

atlan

# The Future of the Modern Data Stack in 2023

Featuring 4 new emerging trends and 6 big trends from last year



# Table of Contents

<b>Introduction</b>	<b>1</b>
<b>4 new trends that will emerge in 2023</b>	<b>2</b>
1. Optimizing data spend will become a major priority	<b>2</b>
2. Data teams will start being run around ROI and metrics	<b>5</b>
3. The modern data stack will start consolidating	<b>7</b>
4. Modern data stack companies will start expanding into on-prem connectors	<b>9</b>
<b>6 trends that will carry through from 2023</b>	<b>10</b>
1. Active metadata will replace the “data catalog” category	<b>10</b>
2. Data contracts and data governance will start shifting “left”	<b>13</b>
3. The semantic layer will enter “adoption mode”, albeit slowly	<b>16</b>
4. Data activation will replace CDPs as marketing spend becomes more important	<b>18</b>
5. The first wave of data mesh implementations will start going live	<b>21</b>
6. Data observability and quality will converge in a “data reliability” category	<b>24</b>
<b>Last thoughts</b>	<b>26</b>

# Introduction

As we close out 2022, it's amazing to see how much the data world has changed.

It was less than a year ago in March that Data Council happened. Yes, it was just an event. But it was *the* event, the first in-person conference since COVID. It was the data world coming alive again and meeting face to face for the first time in two long years.

Since then, we've been busy stirring up controversy with our hot takes, debating our tech and community, raising important conversations, and duking it out on Twitter with Friday fights. We were in growth mode, always searching for the next new thing and vying for a chunk of the seemingly infinite data pie.

Now we're entering a different world, one of recession and layoffs and budget cuts that 98% of CEOs expect will last 12–18 months. Companies are preparing for war, amping up the pressure and shifting from growth mode to *efficiency mode*.

In 2023, we'll face a new set of challenges — improving efficiency, refocusing on immediate impact, and making data teams the most valuable resource in every organization.

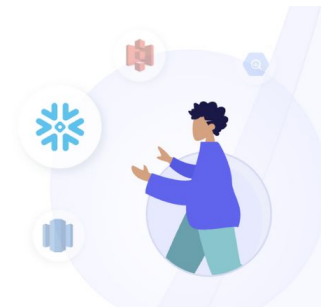
So what does this mean for the data world?

This report breaks down the 10 big trends that we think will happen in the modern data stack this year — 4 emerging trends that will be a big deal in the coming year, and 6 existing trends that are poised to grow even further.

# 4 new trends that will emerge in 2023

With the recent economic downswing, the tech world is looking into 2023 with a new focus on efficiency and cost-cutting. This will lead to four new trends related to how modern data stack companies and data teams operate.

## 1. Optimizing data spend will become a major priority



Storage has always been one of the biggest costs for data teams. For example, Netflix spent \$9.6 million per month on AWS data storage. As companies tighten their budgets, they'll need to take a hard look at these bills.

Snowflake and Databricks have already been investing in product optimization. We'll likely see more improvements to help customers cut costs this year.

For example, in its June conference, Snowflake highlighted product improvements to speed up queries, reduce compute time, and cut costs. It announced 10% average faster compute on AWS, 10-40% faster performance for write-heavy DML workloads, and 7-10% lower storage costs from better compression.

At its June conference, Databricks also devoted part of its keynote to cost-saving product improvements, such as the launches of Enzyme (an automatic optimizer for ETL pipelines) and Photon (a query engine with up to 12x better price to performance).

Later in the year, both Snowflake and Databricks doubled down by investing further in cost optimization features, and more are sure to come next year. Snowflake even highlighted cost-cutting as one of its top data trends for 2023 and affirmed its commitment to minimizing cost while increasing performance.

**In 2023, we'll also see the growth of tooling from independent companies and storage partners to further reduce data costs.**

Dark data, or data that never actually gets used, is a serious problem for data teams. Up to 68% of data goes unused, even though companies are still paying to store it.

This year, we'll see the growth of cost-management tools like Bluesky, CloudZero, and Slingshot designed to work with specific data storage systems like Snowflake and Databricks.

**We'll also see modern data stack partners introduce compatible optimization features, like dbt's incremental models and packages.**

dbt Labs and Snowflake even wrote an entire white paper together on optimizing your data with dbt and Snowflake.

Metadata also has a big role to play here. With a modern metadata platform, data teams can use popularity metrics to find unused data assets, column-level lineage to see when assets aren't connected to pipelines, redundancy features to delete duplicate data, and more.

Much of this can even be automated with active metadata, like automatically optimizing data processing or purging stale data assets.

For example, a data team we work with reduced their monthly storage costs by **\$50,000** just by finding and removing an unused BigQuery table.

Another team deprecated **30,000 unused assets** (or two-thirds of their data estate) by finding tables, views, and schemas that weren't used upstream.



**Bojan Tunguz**

@tunguz

Painting:

"The arrival of the AWS bill."

Oil on canvas.



**Jake Thomas**

@aerialfly

Data engineers should not write ETL.

Instead, they should spend ~80% of their day keeping airflow running, ~10% cramming dbt manifest.json into dags, and ~10% figuring out how to not get yelled at by finance for the Snowflake bill.

♡ 552 2:47 PM - Dec 1, 2022

## 2. Data teams will start being run around ROI and metrics



“[Data Domain and ServiceNow] were built and run for performance, full stop...”

Our companies ran at a higher velocity, with higher standards and a narrower focus than most. Going faster, maintaining higher standards, and with a narrower aperture. Sounds simple? The question is how you go about amping up your organization. How much faster do you run? How much higher are your standards? How hard do you focus?”

— [Frank Sloatman](#)

Frank Sloatman has IPOed three successful tech companies, no small feat in the startup world. He said that his success came down to optimizing team velocity and performance.

In the past few years, data teams have been able to run free with less regulation and oversight.

We have so much belief in the power and value of data that data teams haven't always been required to prove that value.

Instead, they've chugged along, balancing daily data work with forward-looking tech, process, and culture experiments. Optimizing how we work has always been part of the data discussion, but it's often relegated to more pressing concerns like building a super cool tech stack.

Next year, this will no longer cut it.

## As budgets tighten, data teams and their stacks will get more attention and scrutiny.

How much do they cost, and how much value are they providing? Data teams will need to become more like Frank Slootman, focusing on performance and efficiency.

## In 2023, companies will get more serious about measuring data ROI, and data team metrics will start becoming mainstream.

It's not easy to measure ROI for a function as fundamental as data, but it's more important than ever that we figure it out.

This year, we'll see data teams start developing proxy metrics to measure their value. This may include usage metrics like data usage (e.g. DAU, WAU, MAU, and QUA), page views or time spent on data assets, and data product adoption; satisfaction metrics like a d-NPS score for data consumers; and trust metrics like data downtime and data quality scores.



**Sarah Catanzaro**

@sarahcat21

Tell me how you're USING data. I'm sick of hearing about how you're producing data or building data stacks. As one great data scientist once said, the only stacks that matter are those of benjamins.

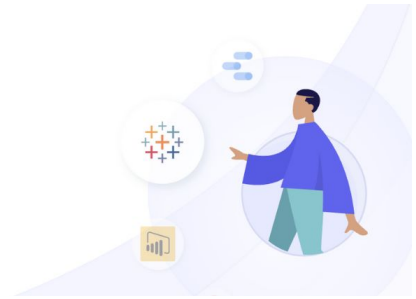


♡ 211 11:08 PM - Mar 29, 2022





### 3. The modern data stack will start consolidating



For years, the modern data stack has been growing. And growing. And growing some more.

As VCs pumped in millions of dollars in funding, new tools and categories popped up every day.

But now, with the economic downturn, this growth phase is over. VC money has already been drying up — just look at the decrease in funding announcements over the last six months.

**We'll see fewer data companies and tools launching next year and slower expansion for existing companies.**

**Ultimately, this is probably good for buyers and the modern data stack as a whole.**

Yes, hypergrowth mode is fun and exciting, but it's also chaotic.

We used to joke that it would suck to be a data buyer right now, with everyone claiming to do everything. The result is some truly wild stack diagrams.

**This lack of capital will force today's data companies to focus on what matters and ignore the rest.**

That means fewer "nice to have" features. Fewer splashy pivots. Fewer acquisitions that make us wonder "Why did they do that?"

With limited funds, companies will have to focus on what they do best and partner with other companies for everything else, rather than trying to tackle every data problem in one platform. This will lead to the creation of the “best-in-class modern data stack”.

As the chaos calms down and data companies focus on their core USPs, the winners of each category will start to become clear.

These tools will also focus on working even better with each other. They’ll act as launch partners, aligning behind common standards and pushing the modern data stack forward.

A couple of examples from last year are Fivetran’s [Metadata API](#) and dbt’s [Semantic Layer](#), where close partners like us built integrations in advance and celebrated the launch as much as Fivetran and dbt Labs.



**Seth Rosen**

@sethrosen

Previously: the simple “modern data stack” when first discovered by data teams was \*magic\*. It made it absurdly easy to quickly centralize, transform, and present data for analytics

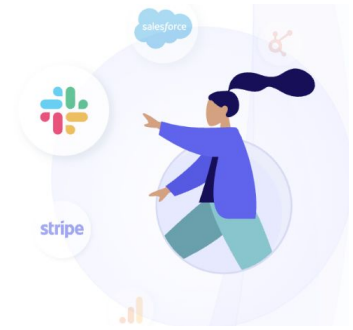
Now: the MDS is a landscape of hundreds of logos across all of “data”

RIP MDS (magic data stack)

♡ 65 5:48 AM - Mar 28, 2022



## 4. Modern data stack companies will start expanding into on-prem connectors



Tech companies are facing new pressure to cut costs and increase revenue in 2023. One way to do this is by focusing on their core functions, as mentioned above. Another way is seeking out new customers.

Guess what the largest untapped source of data customers is today? Enterprise companies with legacy, on-premise data systems. To serve these new customers, modern data stack companies will have to start supporting legacy tools.

**In 2023, the modern data stack will start to integrate with Oracle and SAP, the two enterprise data behemoths.**

This may sound controversial, but it's already begun. The modern data stack started reaching into the on-prem, enterprise data world over a year ago. In October 2021, Fivetran acquired HVR, an enterprise data replication tool. Fivetran said that this would allow it to "address the massive market for modernizing analytics for operational data associated with ERP systems, Oracle databases, and more". This was the first major move from a modern data stack company into the enterprise market.



**Matthew Mullins**  
@mullinsms

Every modern data stack diagram has one dwh at the center, but every enterprise I've ever talked to has at least three. The one that's current, the one they're moving to, and the one they got in an acquisition.

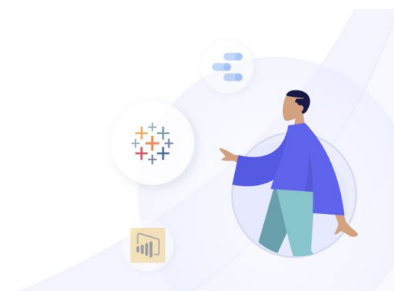
♡ 176 8:39 PM - Aug 16, 2022



# 6 trends that will carry through from 2023

These are six of the big ideas that blew up in the data world last year and only promise to get bigger in 2023.

## 1. Active metadata will replace the “data catalog” category



This was one of the big trends from last year's report, so we're not surprised that it's still a hot topic in the data world. What was surprising, though, was how fast the ideas of active metadata and third-generation data catalogs continued to grow.

**In a major shift from 2021, when these ideas were new and few people were talking about them, many companies are now competing to claim the category.**

Take, for example, Hevo Data and Castor's adoption of the “Data Catalog 3.0” language.

A few companies have the tech to back up their talk. But like the early days of the data mesh, when experts and newbies alike appeared knowledgeable in a space that was still being defined, others don't.

## Last year, analysts latched onto and amplified the idea of active metadata and modern data catalogs.

After its new Market Guide for Active Metadata in 2021, Gartner went all in on active metadata last year. At its August conference, active metadata starred as a key theme in Gartner's keynotes, as well as in what seemed like half of the conference's talks.

G2 released a new "Active Metadata Management" category in the middle of the year, marking a "new generation of metadata". They even called this the "third phase of...data catalogs", in keeping with this new "third-generation" or "3.0" language.

Similarly, Forrester scrapped its Wave report on "Machine Learning Data Catalogs" to make way for "Enterprise Data Catalogs for DataOps", marking a major shift in their idea of what a successful data catalog should look like.

Meanwhile, VCs continued to pump money into metadata and cataloging — e.g. Alation's \$123M Series E, Data.world's \$50M Series C, our \$50M Series B, and Castor's \$23.5M Series A.



@jwills@data-folks.masto.host   
@josh\_wills

To my many friends/followers doing metadata/catalog startups, I have a request: please integrate the metadata info with my BI tool so that I can see it \*while I am doing queries.\*

I have no desire to \*ever\* visit a third website to just "browse the metadata."

♡ 220 8:39 PM - Aug 16, 2022



## Our take on the future of active metadata...

One of the biggest signals from this year was in the new Forrester Wave report. From 2021 to 2022, Forrester upended its Wave rankings. It moved the 2021 Leaders ([Alation](#), [IBM](#), and [Collibra](#)) to the bottom and middle tiers of its 2022 Wave report, and raised previously low or even unranked companies (us, [Data.world](#), and [Informatica](#)) to become the new Leaders.

This is a major sign that the market is starting to separate modern catalogs (e.g. active metadata platforms, data catalogs for DataOps, etc.) from traditional data catalogs.

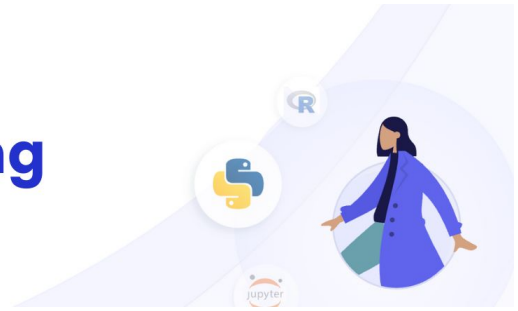
**Our prediction is that active metadata platforms will replace the “data catalog” category in 2023.**

The “data catalog” is just a single use case of metadata: helping users understand their data assets. But that barely scratches the surface of what metadata can do. Activating metadata holds the key to dozens of use cases like observability, cost management, remediation, quality, security, programmatic governance, optimized pipelines, and more — all of which are already being actively debated in the data world. Here are a few real examples:

- **Eventbridge event-based actions:** Allows data teams to create production-grade, event-driven metadata automations, like alerts when ownership changes or auto-tagging classifications.
- **Trident AI:** Uses the power of GPT-3 to automatically create descriptions and READMEs for new data assets, based on metadata from earlier assets.
- **GitHub integration:** Automatically creates a list of affected data assets during each GitHub pull request.

**As the data world aligns on the importance of modernizing our metadata, we’ll see the rise of a distinct active metadata category, likely with a dominant active metadata platform.**

## 2. Data contracts and data governance will start shifting “left”



This started in August with Chad Sanderson’s newsletter on “[The Rise of Data Contracts](#)”. He later followed this up with a [technical guide](#) to data contracts with Adrian Kreuziger.

He then spoke about data contracts on the [Analytics Engineering Podcast](#) — with us! (Shoutout to Chad, Tristan Handy, and Julia Schottenstein for a great chat.)

**The core driver of data contracts is that engineers have no incentive to create high-quality data.**

Because of the modern data stack, the people who create data have been separated from the people who consume it. As a result, we end up with GIGO data systems — garbage in, garbage out.

**The data contract aims to solve this by creating an agreement between data producers and consumers.**

Data producers commit to producing data that adheres to certain rules — e.g. a set data schema, SLAs around accuracy or completeness, and policies on how the data can be used and changed.

After agreeing on the contract, data consumers can create downstream applications with this data, assured that engineers won’t unexpectedly change the data and break live data assets.

After Chad Sanderson’s newsletter went live, this conversation blew up. It spread across Twitter and Substack, where the data community argued whether data

contracts were an important conversation, frustratingly vague or self-evident, not actually a tech problem, doomed to fail, or obviously a good idea. We hosted Twitter fights, created epic threads, and watched battle royales from a safe distance, popcorn in hand.

## **Our take on the future of the data contracts...**

While data contracts are an important issue in their own right, they're part of a larger conversation about how to ensure data quality.

It's no secret that data is often outdated or incomplete or incorrect — the data community has been talking about how to fix it for years. First we said that metadata documentation was the solution, then it was data product shipping standards. Now the buzzword is data contracts.

This is not to dismiss data contracts, which may be the solution we've been waiting for. But it seems more likely that data contracts will be subsumed in a larger trend around data governance.

**In 2023, data governance will start shifting "left", and data standards will become a first-class citizen in orchestration tools.**

For decades, data governance has been an afterthought. It's often handled by data stewards, not data producers, who create documentation long after data is created.

However, we've recently seen a shift to move data governance "left", or closer to data producers. This means that whoever creates the data (usually a developer or engineer) must create documentation and check the data against predefined standards before it can go live.



Major tools have recently made changes that support this idea, and we expect to see even more in the coming year.

- dbt's [yaml files](#) and [Semantic Layer](#), where analytics engineers can create READMEs and define metrics while creating a dbt model
- Airflow's [Open Lineage](#), which tracks metadata about jobs and datasets as DAGs execute
- Fivetran's [Metadata API](#), which provides metadata for data synced by Fivetran connectors
- Atlan's [GitHub extension](#), which creates a list of downstream assets that will be affected by a pull request



**Alex Dean**  
@alexcrdean

It's technically not a Data Contract unless it comes from the Côte d'Ata region of France. Otherwise it's just a sparkling schematization.

♡ 60 11:32 AM - Oct 2, 2022



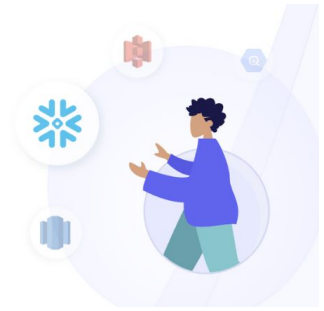
**@jamesdensmore@data-folks.masto.host**  
@jamesdensmore

I don't care if it's just part of the usual hype cycle, as long as the data contracts discussion gets us closer to source teams gaining awareness of the importance of downstream reporting and analytics.

♡ 41 9:56 AM - Sep 22, 2022



### 3. The semantic layer will enter “adoption” mode, albeit slowly



Also called a “metrics layer” or “business layer”, the semantic layer is an idea that’s been floating around the data world for decades.

The semantic layer is a literal term – it’s the “layer” in a data architecture that uses “semantics” (words) that the business user will understand.

Instead of raw tables with column names like “A000\_CUST\_ID\_PROD”, data teams build a semantic layer and rename that column “Customer”. Semantic layers hide complex code from business users while keeping it well-documented and accessible for data teams.

In our previous report, we talked about how companies were struggling to maintain consistent metrics across complex data ecosystems. Last year, we took a big leap forward.

**In October 2022, dbt Labs made a big splash at their annual conference by announcing their new Semantic Layer.**

This was a big deal, spawning excited tweets, in-depth think pieces, and celebrations from partners like us.

The core concept behind dbt’s Semantic Layer: define things once, use them anywhere.

Data producers can now define metrics in dbt, then data consumers can query those consistent metrics in downstream tools. Regardless of which BI tool they use, analysts and business users can look up a stat in the middle of a meeting, confident that their answer will be correct.

The Semantic Layer was a huge step forward for the modern data stack since it paves the way for metrics to become a first-class citizen.

Making metrics part of data transformation intuitively makes sense. Making them part of dbt — the dominant transformation tool, which is already well-integrated with the modern data stack — is exactly what the semantic layer needed to go from idea to reality.

## Our take on the future of the semantic layer...

Since dbt's Semantic Layer launched, progress has been fairly measured — in part because this happened less than three months ago.

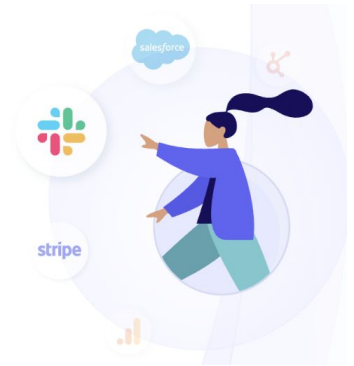
It's also because changing the way that people write metrics is *hard*. Companies can't just flip a switch and move to a semantic layer overnight. The change will take time, likely years rather than months.

**In 2023, the first set of Semantic Layer implementations will go live.**

Many data teams have spent the last couple of months exploring the impact of this new technology — experimenting with the Semantic Layer and thinking through how to change their metrics frameworks.

This process gets easier as more tools in the modern data stack integrate with the Semantic Layer. Seven tools were Semantic Layer-ready at its launch (including us, Hex, Mode, and Thoughtspot). Eight more tools were Metrics Layer-ready, an intermediate step to integrating with the Semantic Layer.

## 4. Data activation will replace CDPs as marketing spend becomes more important



This idea is related to reverse ETL, one of the big trends in last year's report. In 2022, some of the main players in reverse ETL worked to redefine and expand their category. Their latest buzzword is "data activation", a new take on the "customer data platform" (CDP).

A CDP combines data from all customer touchpoints (e.g. website, email, social media, help center, etc). A company can then segment or analyze that data, build customer profiles, and power personalized marketing. For example, they can create an automated email with a discount code if someone abandons their cart, or advertise to people who have visited a specific page on the website and used the company's live chat.

The key idea here is that CDPs are designed around *using* data, rather than simply aggregating and storing it — and this is where data activation comes in. As the argument goes, in a world where data is stored in a central data platform, why do we need standalone CDPs? Instead, we could just "activate" data from the warehouse to handle traditional CDP functions and diverse use cases across the company.

At its core, data activation is similar to reverse ETL, but instead of just sending data back to source systems, you're actively driving use cases with that data.

**We've been talking about data activation in various forms for the last couple of years. However, this idea of data activation as the new CDP took off in 2022.**

For example, Arpit Choudhury analyzed the space in April, Sarah Krasnik broke down the debate in July, Priyanka Somrah included it as a data category in August, and Luke Lin called out data activation in his 2023 data predictions last month.

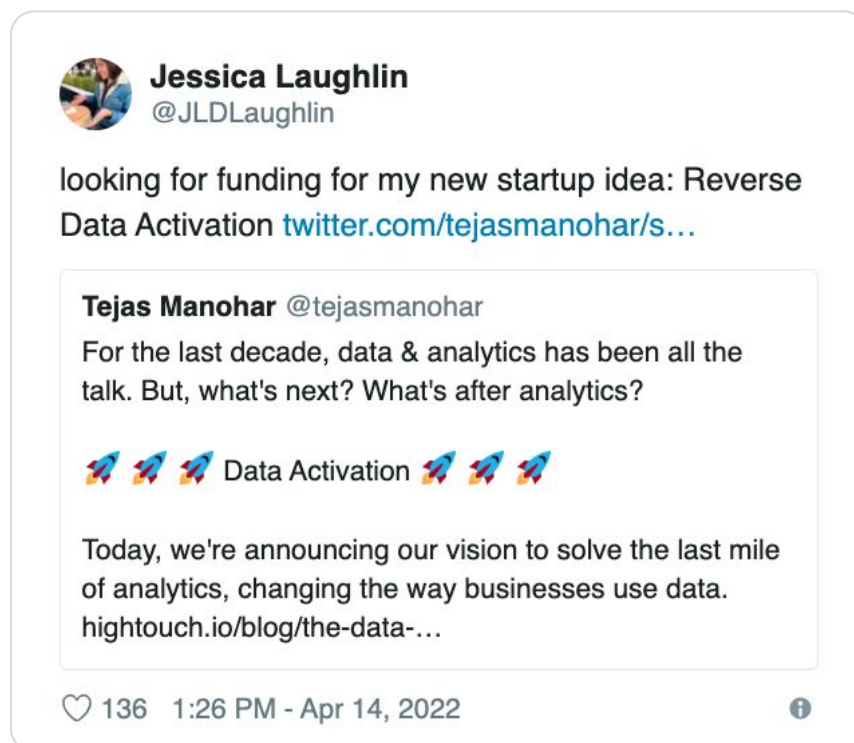
In part, this trend was caused by marketing from former reverse ETL companies, who now brand themselves as data activation products. (These companies still talk about reverse ETL, but it's now a feature within their data activation platform. Notably, Census has resisted this trend, retaining "reverse ETL" across its site.)

For example, Hightouch rebranded itself with a big splash in April, dropping three blogs on data activation in five days:

- Data Activation: The Next Step After Analytics by Pedram Navid
- Hightouch: The Data Activation Platform by Kashish Gupta
- What is Data Activation? by Luke Kline

In part, this can also be traced to the larger debate around driving data use cases and value, rather than focusing on data infrastructure or stacks. As Benn Stancil put it, "Why has data technology advanced so much further than value a data team provides?"

In part, this was also an inevitable result of the modern data stack. Stacks like Snowflake + Hightouch have the same data and functionality as a CDP, but they can be used across a company rather than for only one function.



**Jessica Laughlin** @JLDLaughlin

looking for funding for my new startup idea: Reverse Data Activation [twitter.com/tejasmanohar/s...](https://twitter.com/tejasmanohar/s...)

**Tejas Manohar** @tejasmanohar

For the last decade, data & analytics has been all the talk. But, what's next? What's after analytics?

🚀🚀🚀 Data Activation 🚀🚀🚀

Today, we're announcing our vision to solve the last mile of analytics, changing the way businesses use data. [hightouch.io/blog/the-data-...](https://hightouch.io/blog/the-data-...)

👍 136 1:26 PM - Apr 14, 2022

## Our take on the future of data activation...

CDPs made sense in the past. When it was difficult to stand up a data platform, having an out-of-the-box, perfectly customized customer data platform for business users was a big win.

Now, though, the world has changed, and companies can set up a data platform in under 30 minutes — one that not only has customer data, but also all other important company data (e.g. finance, product/users, partners, etc).

At the same time, data work has been consolidating around the modern data stack. Salesforce once tried to handle its own analytics (called Einstein Analytics). Now it has partnered with Snowflake, and Salesforce data can be piped into Snowflake just like any other data source.

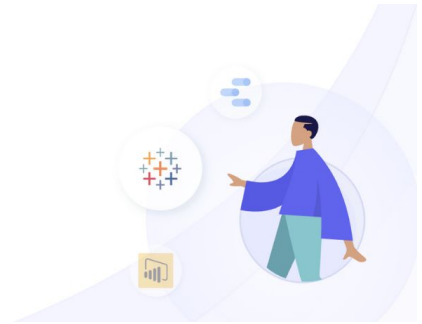
The same thing has happened for most SaaS products. While internal analytics was once their upsell, they are now realizing that it makes more sense to move their data into the existing modern data ecosystem. Instead, their upsell is now syncing data to warehouses via APIs.

**In this new world, data activation becomes very powerful. The modern data warehouse plus data activation will replace not only CDPs but also pre-built, specialized SaaS data platforms.**

With the modern data stack, data is now created in specialized SaaS products and piped into storage systems like Snowflake, where it is combined with other data and transformed in the API layer. Data activation is then crucial for piping insights back into the source SaaS systems where business users do their daily work.

For example, Snowflake acquired Streamlit, which allows people to create pre-built templates and templates on top of Snowflake. Rather than developing their own analytics or relying on CDPs, tools like Salesforce can now let their customers sync data to Snowflake and use a pre-built Salesforce app to analyze the data or do custom actions (like cleaning a lead list with Clearbit) with one click. The result is the customization and user-friendliness of a CDP, combined with the power of modern cloud compute.

## 5. The first wave of data mesh implementations will start going live



This idea came from Zhamak Dehghani — first with [two blogs](#) in 2019, and then with her [O'Reilly book](#) in 2022.

“The shortest summary: treat data as a product, not a by-product. By driving data product thinking and applying domain driven design to data, you can unlock significant value from your data. Data needs to be owned by those who know it best.”

— [Data Mesh Learning Community](#)

There are four pillars to the data mesh:

- **Domain-oriented data decentralization:** Rather than letting data live in a central data warehouse or lake, companies should move data closer to the people who know it best. The marketing team should own website data, RevOps should own finance data, and so on. Each domain would be responsible for its data pipelines, documentation, quality, and so on, with support from a centralized data team.
- **Data as a product:** Data teams should focus on building reusable, reproducible assets (with fundamental product components like SLAs) rather than getting stuck in the “service trap” of ad-hoc work.
- **Self-service data infrastructure:** Rather than one central data platform, companies should have a flexible data infrastructure platform where each data team can create and consume its own data products.
- **Federated computational governance:** Data assets need to work together even when data is distributed. While domain owners should have autonomy



over their data and its localized standards, there should also be a central “federation” of data leaders to create global rules and ensure the company’s data is healthy.

## The data mesh was everywhere in 2021. In 2022, it started to move from abstract idea to reality.

The data mesh conversation has shifted from “What is it?” to “How can we implement it?” As real user stories grew in places like the [Data Mesh Learning Community](#), the implementation debate split into two theories:

- **Via team structures:** Distributed, domain-based data teams are responsible for publishing data products, with support and infrastructure from a central data platforms team.
- **Via “data as a product”:** Data teams are responsible for creating data products — i.e. pushing data governance to the “left”, closer to data producers rather than consumers.

Meanwhile, companies have started branding themselves around the data mesh. So far, we’ve seen this with [Starburst](#), [Databricks](#), [Oracle](#), [Google Cloud](#), [Dremio](#), [Confluent](#), [Denodo](#), [Soda](#), [lakeFS](#), and [K2 View](#), among others.



**John Cutler**  
@johncutlefish

the opposite of data mesh is

data meh

producers do whatever they want, and some poor soul needs to make sense of it all so it can be consumed

♡ 72 10:35 AM - Aug 3, 2022





## Our take on the future of data mesh...

Four years after it was created, we're still in the early phases of the data mesh.

Though more people now believe in the concept, there's a lack of real operational guidance about how to achieve a data mesh.

Data teams are still figuring out what it means to implement the data mesh, and the mesh tooling stack is still premature. While there's been a lot of rebranding, we still don't have a best-in-class reference architecture of how a data mesh can be achieved.

**In 2023, we predict that the first wave of data mesh “implementations” will go live, with “data as a product” front and center.**

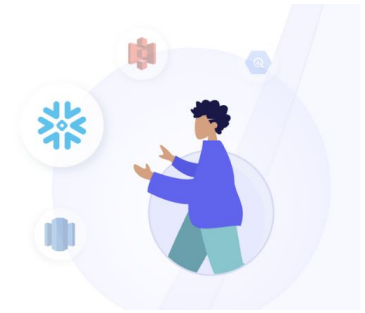
This year, we'll start seeing more and more real data mesh architectures — not the aspirational diagrams that have been floating around data blogs for years, but real architectures from real companies.

**We also expect that the data world will start to converge on a best-in-class reference architecture and implementation strategy for the data mesh.**

This will include the following core components:

- Metadata platform that can integrate into developer workflows (e.g. Atlan's [APIs](#) and GitHub [integration](#))
- Data quality and testing (e.g. [Great Expectations](#), [Monte Carlo](#))
- Git-like process for data producers to incorporate testing, metadata management, documentation, etc. (e.g. dbt)
- All built around the same central data warehouse/lakehouse layer (e.g. Snowflake, Databricks)

## 6. Data observability and quality will converge in a “data reliability” category



One of our big trends from last year, data observability has held its own and continued to grow alongside adjacent ideas like data quality and reliability.

All of these categories have grown significantly over the last year with existing companies getting bigger, new companies going mainstream, and new tools launching every month.

For example, in company news, Databand was acquired by IBM in July 2022. There were also some major Series Ds (Cribl with \$150M, Monte Carlo with \$135M, Unravel with \$50M) and Series Bs (Edge Delta with \$63M, Manta with \$35M) in this space.

In tooling news, Kensu launched a data observability solution, Anomalo launched the Pulse dashboard for data quality, Monte Carlo created a data reliability dashboard, Bigeye launched Metadata Metrics, AWS introduced observability features into Amazon Glue 4.0, and Entanglement spun out another company focused on data observability.

In the thought leadership arena, Monte Carlo and Kensu published major books with O’Reilly about data quality and observability.

In a notable change, this space also saw significant open-source growth in 2022.

Datafold launched an open-source diff tool, Acceldata open-sourced its data platform and data observability libraries, and Soda launched both its open-source Soda Core and enterprise Soda Cloud platforms.



**Sarah Catanzaro**

@sarahcat21

Are we turning a page? I sense a growing recognition among data Twitter that perhaps it's the data and not the tools, people, or titles that's the source of our collective grief?

♥ 68 7:58 AM - Jun 1, 2022



## Our take on the future of data observability, quality, and reliability...

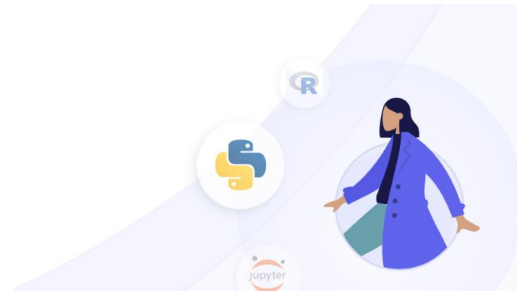
One of our open questions in last year's report was where data observability was heading — towards its own category, or merging with another category like data reliability or active metadata.

We think that data observability and quality will converge in a larger “data reliability” category centered around ensuring high-quality data.

This may seem like a big change, but it wouldn't be the first time this category has changed. It's been trying to settle on a name for several years. [Acceldata](#) started with logs observability but now brands itself as a data observability tool. After starting in the data quality space, [Soda](#) is now a major player in data observability. [Datafold](#) started with data diffs, but now calls itself a data reliability platform. The list goes on and on.

As these companies compete to define and own the category, we'll continue to see more confusion in the short term. However, we're seeing early signs that this will start to settle down into one category in the near future.

# Last thoughts



It feels interesting to welcome 2023 as data practitioners. While there's a lot of uncertainty looming in the air (uncertainty is the new certainty!), we're also a bit relieved.

## 2021 and 2022 were absurd years in the modern data stack.

The hype was crazy, new tools were launching every day, data people were constantly being poached by data startups, and VCs were throwing money at every data practitioner who even hinted at building something. The "modern data stack" was finally cool, and the data world had all the money and support and acknowledgment it needed.

At Atlan, we started as a data team ourselves. As people who have been in data for over a decade, this was a wild time. Progress is generally made in decades, not years. But in the last three years, the modern data stack has grown and matured as much as in the decade before.

It was exciting... yet we ended up asking ourselves existential questions more than once. Is this modern data stack thing real, or is it just hype fueled by VC money? Are we living in an echo chamber? Where are the data practitioners in this whole thing?

## While this hype and frenzy led to great tooling, it was ultimately bad for the data world.

Confronted by a sea of buzzwords and products, data buyers often ended up confused and could spend more time trying to get the right stack than actually using it.

Let's be clear — the goal of the data space is ultimately to help companies leverage data.

Tools are important for this. But they're ultimately an enabler, not the goal.

**As this hype starts to die down and the modern data stack starts to stabilize, we have the chance to take the tooling progress we've made and translate it into real business value.**

We're at a point where data teams aren't fighting to set up the right infrastructure. With the modern data stack, setting up a data ecosystem is quicker and easier than ever. Instead, data teams are fighting to prove their worth and get more results out of less time and resources.


Now that companies can't just throw money around, their decisions need to be targeted and data-driven. This means that data is more important than ever, and data teams are in a unique position to provide real business value.

But to make this happen, data teams need to finally figure out this "value" question.

**Now that we've got the modern data stack down, it's time to figure out the modern data culture stack.**

What does a great data team look like? How should it work with business? How can it drive the most impact in the least time?

These are tough questions, and there won't be any quick fixes. But if we can crack the secrets to a better data culture, we can finally create dream data teams — ones that will not just help their companies survive during the next 12-18 months, but propel them to new heights in the coming decades.

This report was created with  by Atlan. It was published in January 2023.

*Authors:*

Prukalpa Sankar (Co-Founder)  
Christine Garcia (Director of Content)

**atlan**

Stay in touch to get our latest updates:

   @AtlanHQ

 Metadata Weekly

---

The text, images, or a combination of both, as described in this material, cannot be copied, modified, published or distributed without prior written permission from Atlan (Peeply Technologies Pvt Ltd) and its respective authors.

The names, logos and brand marks of all data software, platform and tools other than Atlan's which are mentioned in this report are the properties of their respective owners. No copyright infringement is intended. Should there be any question or concern, you can write to [hello@atlan.com](mailto:hello@atlan.com).



The leading active metadata platform for modern data teams



Built by a data team for data teams, Atlan is the active metadata platform for DataOps. Our platform activates metadata to help data-driven enterprises discover, understand, trust, and collaborate on their data. With intelligent bots, column-level lineage, and personalized experiences, Atlan creates a single source of truth and brings context back into the tools where data teams live. Just three years after launch, Atlan is the tool of choice for a growing list of modern data teams around the world, including WeWork, Plaid, Postman, Scripps Health, TechStyle, Snapcommerce, and Delhivery.

### Pioneering the Active Metadata and DataOps categories



Named a **Leader** in the **Forrester Wave™**: Enterprise Data Catalogs for DataOps, Q2 2022



Recognized as a **Top 5 Global Innovator** in DataOps by IDC in 2022



Named in Gartner's inaugural **Market Guide for Active Metadata**, 3 Hype Cycles, and 7 reports in 2021



Recognized as a **Gartner Cool Vendor** in DataOps in 2020

### Deep partnerships and integrations across the modern data stack



First data catalog validated as a **Snowflake Ready Technology Partner**



Native integration with **Unity Catalog**, including column-level lineage



Named an AWS **Advanced Technology Partner** and Marketplace Seller



great\_expectations



SEE A DEMO

LEARN MORE