The Future of the Modern Data Stack in 2022

atlan

By Prukalpa Sankar & Christine Garcia

Introduction

Is it just us, or did data go through five years' worth of change in 2021? With so much hype and rapid change, it's hard to know what trends are here to stay and which will disappear just as quickly as they arose.

This guide breaks down the six ideas you should know about the modern data stack going into 2022 — the ones that exploded in the data world last year and don't seem to be going away.

atlan

This report was created with 💛 by Atlan. It was last updated in January 2022.

Authors:

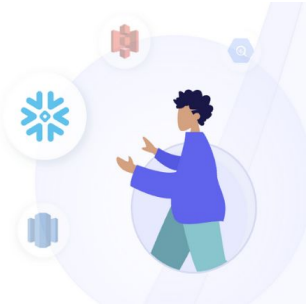
Prukalpa Sankar (Co-Founder)
Christine Garcia (Director of Content)

Stay in touch to get our latest updates:

@AtlanHQ

SIX BIG IDEAS FROM 2021

1. Data Mesh



You probably know this term by now, even you don't exactly know what it means.

The idea of the "data mesh" came from two 2019 blogs by Zhamak Dehghani, Director of Emerging Technologies at Thoughtworks:

1. How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh
2. Data Mesh Principles and Logical Architecture

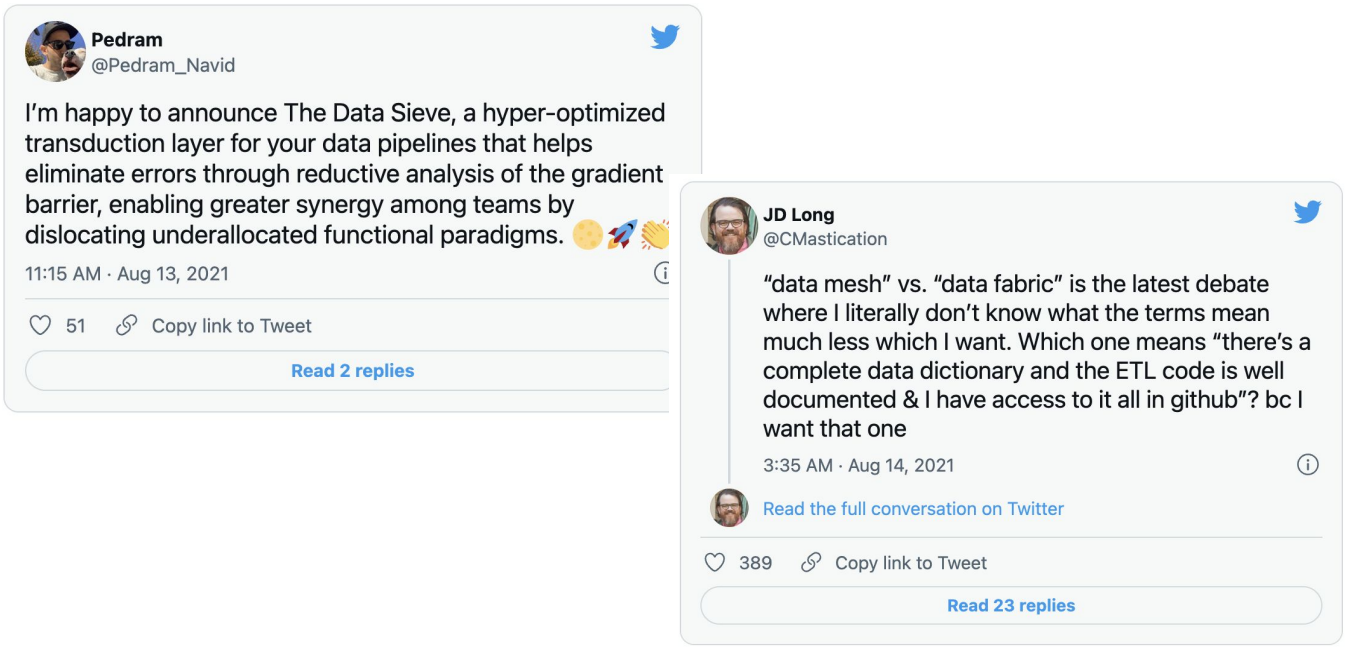
Its core idea is that companies can become more data-driven by moving from centralized data warehouses and lakes to a "domain-oriented decentralized data ownership and architecture" driven by self-serve data and "federated computational governance".

As you can see, the language around the data mesh gets complex fast, which is why there's no shortage of "what actually is a data mesh?" articles.

The idea of the data mesh has been quietly growing since 2019, until suddenly it was everywhere in 2021.

The Thoughtworks Techology Radar moved Data Mesh's status from "Trial" to "Assess" in just one year. The Data Mesh Learning Community launched, and their Slack group got over 1,500 signups in 45 days. Zalando started doing talks about how it moved to a data mesh.

Soon enough, hot takes were flying back and forth on Twitter, with data leaders arguing over whether the data mesh is revolutionary or ridiculous.



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

In 2022, we think we'll see a ton of platforms rebrand and offer their services as the "ultimate data mesh platform".

But the thing is, the data mesh isn't a platform or a service that you can buy off the shelf. It's a design concept with some wonderful concepts like distributed ownership, domain-based design, data discoverability, and data product shipping standards — all of which are worth trying to operationalize in your organization.

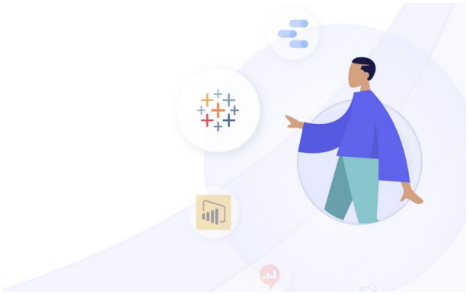
So here's our advice: As data leaders, it is important to stick to the first principles at a conceptual level, rather than buy into the hype that you'll inevitably see in the market soon.

We wouldn't be surprised if some teams (especially smaller ones) can achieve the data mesh architecture through a fully centralized data platform built on Snowflake and dbt, whereas others will leverage the same principles to consolidate their "data mesh" across complex multi-cloud environments.

SIX BIG IDEAS FROM 2021

2. Metrics Layer

aka "headless BI" or "metrics store"



Metrics are critical to assessing and driving a company's growth, but they've been struggling for years. They're often split across different data tools, with different definitions for the same metric across different teams or dashboards.

In 2021, people finally started talking about how the modern data stack could fix this issue. It's been called the metrics layer, metrics store, headless BI, and even more names than we can list here.

It started in January, when Base Case proposed "Headless Business Intelligence", a new approach to solving metrics problems. A couple months later, Benn Stancil from Mode talked about the "missing metrics layer" in today's data stack.

That's when things really took off. Four days later, Mona Akmal and Aakash Kambuj from Falkon published articles about making metrics first-class citizens and the "modern metrics stack".

Two days after that, Airbnb announced that it had been building a home-grown metrics platform called "Minerva" to solve this issue. Other prominent tech companies soon followed suit, including LinkedIn's Unified Metrics Platform, Uber's uMetric, and Spotify's metrics catalog in their "new experimentation platform".

Just when we thought this fervor had died down, Drew Banin (CPO and Co-Founder of dbt) opened a PR on dbtcore in October. He hinted that dbt would be incorporating a metrics layer into its product, and even included links to those foundational blogs by Benn and Base Case.

The PR blew up and reignited the discussion around building a better metrics layer in the modern data stack.

Meanwhile, a bunch of early stage startups have launched to compete for this space. Transform is probably the biggest name so far, but Metriql, Lightdash, Supergrain, and Metlo also launched this year. Some bigger names are also pivoting to compete in the metrics layer, such as GoodData's foray into Headless BI.

Our take on the future of the metrics layer...

We are extremely excited about the metrics layer finally becoming a thing. A few months ago, George from Fivetran surfaced an unpopular opinion that all metrics stores will evolve into BI tools.

While we don't fully agree, we do believe that a metrics layer that isn't tightly integrated with BI is unlikely to ever become commonplace.

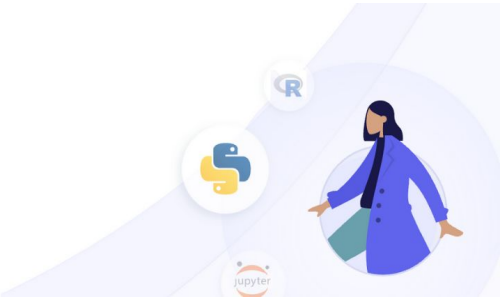
However, existing BI tools aren't really incentivized to integrate an external metrics layer into their tools... which makes this a chicken and egg problem. Standalone metrics layers will struggle to encourage BI tools to adopt their frameworks, and will be forced to build BI like Looker was forced to many years ago.

This is why we're really excited about dbt announcing their foray into the metrics layer. dbt already has enough distribution to encourage at least the modern BI tools (e.g. Preset, Mode, Thoughtspot) to integrate deeply into the dbt metrics API, which may create competitive pressure for the larger BI players.

We also think that metrics layers are so deeply intertwined with the transformation process that intuitively this makes sense.

Our prediction is that we'll see metrics become a first-class citizen in more transformation tools in 2022.

3. Reverse ETL



For years, ETL (Extract, Transform, Load) was how data teams populated their systems. First, they'd pull data from third-party systems, clean it up, and then load it into their warehouses.

This was great because it kept data warehouses clean and orderly, but it also meant that it took forever to get data into warehouses. Sometimes, data teams just wanted to dump raw data into their systems and deal with it later.

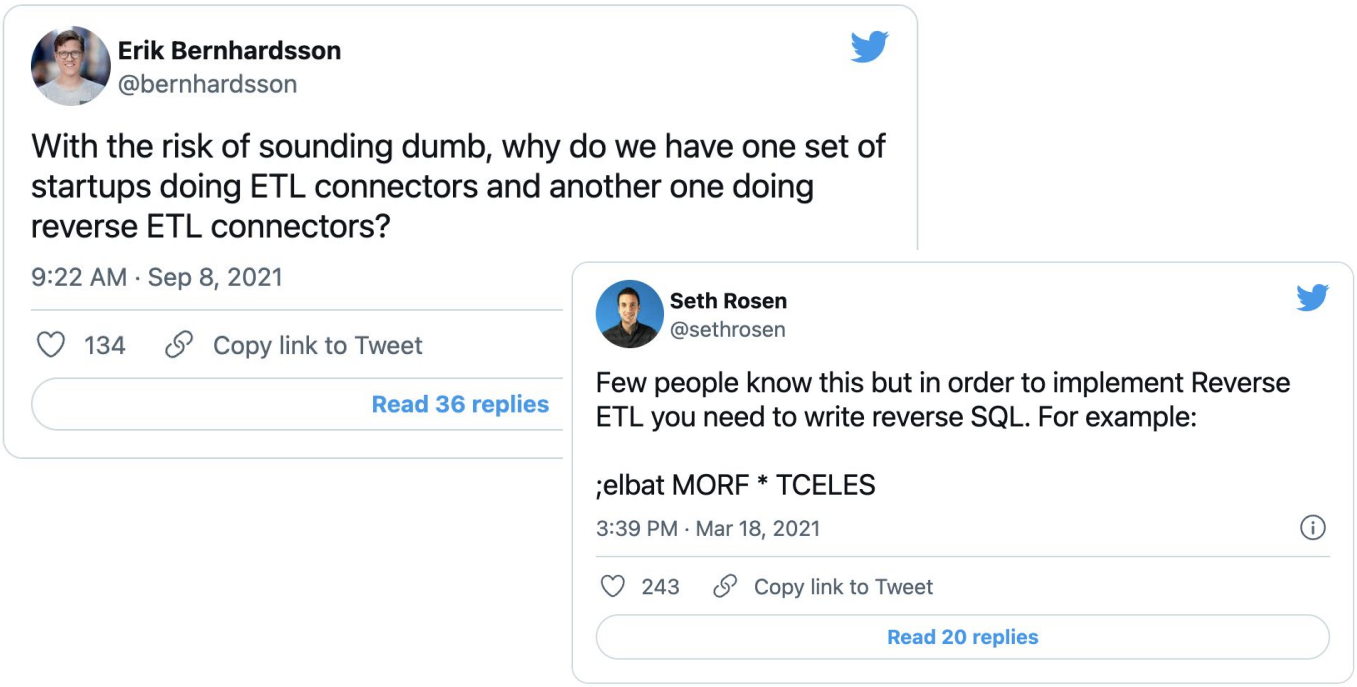
That's why many companies moved from ETL to ELT (Extract, Load, Transform) a couple of years ago. Instead of transforming data first, companies would send raw data into a data lake, then transform it later for a specific use case or problem.

In 2021, we got another major evolution in this idea — reverse ETL.

This concept first started getting attention in February, when Astasia Myers (Founding Enterprise Partner at Quiet Capital) wrote an article about the emergence of reverse ETL.

Since then, Hightouch and Census (both of which launched in December 2020) have set off a firestorm as they've battled to own the reverse ETL space. Census announced that it raised a $16 million Series A in February and published a series of benchmarking reports targeting Hightouch. Hightouch countered with three raises of a total $54.2 million in less than 12 months.

Hightouch and Census have dominated the reverse ETL discussion this year, but they're not the only ones in the space. Other notable companies are Grouparoo, HeadsUp, Polytomic, Rudderstack, and Workato (who closed a $200m Series E in November). Seekwell even got acquired by Thoughtspot in March.



Our take on the future of reverse ETL...

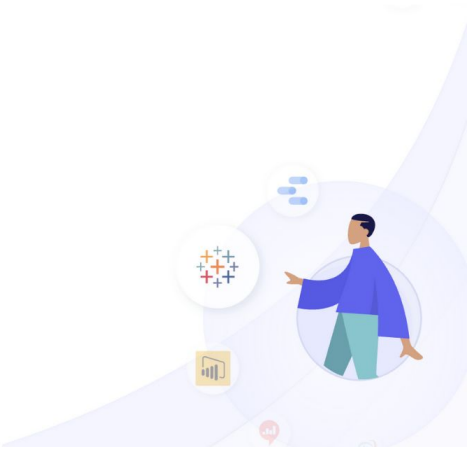
We're pretty excited about everything that's solving the "last mile" problem in the modern data stack.

We're now talking more about how to use data in daily operations than how to warehouse it — that's an incredible sign of how mature the fundamental building blocks of the data stack (warehousing, transformation, etc) have become!

What we're not so sure about is whether reverse ETL should be its own space or just be combined with a data ingestion tool, given how similar the fundamental capabilities of piping data in and out are.

Players like Hevodata have already started offering both ingestion and reverse ETL services in the same product, and we believe that we might see more consolidation (or deeper go-to-market partnerships) in the space soon.

4. Active Metadata & Third-Gen Data Catalogs



In the last couple of years, the debate around data catalogs was, "Are they obsolete?" And it would be easy to think the answer is yes.

In a couple of well-known articles, Barr Moses argued that data catalogs were dead, and Michael Kaminsky argued that we don't need data dictionaries.

On the other hand, there's never been so much buzz about data catalogs and metadata. There are so many data catalogs that Rohan from our team created thedatacatalog.com, a "catalog of catalogs", which feels both ridiculous and completely necessary. So which is it — are data catalogs dead or stronger than ever?

This year, data catalogs got new life with the creation of two new concepts — third-generation data catalogs and active metadata.

At the beginning of 2021, we wrote an article on modern metadata for the modern data stack. It introduced the idea that we're entering the third-generation of data catalogs, a fundamental transformation from the prevalent old-school, on-premise data catalogs.

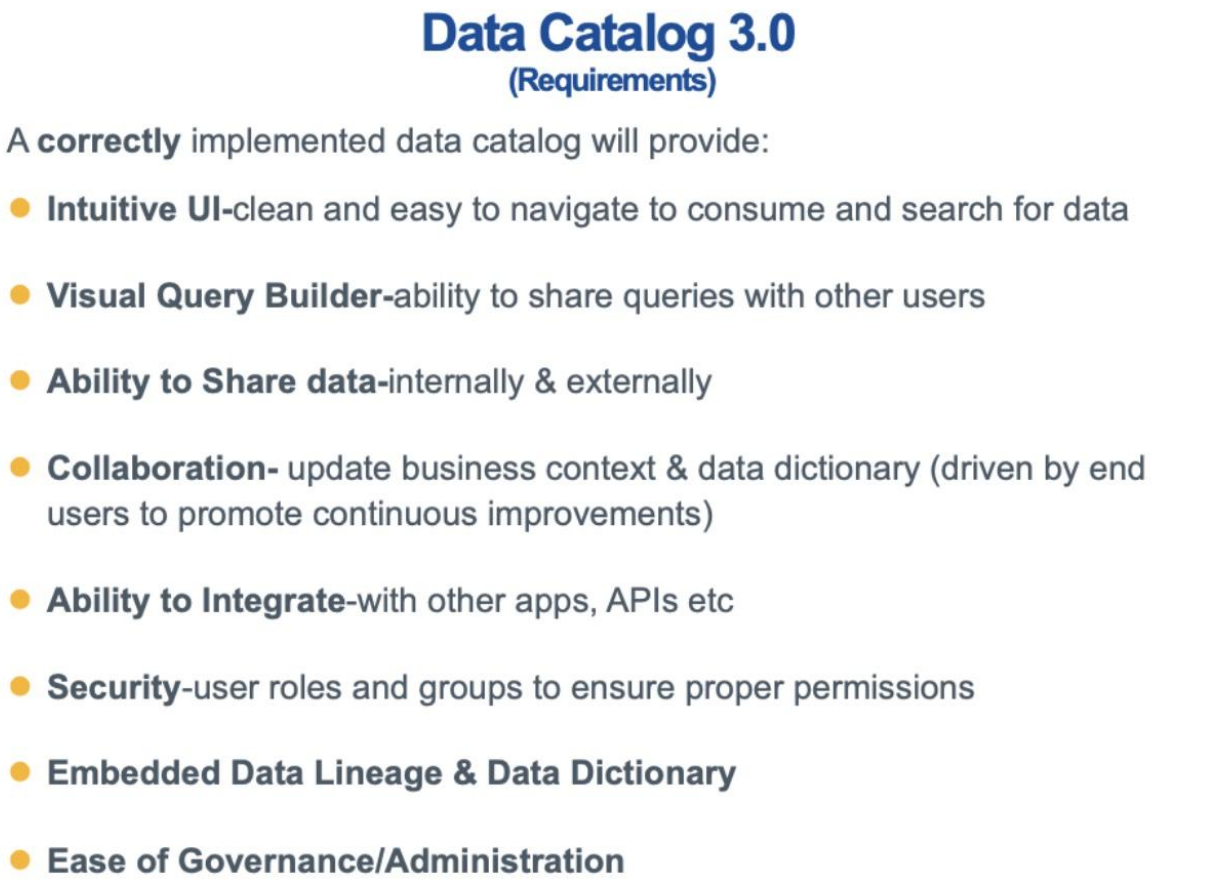
These new data catalogs are built around diverse data assets, "big metadata", end-to-end data visibility, and embedded collaboration.

This idea got amplified by a huge move Gartner made this year — scrapping its Magic Quadrant for Metadata Management Solutions and replacing it with the Market Guide for Active Metadata. In doing this, they introduced "active metadata" as a new category in the data space.

What's the difference? Old-school data catalogs collect metadata and bring them into a siloed "passive" tool, aka the traditional data catalog.

Active metadata platforms act as two-way platforms. They not only bring metadata together into a single store like a metadata lake, but also leverage "reverse metadata" to make metadata available in daily workflows.

Since the first time we wrote about third-generation catalogs, they've become part of the discourse around what it means to be a modern data catalog. We even saw the terms pop up in RFPs!



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

At the same time, VCs have been eager to invest in this new space. Metadata management has grown a ton with raises across the board — e.g. Collibra's $250m Series G, Alation's $110m Series D, and our $16m Series A. Seed-stage companies like Stemma and Acryl Data also launched to build managed metadata solutions on existing open-source projects.

Our take on the future of third-gen data catalogs...

The data world will always be diverse, and that diversity of people and tools will always lead to chaos.

We're probably biased, given that we've dedicated our lives to building a company in the metadata space. But we truly believe that the key to bringing order to the chaos that is the modern data stack lies in how we can use and leverage metadata to create the modern data experience.

Gartner summarized the future of this category in a single sentence:

> "The stand-alone metadata management platform will be refocused from augmented data catalogs to a metadata 'anywhere' orchestration platform."

Where data catalogs in the 2.0 generation were passive and siloed, the 3.0 generation is built on the principle that context needs to be available wherever and whenever users need it. Instead of forcing users to go to a separate tool, third-gen catalogs will leverage metadata to improve existing tools like Looker, dbt, and Slack, finally making the dream of an intelligent data management system a reality.

While there's been a ton of activity and funding in the space in 2021, we're quite sure we'll see the rise of a dominant and truly third-gen data catalog (aka an active metadata platform) in 2022.

5. Data Teams as Product Teams

As the modern data stack goes mainstream and data becomes a bigger part of daily operations, data teams are evolving to keep up. They're no longer "IT folks", working separately from the rest of the company.

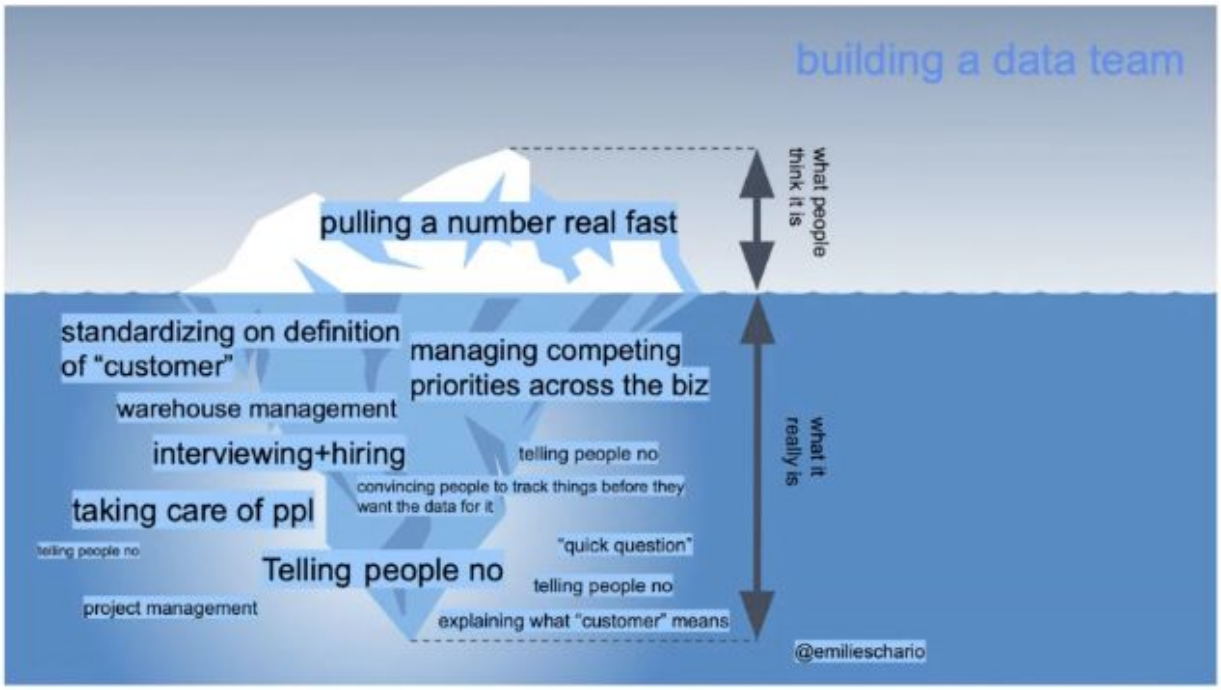
But this raises the question, how should data teams work with the rest of the company? Too often, they get stuck in the "service trap" — never-ending questions and requests for creating stats, rather than generating insights and driving impact through data.

In 2021, Emilie Schario from Amplify Partners, Taylor Murphy from Meltano, and Eric Weber from Stitch Fix talked about a way to break data teams out of this trap — rethinking data teams as product teams. They first explained this idea with a blog on Locally Optimistic, followed by great talks at conferences like MDSCON, dbt Coalesce, and Future Data.

A product isn't measured on how many features it has or how quickly engineers can quash bugs — it's measured on how well it meets customers' needs.

Similarly, data product teams should be centered on the users (i.e. data consumers throughout the company), rather than questions answered or dashboards built. This allows data teams to focus on experience, adoption, and reusability, rather than ad-hoc questions or requests.

This focus on breaking out of the service trap and reorienting data teams around their users really resonated with the data world this year. More people have started talking about what it means to build "data product teams", including plenty of hot takes on who to hire and how to set goals.



Emilie Schario's iconic image on the reality of working on a data team.

Our take on the future of data teams as product teams...

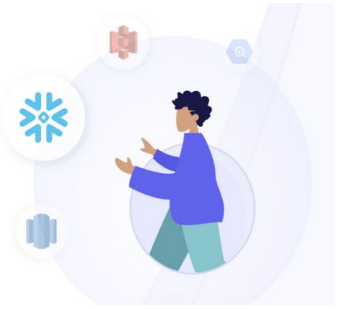
Of all the hyped trends in 2021, this is the one we're most bullish on.

We believe that in the next decade, data teams will emerge as one of the most important teams in the organization fabric, powering the modern, data-driven companies at the forefront of the economy.

However, the reality is that data teams today are stuck in a service trap, and only 27% of their data projects are successful.

We believe the key to fixing this lies in the concept of the "data product" mindset, where data teams focus on building reusable, reproducible assets for the rest of the team. This will mean investing in user research, scalability, data product shipping standards, documentation, and more.

6. Data Observability



This idea came out of "data downtime", which Barr Moses from Monte Carlo first spoke about in 2019 saying, "Data downtime refers to periods of time when your data is partial, erroneous, missing or otherwise inaccurate".

It's those emails you get the morning after a big project, saying "Hey, the data doesn't look right..."

Data downtime has been a part of normal life on a data team for years. But now, with many companies relying on data for literally every aspect of their operations, it's a huge deal when data stops working. Yet everyone was just reacting to issues as they cropped up, rather than proactively preventing them.

This is where data observability — the idea of "monitoring, tracking, and triaging of incidents to prevent downtime" — came in.

We still can't believe how quickly data observability has gone from being just an idea to a key part of the modern data stack.

As it's evolved quickly, this category even started being called "data reliability" or "data reliability engineering".

The space went from being non-existent to hosting a bunch of companies, with a collective $200m of funding raised in 18 months.

This includes Acceldata, Anomalo, Bigeye, Databand, Datafold, Metaplane, Monte Carlo, and Soda. People even started creating lists of new "data observability companies" to help keep track of the space.

Our take on the future of data observability...

We believe that in the past two years, data teams have realized that tooling to improve productivity is not a good-to-have but a must-have. After all, data professionals are one of the most sought-after hires you will ever make, so they shouldn't be wasting their time on troubleshooting pipelines.

So will data observability be a key part of the modern data stack in the future? Absolutely.

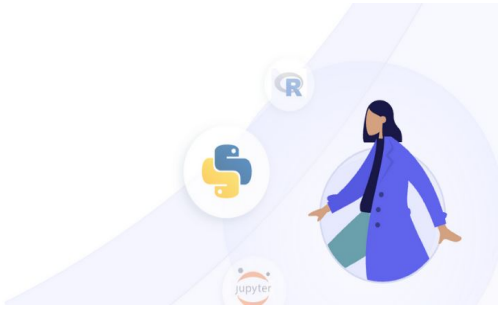
But will data observability continue to exist as its own category or will it be merged into a broader category (like active metadata or data reliability)? This is what we're not so sure about.

Ideally, if you have all your metadata in one open platform, you should be able to leverage it for a variety of use cases (like data cataloging, observability, lineage and more). We wrote about that idea last year in an article on the metadata lake.

That being said, today, there's a ton of innovation that these spaces need independently.

Our sense is that we'll continue to see fragmentation in 2022 before we see consolidation in the years to come.

Last thoughts



It may feel chaotic and crazy at times, but today is a golden age of data.

In the last eighteen months, our data tooling has grown exponentially. We all make a lot of fuss about the modern data stack, and for good reason — it's so much better than what we had before.

The earlier data stack was frankly as broken as broken could get, and this gigantic leap forward in tooling is exactly what data teams needed.

In our opinion, the next "delta" on the horizon for the data world is the modern data culture stack — the best practices, values, and cultural rituals that will help us diverse humans of data collaborate effectively and up our productivity as we tackle our new data stacks.

However, we can only think about working together better with data after we've nailed, well, working with data.

We're on the cusp of getting the modern data stack right, and we can't wait to see what new developments and trends 2022 will bring!