atlan

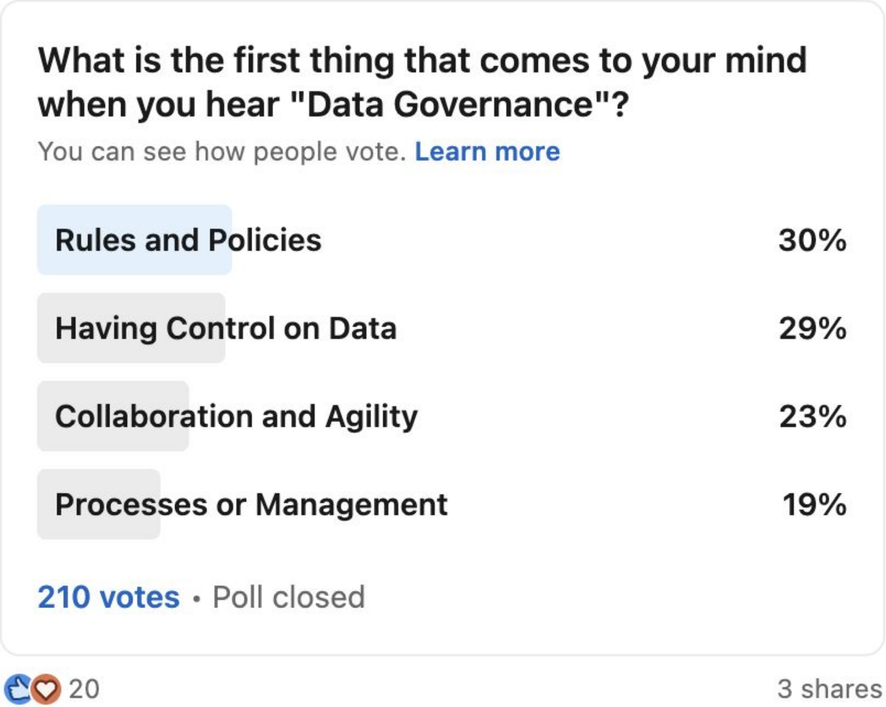
Data Governance for the Lakehouse Era: Community-Led and Bottom-Up

Prukalpa Sankar
Cofounder, Atlan

Data governance is pretty misunderstood…

MYTH #1

Data governance is all about control and rules



What is the first thing that comes to your mind when you hear "Data Governance"?

You can see how people vote. Learn more

Rules and Policies	30%
Having Control on Data	29%
Collaboration and Agility	23%
Processes or Management	19%

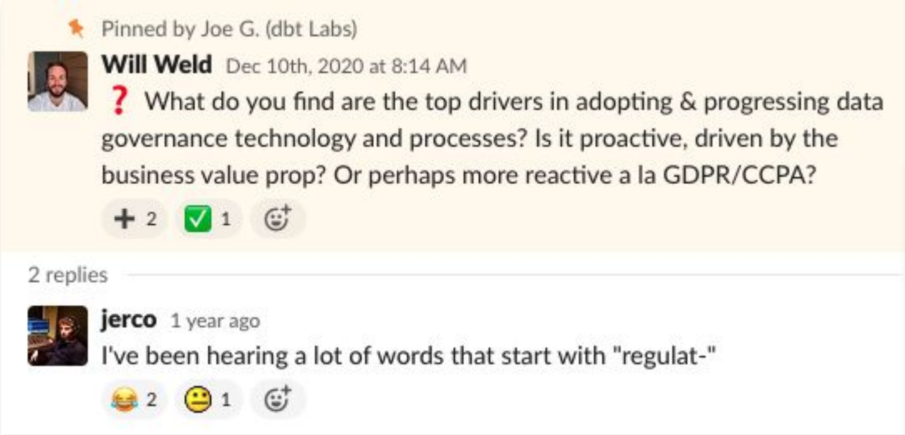
210 votes • Poll closed

20

3 shares

MYTH #2

Data governance = data security or protection



Pinned by Joe G. (dbt Labs)

Will Weld Dec 10th, 2020 at 8:14 AM

❓ What do you find are the top drivers in adopting & progressing data governance technology and processes? Is it proactive, driven by the business value prop? Or perhaps more reactive a la GDPR/CCPA?

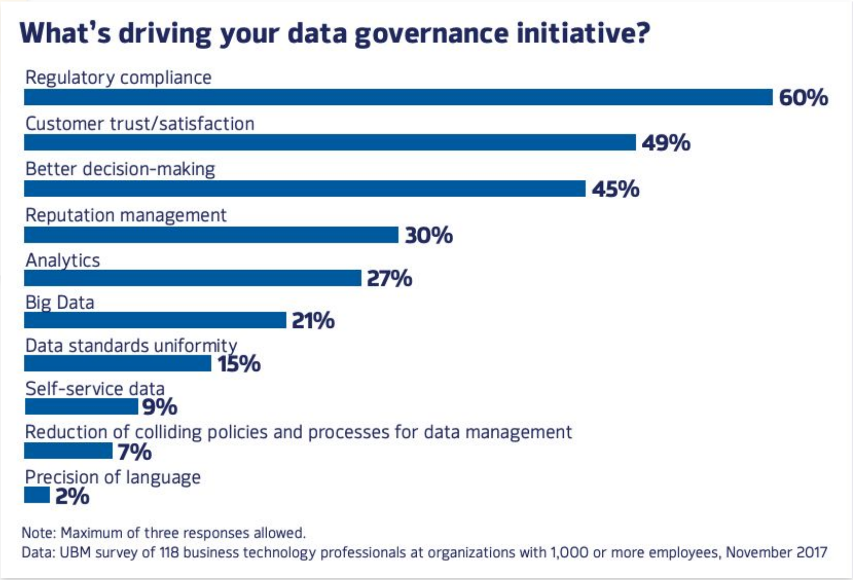
➕ 2 ✅ 1

2 replies

jerco 1 year ago

I've been hearing a lot of words that start with "regulat-"

😂 2 😕 1



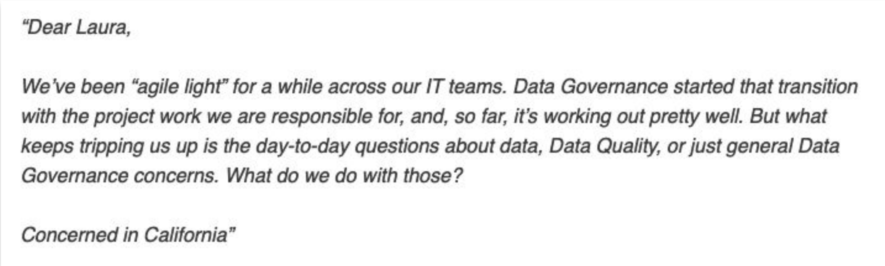
What's driving your data governance initiative?

Note: Maximum of three responses allowed.
Data: UBM survey of 118 business technology professionals at organizations with 1,000 or more employees, November 2017

Graph from Erwin & Dataversity, "The 2018 State of Data Governance"

MYTH #3

Data governance slows you down

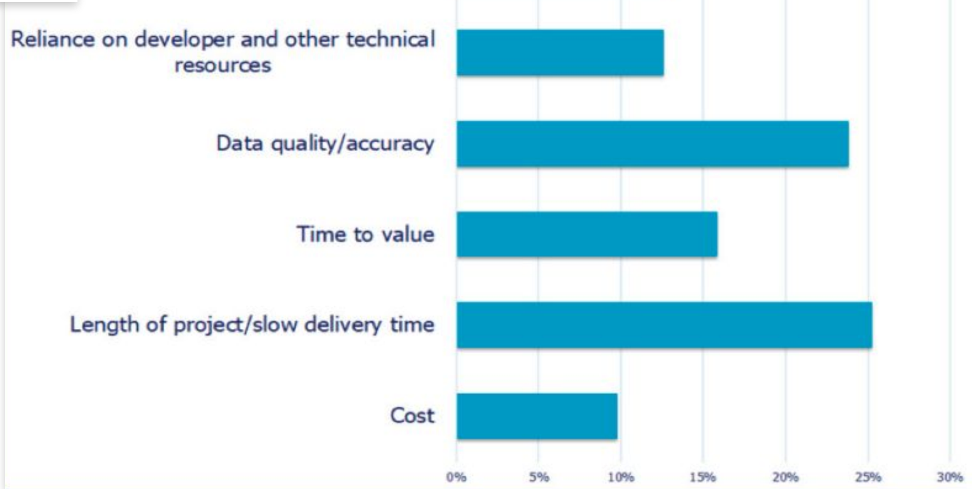


"Dear Laura,

We've been "agile light" for a while across our IT teams. Data Governance started that transition with the project work we are responsible for, and, so far, it's working out pretty well. But what keeps tripping us up is the day-to-day questions about data, Data Quality, or just general Data Governance concerns. What do we do with those?

Concerned in California"

What is the most significant challenge to your organization's data preparation/ data governance/ data intelligence efforts?



Graph from Erwin & Dataversity, "The 2020 State of Data Governance and Automation"

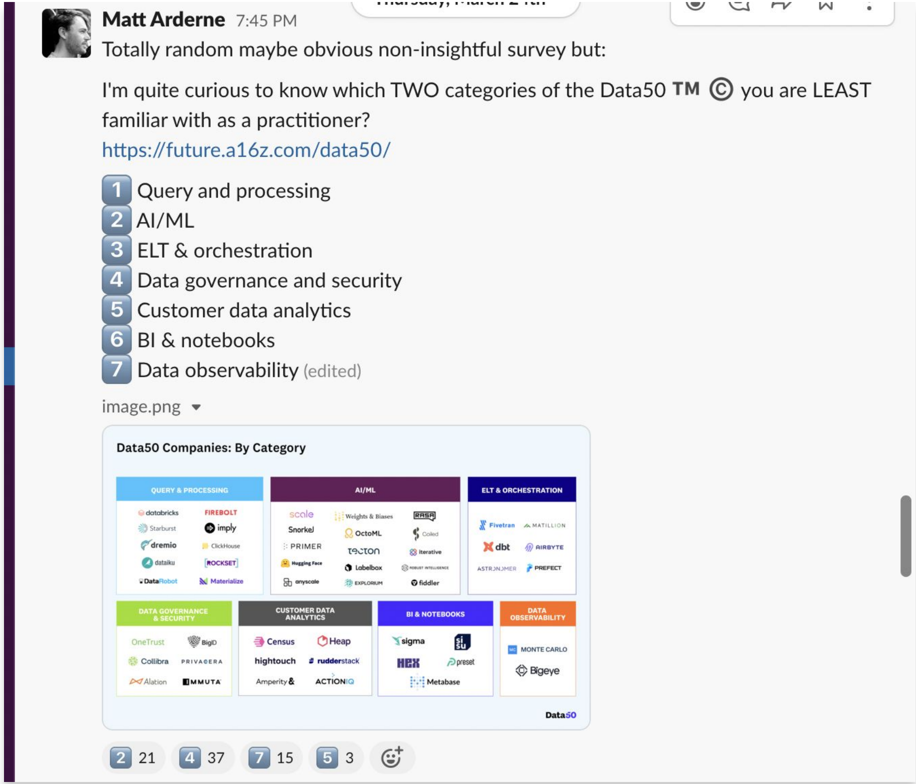
People see data governance as a **monarchy** 👑 with bureaucratic, ineffective rules dropped down from on high…

> “**Most governance programs today are ineffective**. The issue frequently starts at the top, with a C-suite that doesn’t recognize the value-creation potential in data governance.
>
> As a result, it becomes a set of policies and guidance **relegated to a support function** executed by IT and **not widely followed** — rendering the initiatives that data powers equally ineffective.”

McKinsey Digital

“Designing data governance that delivers value” (June 2020)

Or they don’t even know what data governance actually is 🤷‍♂️



Matt Arderne 7:45 PM

Totally random maybe obvious non-insightful survey but:

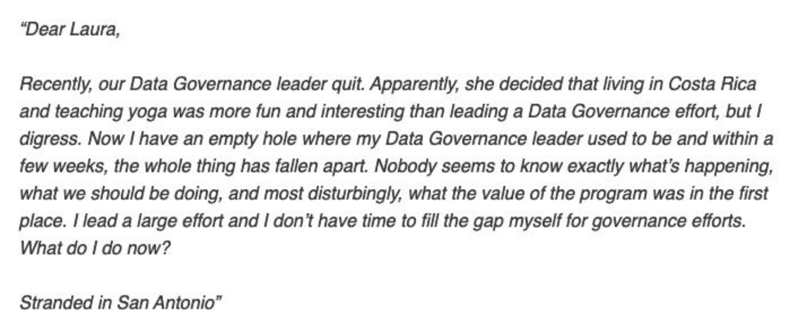
I'm quite curious to know which TWO categories of the Data50 TM © you are LEAST familiar with as a practitioner?

https://future.a16z.com/data50/

1 Query and processing
2 AI/ML
3 ELT & orchestration
4 Data governance and security
5 Customer data analytics
6 BI & notebooks
7 Data observability (edited)

image.png ▾

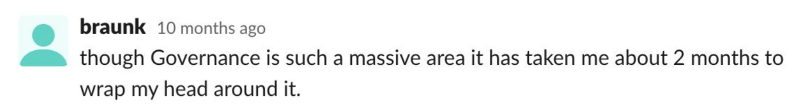
2 21 4 37 7 15 5 3



“Dear Laura,

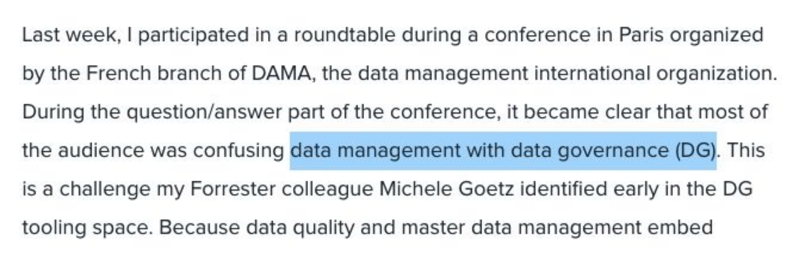
Recently, our Data Governance leader quit. Apparently, she decided that living in Costa Rica and teaching yoga was more fun and interesting than leading a Data Governance effort, but I digress. Now I have an empty hole where my Data Governance leader used to be and within a few weeks, the whole thing has fallen apart. Nobody seems to know exactly what’s happening, what we should be doing, and most disturbingly, what the value of the program was in the first place. I lead a large effort and I don’t have time to fill the gap myself for governance efforts. What do I do now?

Stranded in San Antonio”



braunk 10 months ago

though Governance is such a massive area it has taken me about 2 months to wrap my head around it.



Last week, I participated in a roundtable during a conference in Paris organized by the French branch of DAMA, the data management international organization. During the question/answer part of the conference, it became clear that most of the audience was confusing data management with data governance (DG). This is a challenge my Forrester colleague Michele Goetz identified early in the DG tooling space. Because data quality and master data management embed

I’ll admit it…

For a while, I had no idea what data governance was.

Hi, I’m Prukalpa 👋

Lifelong data practitioner

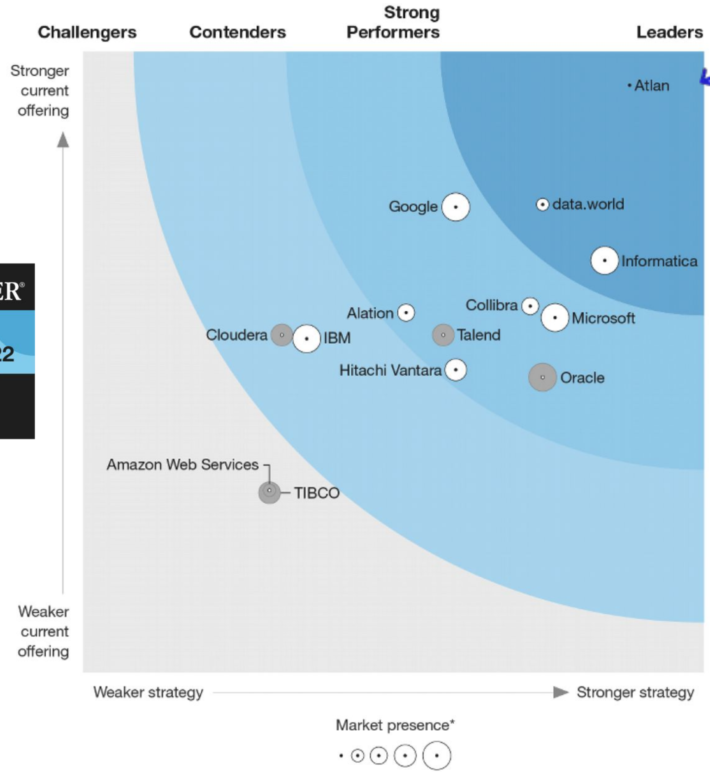
Co-founder of atlan

Pioneering the **Active Metadata** category, the “collaboration layer” for the modern data stack

Powering teams like **Postman, Plaid, WeWork, News Corp, Unilever, and Juniper**

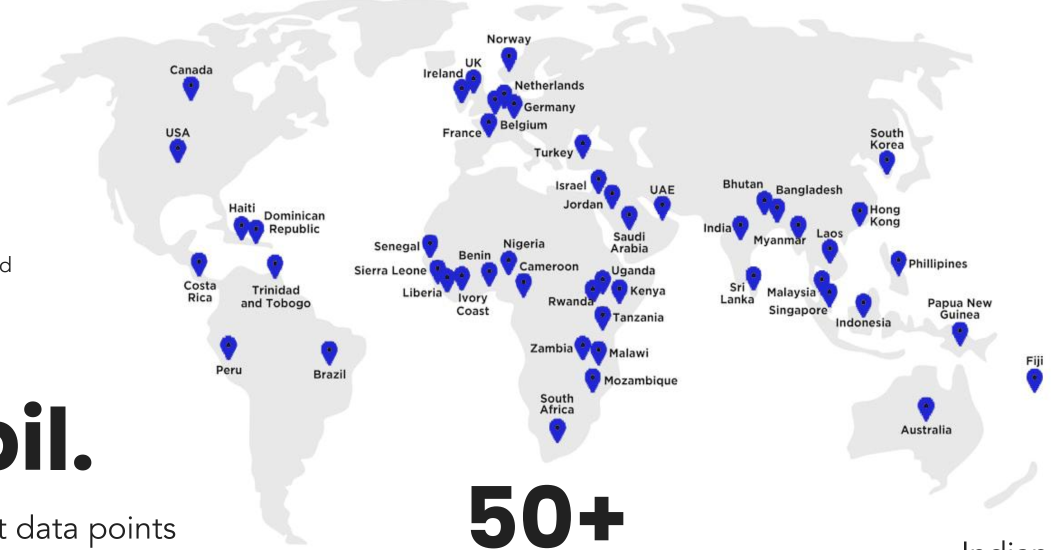
Named a **Gartner Cool Vendor** in DataOps and **IDC Innovator**

FORRESTER®
WAVE LEADER 2022
Enterprise Data Catalogs For DataOps



Launch partner for **Databricks Unity Catalog**

We started as a data team ourselves using **data science** for **social good**



110 bil.

external data points

ingested, cleaned, and visualized

6.5 bil.

satellite imagery

pixels processed

1.5 bil.

government data points

aggregated in real-time

50+

countries

with a diverse set of organizations

500 mil.

Indian citizens' data processed

THE WORLD BANK

United Nations

BILL & MELINDA GATES foundation

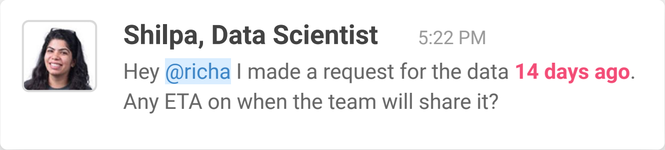
भारत सरकार

Government of INDIA

Every day was **chaos**. I didn't know this had anything to do with **governance**.

#team-datascience

Data discovery

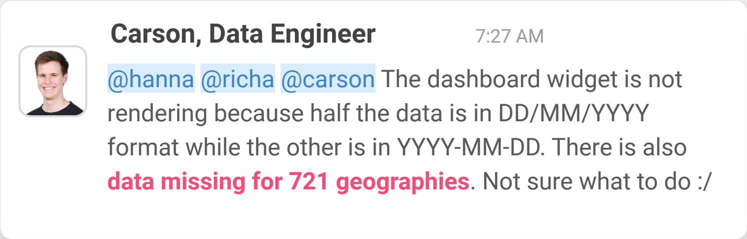


Shilpa, Data Scientist 5:22 PM

Hey @richa I made a request for the data **14 days ago**. Any ETA on when the team will share it?

#team-frontend

Data visibility

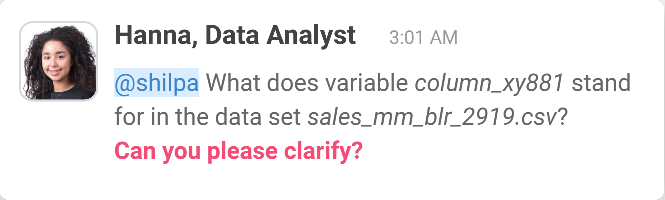


Carson, Data Engineer 7:27 AM

@hanna @richa @carson The dashboard widget is not rendering because half the data is in DD/MM/YYYY format while the other is in YYYY-MM-DD. There is also **data missing for 721 geographies**. Not sure what to do :/

Human tribal knowledge

Private Chat

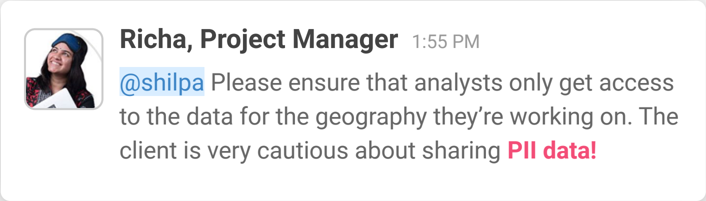


Hanna, Data Analyst 3:01 AM

@shilpa What does variable *column_xy881* stand for in the data set *sales_mm_blr_2919.csv*? **Can you please clarify?**

Data governance

#project-gb-data



Richa, Project Manager 1:55 PM

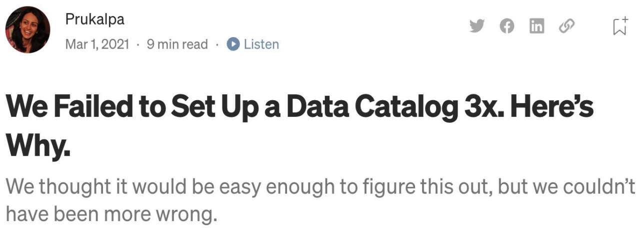
@shilpa Please ensure that analysts only get access to the data for the geography they're working on. The client is very cautious about sharing **PII data!**

The dreaded question… 😓

“That number doesn’t look right…”

That’s how we started the **Assembly Line Project**.

We tried to buy a solution.



Prukalpa

Mar 1, 2021 · 9 min read · Listen

We Failed to Set Up a Data Catalog 3x. Here’s Why.

We thought it would be easy enough to figure this out, but we couldn’t have been more wrong.

Building The World's Largest Government Data Lake - DISHA Platform

Our team became **6X more agile**.

12
months to build

8
member team

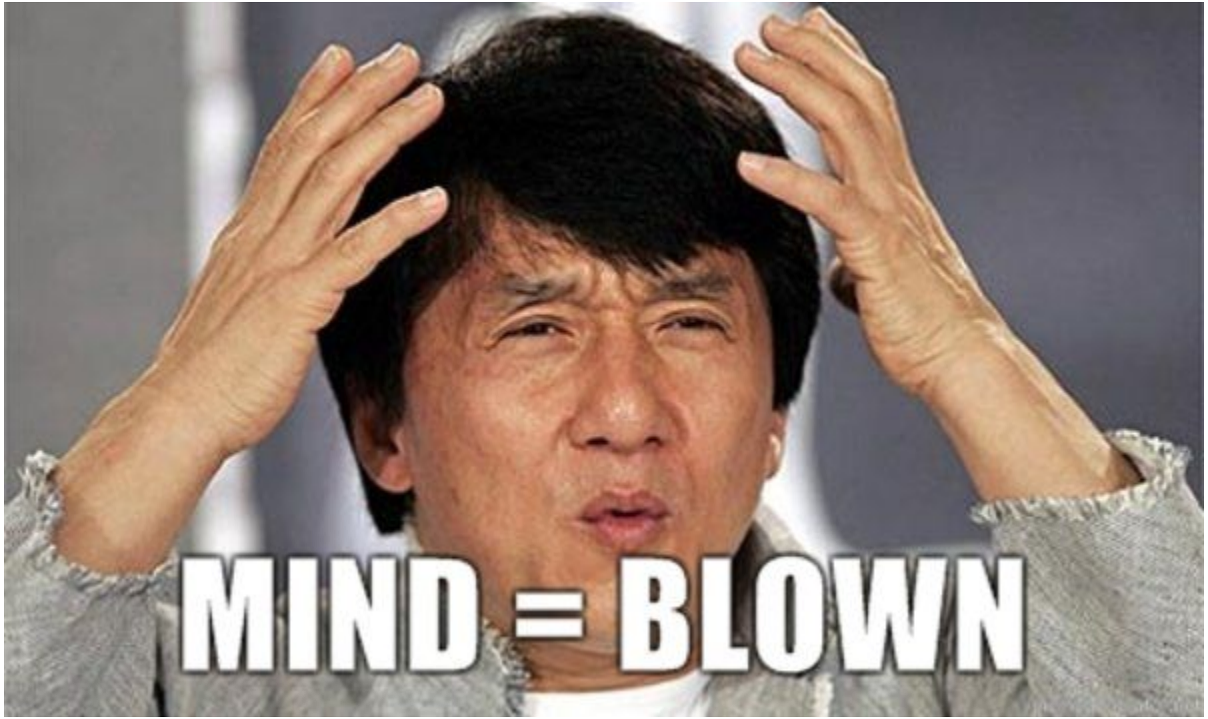
12
master data hierarchies

3.5b
dynamic data points

42
data portals connected

After a demo someone told me,

“Oh, you are a modern data governance tool!”



Data governance is changing…

… because our world is changing

TREND #1

People expect purpose and autonomy in their work

"With Generation Y coming into the business, **hierarchies have to disappear**. Generation Y expects to work in **communities of mutual interest and passion** — not structured hierarchies. Consequently, people management strategies will have to change so they look… less like the pyramid structures we are used to."

–Vineet Nayar, Vice Chairman and CEO, HCL Technologies

pwc "Millennials at work: Reshaping the workplace." PwC, 2011.

"Today's young workers have shifted toward interests in doing valuable work and finding meaning in their day-to-day job functions.

Leaders and managers are the ones who have the power to help foster that connection of meaningful work…

There's a giant risk for employers if they don't help employees have a sense of purpose and a sense of well-being and engagement."

University of Missouri "The Effect of Respect." LaGree, Houston, Duffy, Shin. Sage Journals, May 2021.

TREND #2

Consumerization of enterprise

aka: people expect work tech to be as cool as personal tech

> “It used to be enough to provide tools that improved productivity for the business in a measurable manner like enhancing processes and workflows.
>
> Now, every tool that is used by employees must provide a world-class user experience. **Employees will not adopt tools without a memorable experience.**
>
> These tools and technologies will eventually be phased out of the organization because enterprise budget-holders cannot realize productivity without adoption.”

Forbes “The Consumerization Of Enterprise Technology”, Dec 2017

- ✅ Intuitive experience
- ✅ Anytime, anywhere
- ✅ Heavily personalized
- ✅ A sense of community
- ✅ Multiple modalities
- ✅ Quick and snappy
- ✅ Alive and changing

TREND #3

The rise of automation in software

Hyper-automation industry forecasted to reach **$600 billion** by 2022, said Gartner.

🤖 Robotic Process Automation (RPA)

The automation of repeatable and redundant, rule-based human action through software bots

🧠 Intelligent Automation (IA)

The automation of nonroutine tasks through artificial intelligence

🤔 Cognitive Process Automation (CPA)

The ability for bots to replicate decisions requiring human judgment

🧑‍💻 Low-Code Automation (LCA)

Rapid application delivery with minimal coding and less reliance on developers/engineers

TREND #4

Protecting data → Getting value from data

99% of companies report that they are investing in data initiatives

24% of these companies say they've actually become data-driven

49% of these companies say they're actually driving innovation with data

ORGANIZATIONS
MAKE BETTER DECISIONS USING DATA
COMPLEX GOVERNING ECOSYSTEM
ORGANIZATIONS
MAKE BETTER DECISIONS USING DATA
imgflip.com

So what should the future of data governance look like?

The evolution of governance

Governance 1.0

Monarchy

Rule from above by "the one" (king/queen)

Governance 2.0

Aristocracy

Rule by "the few" (a group of elites) with input from above

Governance 3.0

Democracy

"The many" get a say in the policies governing them

1990-2010

The rise of traditional data warehousing

ORACLE

DATA GOVERNANCE 1.0

Monarchy

Fundamentally built for IT users, acted as a “data inventory”

2008-2018

Data lakes gain prominence as the architecture of choice

hadoop

DATA GOVERNANCE 2.0

Aristocracy

“Data Stewardship” tools built for top-down governance programs

2016-2020

The modern data stack goes mainstream with key capabilities like pay as you go, elastic compute, and 30-minute quick start

databricks

DATA GOVERNANCE 3.0

?

Democracy

1

Build a data community, not a data governance program

Lead with your “why”

Ask your team...

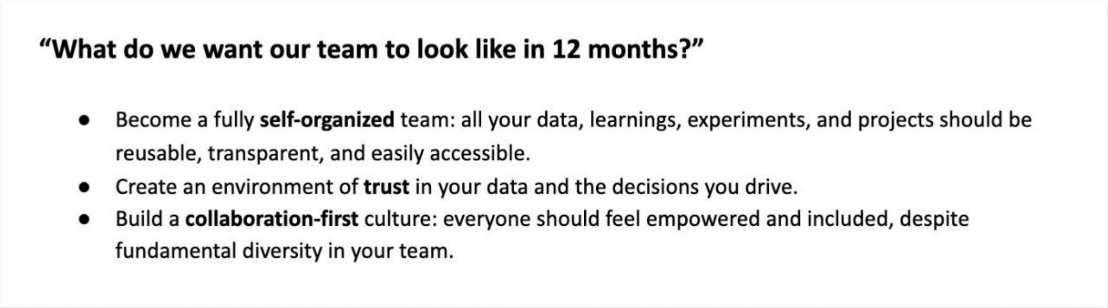
What do we want our data culture to look like in 12 months?

DELHIVERY

1.2 TB
of data ingested per day

66k
Events per second in their data pipeline

1 mil
Packages fulfilled per day, 365 days a year



“What do we want our team to look like in 12 months?”

- Become a fully **self-organized** team: all your data, learnings, experiments, and projects should be reusable, transparent, and easily accessible.
- Create an environment of **trust** in your data and the decisions you drive.
- Build a **collaboration-first** culture: everyone should feel empowered and included, despite fundamental diversity in your team.

Rally your team around a “data product” mindset

Services	Product

Rally your team around a “data product” mindset

Services

- Start at a client problem, and engage a team to fully solve/ implement that problem.
- Get **paid by the client** for implementation.

McKinsey & Company

Infosys

Mu Sigma

accenture

wipro

Product

- Build one solution that can be reused by multiple customers/users to solve a problem.
- Get **paid by customers** for "**usage**" of the product.

Microsoft

salesforce

ORACLE

Spotify

Rally your team around a “data product” mindset

	Data Services	Data Product
Success Criteria	Successful implementation — i.e. did we deliver on time?	Successful usage — i.e did it solve the problem for users, and do they use it regularly?
Reusability	Single use: Build once, for use by one client	Scalability & reusability — build once, for use by many
Requirements & Scoping	Build what the customer asks you to build	Understand commonalities in problems across the customer base and build accordingly
Gratification	Instant gratification — get paid on day zero	Delayed gratification for higher rewards — takes longer to get to usage ("Product Market Fit"), but when it does, it can scale incredibly quickly
Investment	No up-front investment necessary	Up-front investment necessary (time and resources)

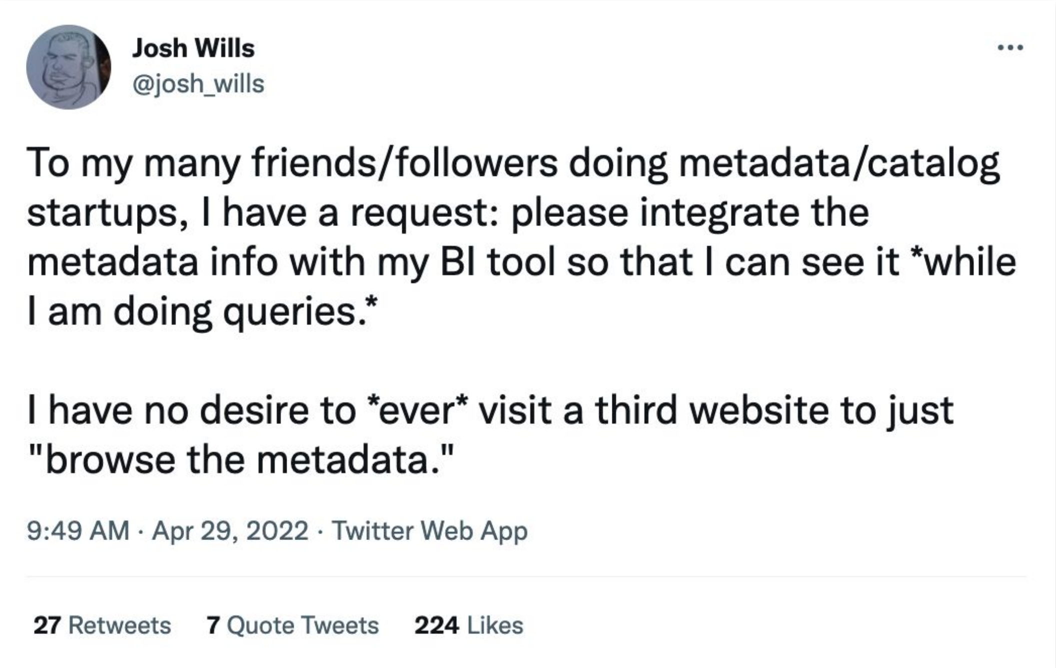
Fundamental principles in treating data as product

- Reusable
- Reproducible
- Well-documented
- Accessible
- Enables self-service for end users
- Scalable (built for more than one user)
- Focused on impact not inputs (end user adoption)

2

Collaboration, not control

Embed collaboration in daily work,
rather than creating another siloed tool or workflow



Josh Wills
@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.*

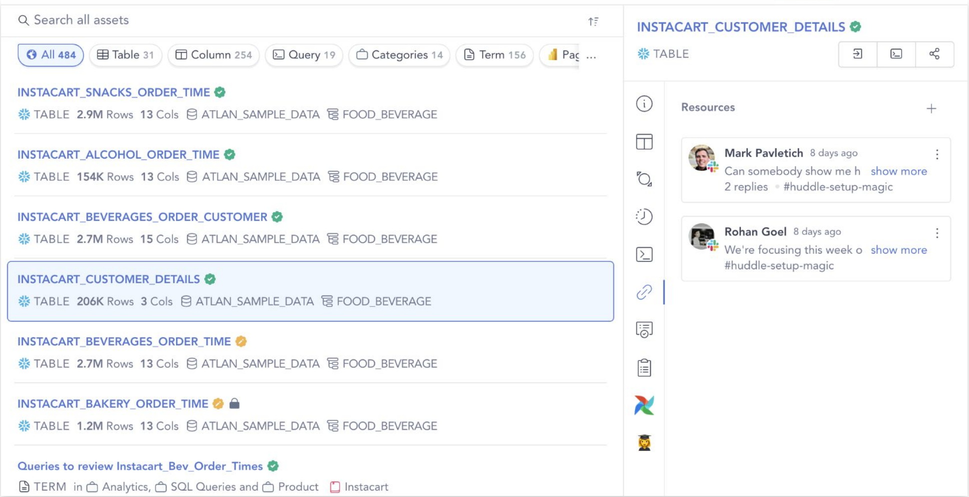
I have no desire to *ever* visit a third website to just "browse the metadata."

9:49 AM · Apr 29, 2022 · Twitter Web App

27 Retweets **7** Quote Tweets **224** Likes

Activate metadata to collaborate where you work all day

Add Slack conversations to data context



Search all assets

All 484 | Table 31 | Column 254 | Query 19 | Categories 14 | Term 156 | Pag ...

INSTACART_SNACKS_ORDER_TIME
TABLE 2.9M Rows 13 Cols ATLAN_SAMPLE_DATA FOOD_BEVERAGE

INSTACART_ALCOHOL_ORDER_TIME
TABLE 154K Rows 13 Cols ATLAN_SAMPLE_DATA FOOD_BEVERAGE

INSTACART_BEVERAGES_ORDER_CUSTOMER
TABLE 2.7M Rows 15 Cols ATLAN_SAMPLE_DATA FOOD_BEVERAGE

INSTACART_CUSTOMER_DETAILS
TABLE 206K Rows 3 Cols ATLAN_SAMPLE_DATA FOOD_BEVERAGE

INSTACART_BEVERAGES_ORDER_TIME
TABLE 2.7M Rows 13 Cols ATLAN_SAMPLE_DATA FOOD_BEVERAGE

INSTACART_BAKERY_ORDER_TIME
TABLE 1.2M Rows 13 Cols ATLAN_SAMPLE_DATA FOOD_BEVERAGE

Queries to review Instacart_Bev_Order_Times
TERM in Analytics, SQL Queries and Product Instacart

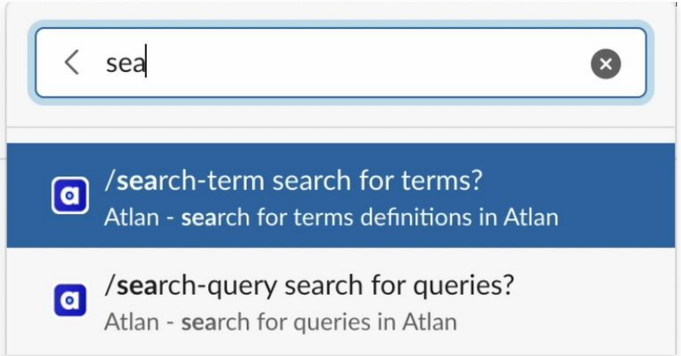
INSTACART_CUSTOMER_DETAILS
TABLE

Resources +

Mark Pavletich 8 days ago
Can somebody show me h show more
2 replies • #huddle-setup-magic

Rohan Goel 8 days ago
We're focusing this week o show more
#huddle-setup-magic

Search for metrics in Slack

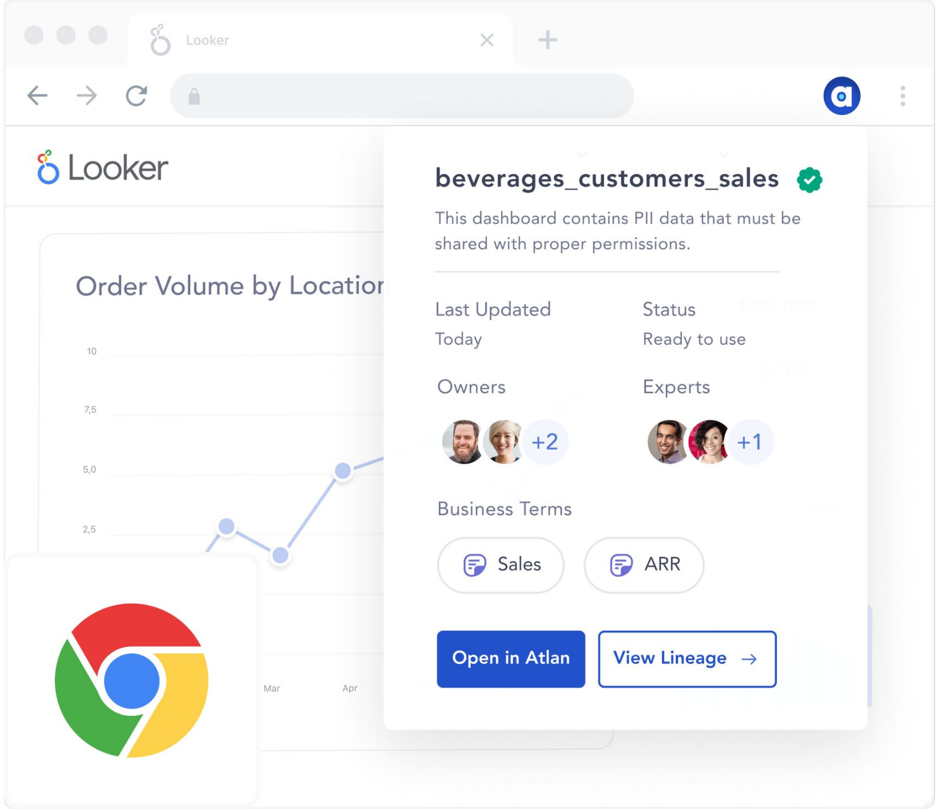


sea

/search-term search for terms?
Atlan - search for terms definitions in Atlan

/search-query search for queries?
Atlan - search for queries in Atlan

Bring context into BI to show the value of governance



beverages_customers_sales ✓

This dashboard contains PII data that must be shared with proper permissions.

Last Updated: Today

Status: Ready to use

Owners +2

Experts +1

Business Terms

3

Automate wherever possible

DELHIVERY

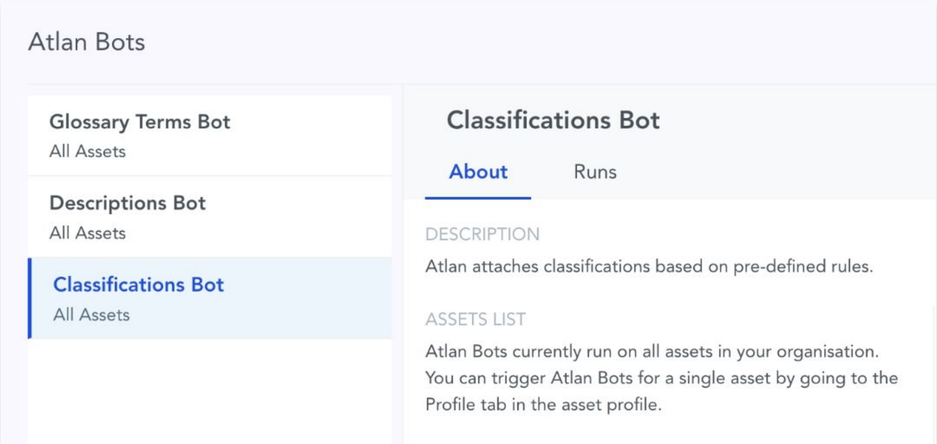
aws

Auto-assigning owners to assets
Auto-attaching column descriptions

Delhivery deploys a bot to automatically scan query log history & custom metadata to find the best owner for every asset.

This helps developers at Delhivery drive documentation volume and standards, and reduce time-to-ownership for assets.

90% of column descriptions were automatically deduced by the bot



Atlan Bots

Glossary Terms Bot
All Assets

Descriptions Bot
All Assets

Classifications Bot
All Assets

Classifications Bot

DESCRIPTION

Atlan attaches classifications based on pre-defined rules.

ASSETS LIST

Atlan Bots currently run on all assets in your organisation. You can trigger Atlan Bots for a single asset by going to the Profile tab in the asset profile.

Custom classifications bot

Delhivery uses a bot to auto-classify its assets based on PII and GDPR restrictions.

53% of users reported saving time with Atlan as bot-automation eliminated routine tasks.

Atlan **integrated into CI/CD pipeline** for auto-creation and enrichment of assets

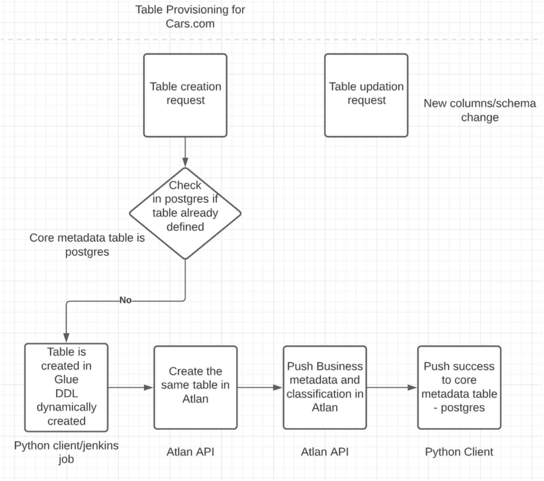
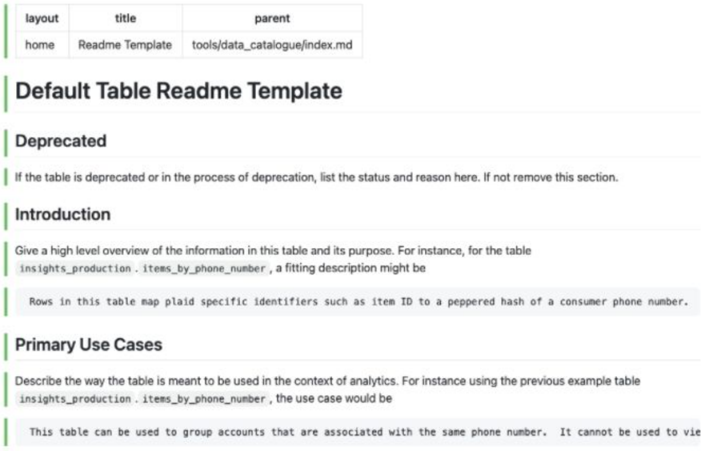


Table Provisioning for Cars.com
Table creation request
Table updation request
New columns/schema change
Check in postgres if table already defined
Core metadata table is postgres
No
Table is created in Glue DDL dynamically created
Create the same table in Atlan
Push Business metadata and classification in Atlan
Push success to core metadata table - postgres
Python client/jenkins job
Atlan API
Atlan API
Python Client

Integrated Atlan into their development workflow to **programmatically generate standardized developer documentation** (automated templates) via APIs



layout	title	parent
home	Readme Template	tools/data_catalogue/index.md

Default Table Readme Template

Deprecated

If the table is deprecated or in the process of deprecation, list the status and reason here. If not remove this section.

Introduction

Give a high level overview of the information in this table and its purpose. For instance, for the table `insights_production` . `items_by_phone_number` , a fitting description might be

```
Rows in this table map plaid specific identifiers such as item ID to a peppered hash of a consumer phone number.
```

Primary Use Cases

Describe the way the table is meant to be used in the context of analytics. For instance using the previous example table `insights_production` . `items_by_phone_number` , the use case would be

```
This table can be used to group accounts that are associated with the same phone number.  It cannot be used to vie
```

4 Go from data governance to DataOps

Every other domain in our organizations has a focused enablement function

SalesOps & Sales Enablement

Focused on improving enablement, productivity, ramp time and success of the sales team

Sales Rep Ramp Time
Win Rate

DevOps & Developer Productivity Eng.

Focused on improving collaboration between software teams, and productivity of developers

Developer Productivity
Deployment Time

ProductOps & Agile

Focused on improving collaboration between product, GTM teams, and customers

Velocity
Cycle Time

> “DataOps is a collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers and data consumers across an organization.”

Gartner

🔄 Lean

🏃 Agile

🤝 DevOps

	Learning	**Applying to Data Governance**
🔄 **Lean Manufacturing**	Centered around “value streams”, and minimising waste through process mapping	How can we align to “value” for end users and the business?
🏃 **Agile**	Moved development process from waterfall to *Agile*	Can we ship “data products” like “software products”? How do we ship fast and involve end users in implementation?
🤝 **DevOps**	Went from siloed teams of software development (shipping software) and ITOps (maintaining software) to integrated dev proces	How do we integrate governance into how we work with data on a daily basis? How should DataOps teams be structured?

DataOps enables the rest of the organization to become data-driven.

This function doesn’t actually execute data or analytics projects. Instead, it focuses on **Tools, Processes & Culture** that will make the rest of organization more data-driven.

Stakeholders of DataOps

1. Data team

Analysts, Analytics Engineers, Scientists, Data Engineers

Impact:

- Improve productivity of data team
- Increase time to value / speed of delivery
- Reduce ramp time of a new joinee
- Reduce attrition

2. Data Consumers

Executives, Business Users, Product Managers, Compliance, Finance etc.

Impact:

- Enable self service
- Reduce dependencies on data team
- Improve speed of decision making

Team work makes the dream work!

Central DataOps Function

Emily Lazio

DataOps Enablement

1) **Masters in Information and Library Sciences**
Understands taxonomy and structure

2) **Children's Librarian**
Energetic, extroverted, great at bringing people together

3) **Information Architect in WeWork's Design team**
Understands data ecosystem and understands user research

Yong Lu

DataOps Engineering

1) **Masters in Computer Science**
Understands data and technology

2) **Engineer & Data Management Leader**
Systems thinker, great at simplifying complex problems

3) **Data Engineering Lead in WeWork's Engineering**
Internal data "guru", able to identify patterns for automation

All other Data Pods

Data Teams
Analysts, Analytics Engineers

Data Product Managers
Liaison between data team & data consumers

Data Consumers
Business users, etc.

Create data products for data consumers

Make Data Governance 3.0 a reality with

atlan + databricks

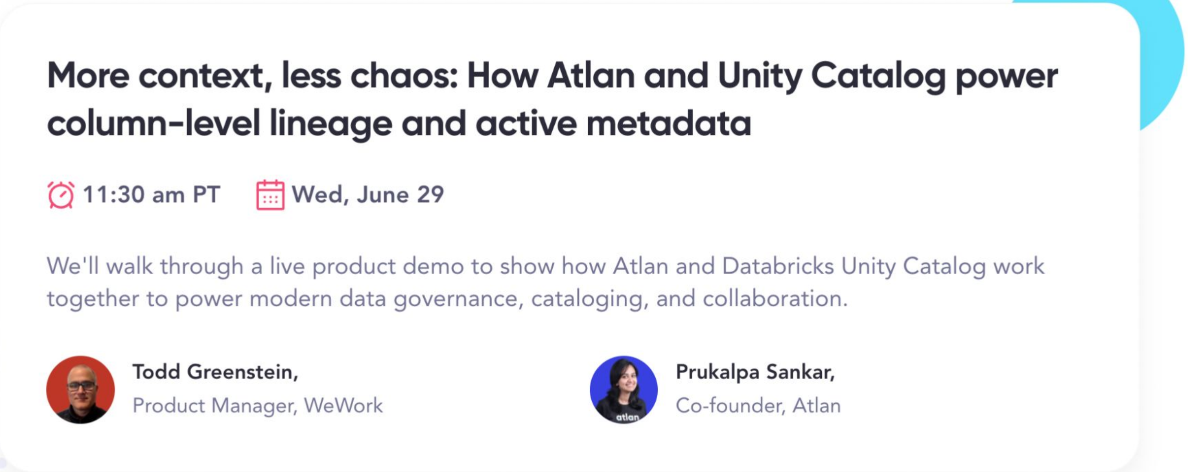
atlan + databricks

Data Governance 3.0 for the Lakehouse Era

- Discovery & catalog
- Data asset 360°
- Column-level lineage
- Active data governance
- Embedded collaboration
- Glossary

💻 **bit.ly/atlan-databricks**

Attend our joint product demo with databricks



More context, less chaos: How Atlan and Unity Catalog power column-level lineage and active metadata

⏰ 11:30 am PT 📅 Wed, June 29

We'll walk through a live product demo to show how Atlan and Databricks Unity Catalog work together to power modern data governance, cataloging, and collaboration.

Todd Greenstein,
Product Manager, WeWork

Prukalpa Sankar,
Co-founder, Atlan

p@atlan.com

@prukalpa

LI: Prukalpa

metadataweekly.substack.com

www.atlan.com