalogs immediately became a constant struggle for IT folks.

2010s



Data Catalog 2.0

Data inventories powered by data stewards

Amidst the rise of data management tools and data lakes, “data stewardship” tools were built for top-down data catalogs and governance programs.

Built for the Hadoop era, these data catalogs proved difficult to set up and maintain. Technical debt grew, and data catalogs steadily fell behind the rest of the data stack.

2020s



Data Catalog 3.0

Modern metadata for the modern data stack

Most early adopters of the modern data stack ended up building internal tools for metadata management to fill this gap. However, not all companies can do this, and it’s inefficient to create dozens of similar tools.

It’s time for a modern metadata solution, one that is just as fast, flexible, and scalable as the rest of the modern data stack.

Key capabilities:



Programmable
bots



End-to-end
visibility



Embedded
collaboration



Open by
default

The evolution from Data Catalog 1.0 → 2.0 → 3.0

The 4 pillars of third-generation data catalogs

The fundamental principles behind today's new third generation of data catalogs

1. Programmable bots

In the past few years, “augmented” data catalogs (which use machine learning to automate a set of manual tasks) have become more and more popular.

This is a step in the right direction, but it’s not a silver bullet. That’s because every company is unique. Every industry is unique. Every individual team’s data is unique. 😊

No matter how good it is, no ML algorithm can fully solve the world’s data management problems. **No one algorithm can magically create context, identify anomalies, and achieve the intelligent data management dream — 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.

1st Gen

Manual, human-led documentation

2nd Gen

Augmented catalogs with “one size fits all” approach

3rd Gen

Programmable bots

Like bots on



Column-description recommendation bots



Regulation & risk-reduction bots

2. Embedded collaboration

Because of the fundamental diversity in data teams, data tools need to be designed to integrate seamlessly with teams' daily workflows.

This is where the idea of embedded collaboration really comes alive. Embedded collaboration is about work happening where you are, with the least amount of 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.

“

Designing the interface and user experience of a data tool should not be an afterthought.¹

”



1st Gen

Data inventories with solutions built by IT users

2nd Gen

Data catalogs with some business-facing UIs

3rd Gen

Embedded collaboration

Integrated into teams' daily workflows



¹ Chris Williams & John Bodley, “[Democratizing Data at Airbnb](#)”, Graph Connect Europe (2017)

Examples of these principles in action

Embedded collaboration

Programmable bots



Security & compliance bots

As security requirements go mainstream, 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.

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



Classification bots

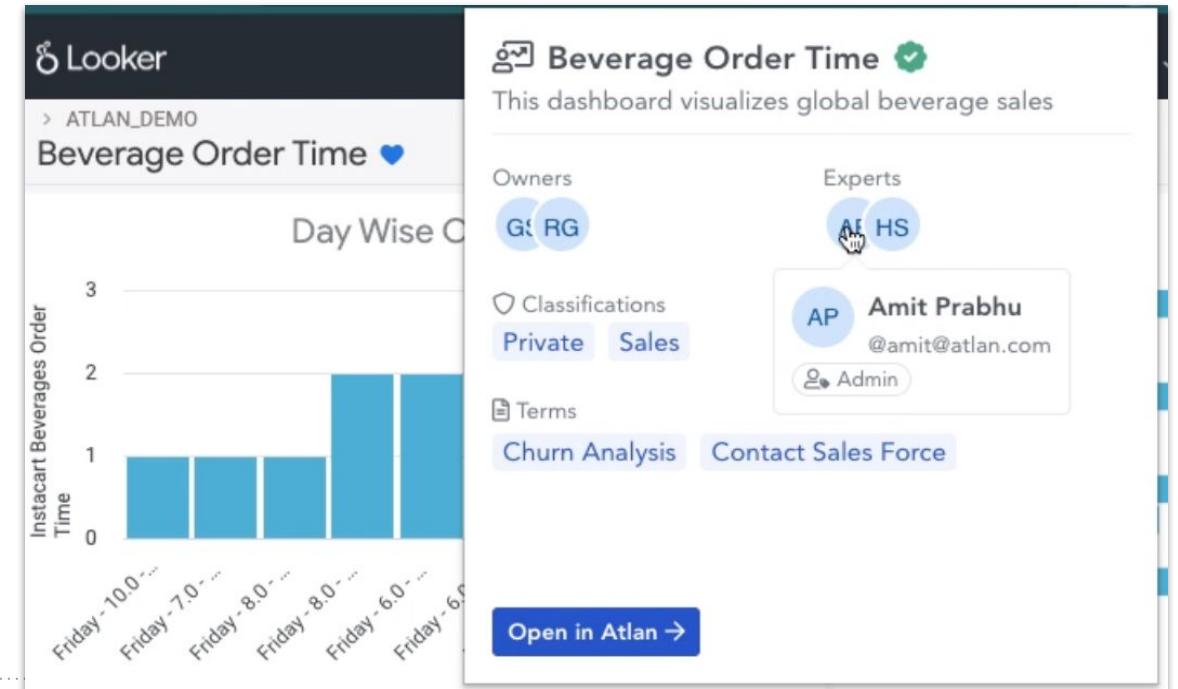
Companies with specific naming conventions for their data sets can create bots to automatically organize, classify, and tag their data ecosystem based on preset rules.

For 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.



Reverse metadata: Make context available in daily tools



#team-sales



Sam - Data Lake Owner

5:16 PM

@andrew hey, why is there a spike in missing values on #sales-nov-agg?



Let me check, a new DAG was scheduled to run on #sales-nov-agg earlier today



5 replies



#team-dataadmins



New Access Request

5:22 PM

@damien requested to view #sales_orders

Approve



#alerts-dataquality



⚠ 60% increase in column missing values detected in dataset #sales-nov-agg

View

Comment

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 information about data assets is spread across different places** — tools for data lineage, quality, prep, and more. To get the full picture for a data asset, users need to ask multiple people and check information across multiple tools.

With complete visibility into every data asset, **third-generation data catalogs will 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 data set, auto-generated from query history
- where it comes from, from automated lineage
- whether it is trustworthy, from quality scores and how recently the data was updated
- which columns are used most, and how people use them
- and most importantly, a preview of the data itself!

All available in one seamless experience. No more constant toggling between all the tools in your modern data stack.

1st Gen

Piecemeal solutions for quality, lineage, catalog, etc.

2nd Gen

Some partnerships between piecemeal solutions (e.g. data lineage ↔ catalog)

3rd Gen

End-to-end visibility

- ✓ Visual data previews & related queries
- ✓ Column-level lineage
- ✓ Data classifications, access control & governance
- ✓ Custom metadata to bring in context from ETL tools, orchestration tools, and more

Like code has a profile on **GitHub**

and customers have a profile on 

4. Open by default

We've already talked about how the modern data stack is built on open standards and tools, so this principle shouldn't be a surprise.

We're 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 or automatically creating column-level lineage from SQL code in query logs.

Here's one example: 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.

1st Gen

Closed solutions

2nd Gen

Closed solutions
with open APIs

3rd Gen

Open by default

- ✓ Metadata lake
- ✓ Open APIs
- ✓ Knowledge graph-based

Metadata Plugin
Marketplace

Like design plugins on  Figma
or code sharing on  npm

Examples of these principles in action

Open by default

Custom plugins for data asset integrations



for new data asset syncs

Embedded collaboration plugins



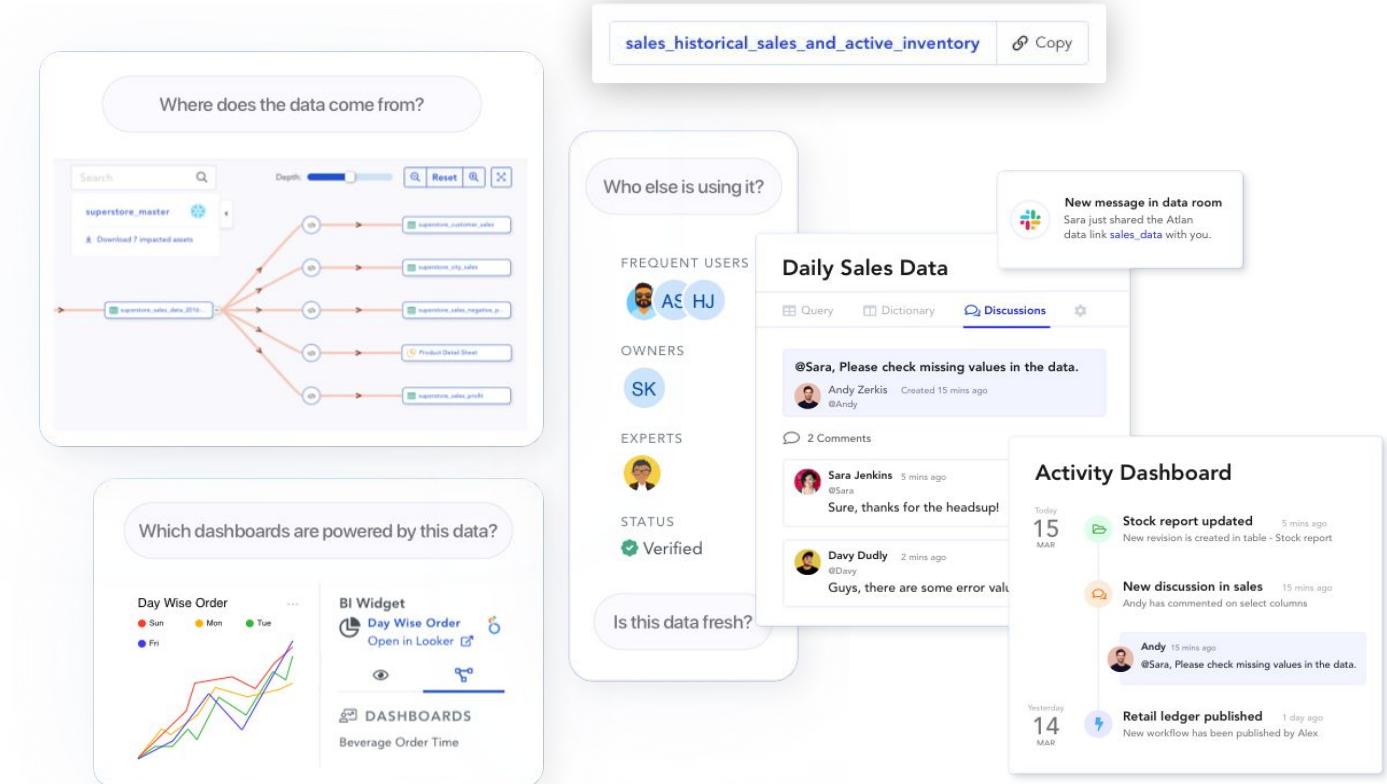
for issue management

Programmable bots



for classifying CIA ratings

End-to-end visibility



Knowledge & Context



Trust & Visibility



Meet atlan

The ultimate third-generation data catalog
for modern data teams

Backed by marquee investors



and the top founders & CEOs pioneering the modern data stack



Leading the DataOps & Active Metadata categories



Named in [3 Gartner Hype Cycles in 2021](#) — the only company to be named in the Active Metadata and DataOps categories



Named in Gartner's [Inaugural Market Guide For Active Metadata Management](#)



Named a [Cool Vendor in DataOps](#) in Gartner's inaugural DataOps report

A screenshot of the Atlan data catalog interface. The top navigation bar shows "Atlan | Hope you are having a good day, Adam!" and "What's New". The main area is titled "Catalog" with a "NEW TABLE" button. On the left, there are filters for "Table Owners" (Sarah, Sandra, James, Darwin, Geeta) and "Status" (Verified). The catalog lists several datasets:

- Motor Insurance: Internal Claims Data** (verified): Motor insurance claims data for fraud analytics and customer exp. Fraud, Analytics, Customer Experience. 98,088 rows, 39 columns, Ujwal, 9,576 rows, 9 hours.
- AAA Rated Insured** (verified): This is a dataset of highest rated customers of the company. customer, ratings, Health. 21,600 rows, 39 columns, Ujwal, 6,048 rows, 9 hours.
- Outlet sales raw 2018**: Click to add description. Bootcamp. 27,751,391 rows, 7 columns, Ankita, 27,751,391 row.
- Outlet sales raw 2019**: Click to add description.

- **Cloud-native** data catalog
- **24 hours** to get up and running
- **Democratization** for businesses
- **Governance** for IT teams

[SEE A DEMO](#)

[LEARN MORE](#)