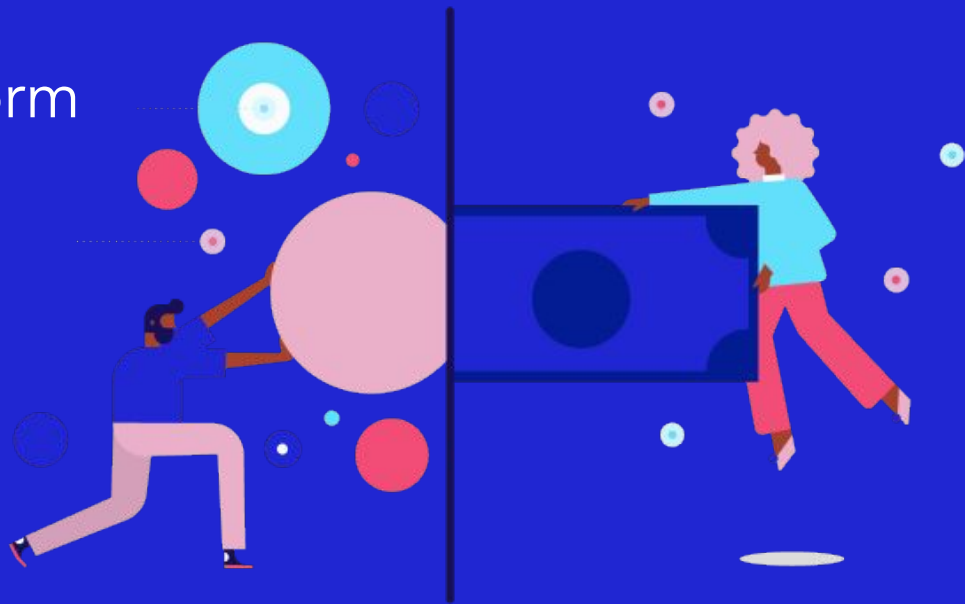


# Meet **atlan**

The active metadata platform

Dataversity: September 7, 2022

Presented by: Andrew Ermogenous



# atlan

Pioneering the active metadata category



wework

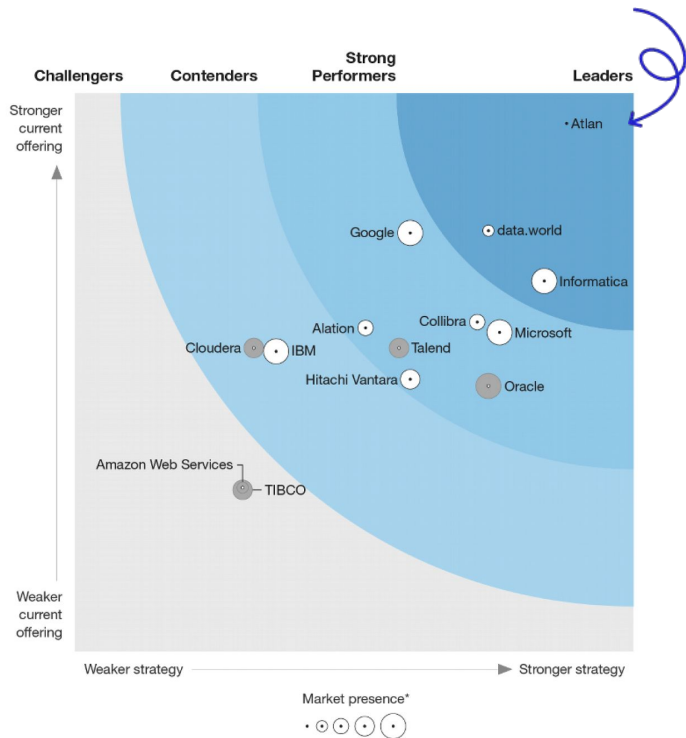
MONSTER

BRITISH AIRWAYS

RALPH LAUREN

News Corp

JUNIPER NETWORKS



Backed by marquee investors

INSIGHT PARTNERS

SEQUOIA



and the top founders & CEOs pioneering the modern data stack

Fivetran

Stitch  
A Talend Company



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

110 bil.

external data points  
ingested, cleaned, and visualized

1.5 bil.

government data points  
aggregated in real time

50+

countries with a diverse  
set of organizations

6.5 bil.

satellite imagery  
pixels processed

500 mil.

Indian citizens' data processed



# Every day was chaos as a data team

 #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?


## 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?**

 #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 :/

## 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!**

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.

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



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

12

months  
to build

8

member  
team

12

master data  
hierarchies

3.5b

dynamic data  
points

42

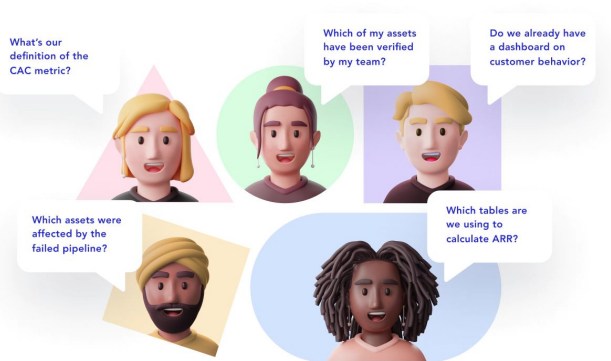
data portals  
connected

# Traditional data catalogs fail because they are “passive”



## Siloed

Doesn't give consumers context where they are, when they need it.



## Generic

Context means different things to different users: data analysts, engineers, scientists, business

Column description in one column should be inherited to all derived columns

Name, email → Auto-tag to "PII"

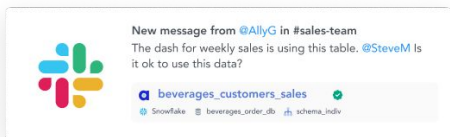
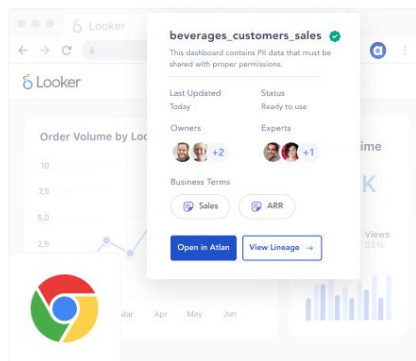
If 90% users login to dashboards at 10 am on Monday morning, we should auto-scale up compute.

## Manual

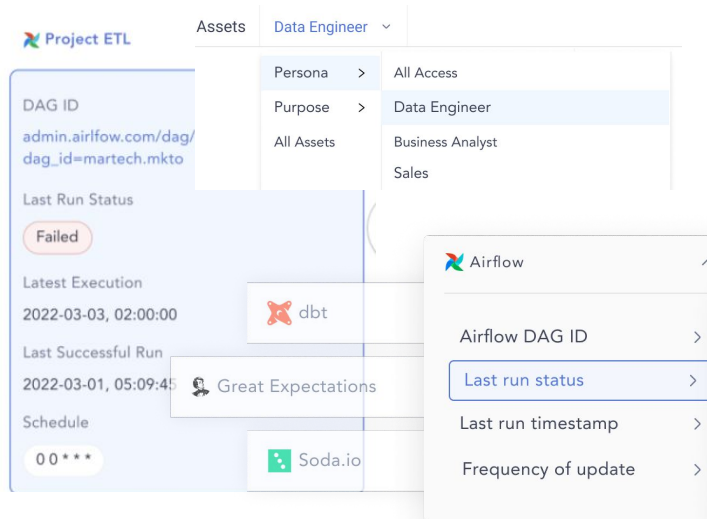
Relatively manual setup, with no programmatic value creation through metadata

# From ~~Passive~~ to Active

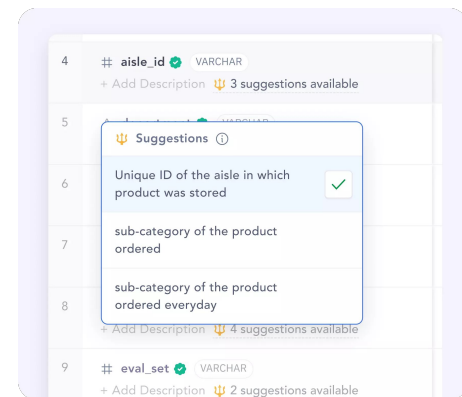
Siloed to  
**Embedded**



Generic to  
**Personalized**



Manual to  
**Autonomous**



PERSONALIZED, CUSTOMIZED EXPERIENCES



**Collaboration Workspace**

Discovery   Lineage   Glossary  
Governance   Querying

Reusable 360° Asset Profiles as the **SINGLE VERSION OF TRUTH**  
Like code has a profile on **GitHub**

**Central Active Metadata Platform**

IN-FLOW EMBEDDED EXPERIENCES

Embedded context   Contextual discussions  
Deep in-flow integrations

Metadata Activation

AUTOMATION WORKFLOWS

Quality   Observability  
Alerts   Cost optimization  
Security/audits   CI/CD  
Programmatic governance

Data Lakes & Warehouses

BI & Data Analytics

Pipelines / Code / Transformation Assets

Custom Assets

APIs   Notebooks  
Features

ORCHESTRATION LAYER





People not speaking the same language at WeWork has always been a problem. So we started with Public KPIs to make sure that we were all aligned on what they are."



Harel Shein,  
Head of Data Engineering

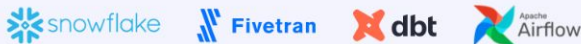


### How WeWork built trust in data with Column-Level Lineage

WeWork was going public, so their publicly reported data had to be fully compliant and 100% consistent.

- The data engineering team added a "Critical" classification to any publicly reported dashboards.
- They propagated the classification to upstream source tables through column-level lineage.

WeWork's modern data stack



The glossary is how you can create automated governance workflows. So someone who knows the data will get automated description suggestions, which they can approve."



Sara Swart,  
Chief of Staff to CTO



### How Monster created a knowledge layer with Glossary

Monster created Glossaries for multiple domains:

- Business departments (Marketing and Sales)
- Data sources (Salesforce and Google Analytics)
- Enterprise metrics (KPIs)

The Data Governance Team linked all their assets to definitions and metrics in the glossary, creating a knowledge layer.

Monster's modern data stack



"Snowflake's metadata is phenomenal, so it's cool that you can tap into that. End-to-end access to your metadata is really powerful. And the integration with PowerBI has been great."



Victor Wilson,  
Data Architect



### Why Scripps chose Atlant as its modern data catalog

- Enterprise-wide collaboration: Scripps was bringing analysts from finance, supply chain, and hospitals onto one platform.
- Assured security: Scripps needed programmatic governance like automated PII detection for sensitive healthcare data.

Scripps's modern data stack



Thank you!

Andrew Ermogenous  
Head of North America  
andrew@atlan.com