



# Navigating the Impact of Generative AI and ChatGPT on Metadata Management

Juan Sequeda, Ph.D. - Principal Scientist and Head of AI Lab  
juan@data.world

# Today's agenda

- **Opportunities, Concerns and Landscape**
- **Generative AI and Data Catalog**
- **How to get started right now?**



# data.world



- This is where your content would go.
- Use this slide if you want to have a bulleted list or
- **The Data Catalog Platform** mpanied by a
- Most used data catalog in the world
- Catalog, Governance, DataOps
- Built on a knowledge graph architecture
- Deep integrations with modern data stack

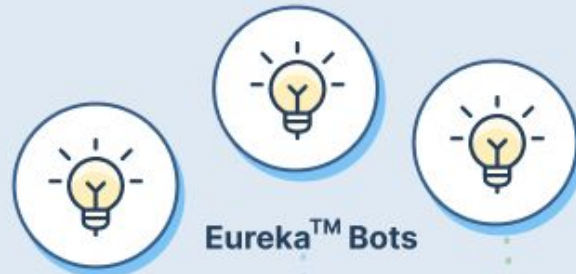


# The Data Catalog Platform

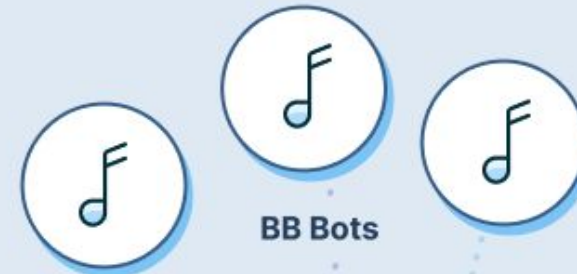
Introducing The Parliament - a team of embeddable data.world AI bots



Curations powered by LLMs

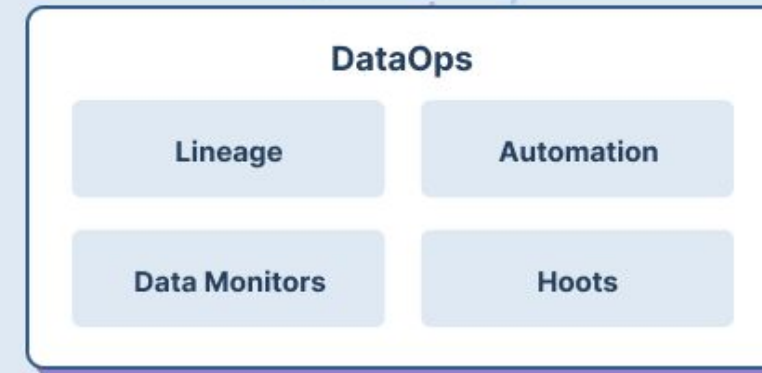
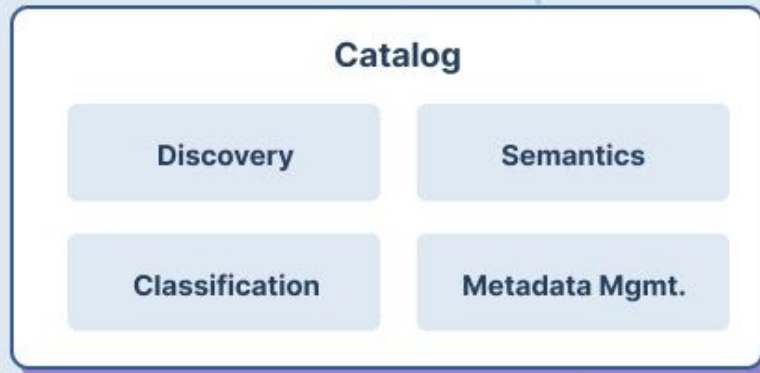


Automations powered by Policies



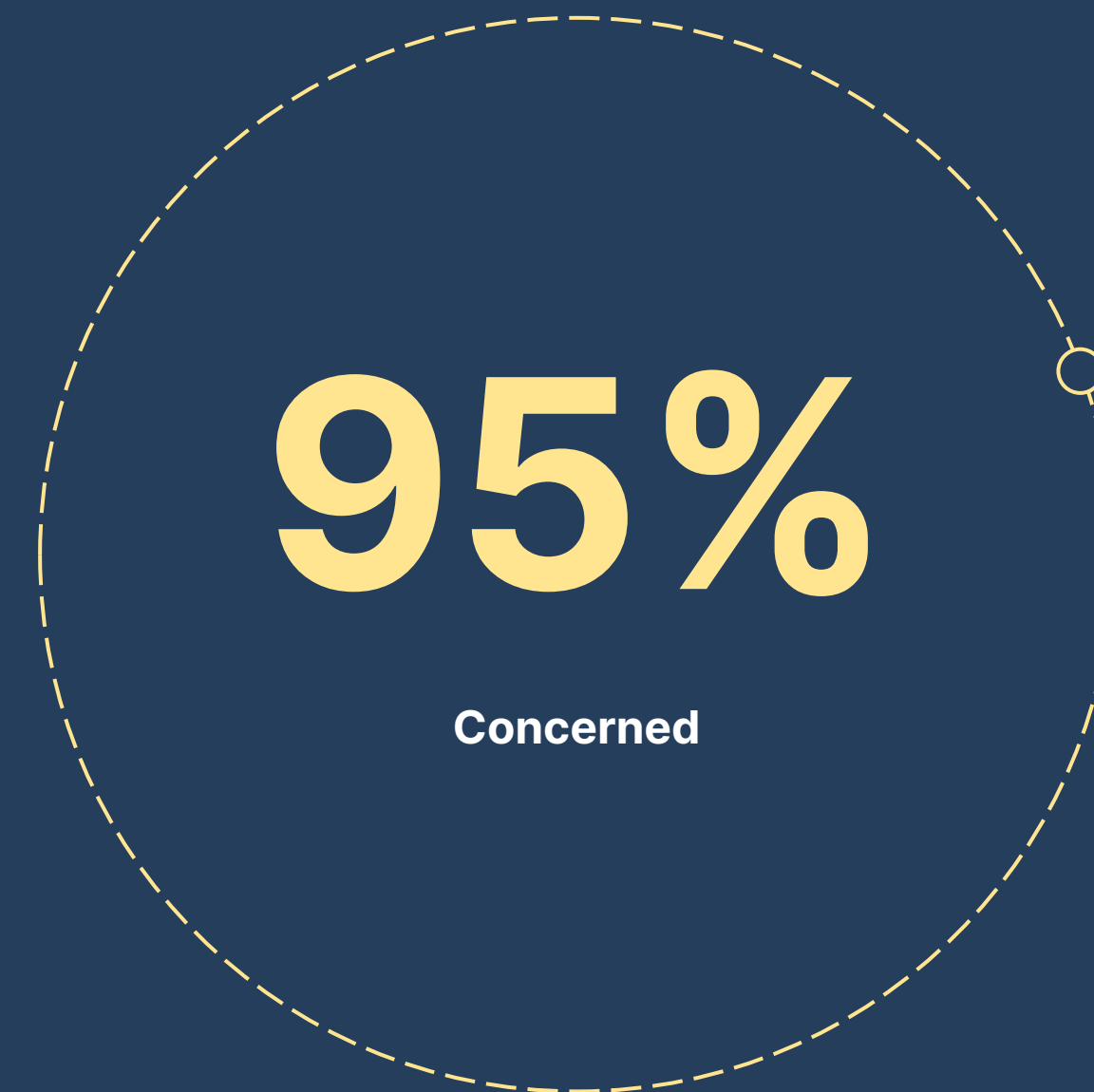
Data quality powered by monitoring

## data.world applications



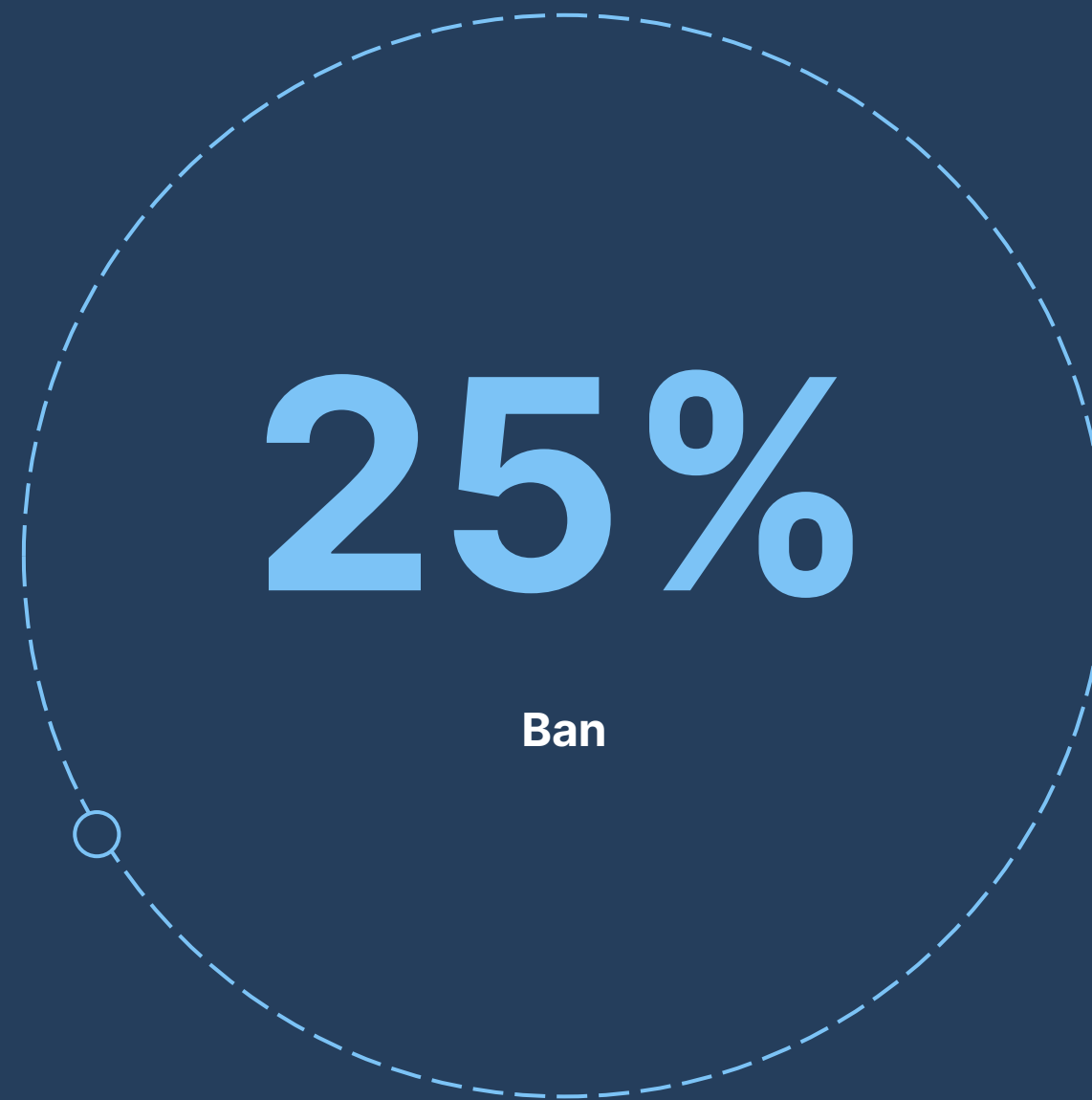
Dataversity SIG Gen AI, ChatGPT, Data Governance - June 2023

## How do you feel about AI?



Dataversity SIG Gen AI, ChatGPT, Data Governance - June 2023

## Organizational ChatGPT usage



Dataversity SIG Gen AI, ChatGPT, Data Governance - June 2023

# Exploring ChatGPT for



# Opportunities



## Technical

- **Code:** Generate and Optimize code
- **Metadata:** Generate metadata and descriptions
- **Natural language** to SQL, SPARQL, Cypher



## Business

- **Chatbot** for customer/technical support
- **Writing Assistant** that generates an initial draft of content/letter which can then be reviewed by users
- **Summarization** documents, financial reports, extract metrics/kpis



# Concerns



## Social

- **Abuse:** Writing bad code, using ChatGPT to answer interview questions
- **Long term effect of skills:** Lack of critical thinking. Generating code without understanding what it means and taking the output literally. Thinking, “it’s good enough.”
- **Intellectual property issues**
- **Deep fake:** Can’t differentiate between human and non-human



## Technical

- **Hallucinations:** Makes things up – unreliable.
- **Security:** Leaking information. Users providing PII data – exposing vulnerabilities. Who keeps all these prompts? What happens if they get leaked?
- **Bias in the System**
- **Not understanding** how the models were trained

# Governance for AI: safe usage (You can do this now)

## → Establish policies on when you can use it

- Guardrails with common sense rules
- Call out when AI was used and how
- Legal: Understand who is accountable

## → Education

- First, get the right SMEs who understands it. Do it now, otherwise it will be overwhelming later.
- Educate on critical thinking. It's not that different from, "I saw it on the internet".

## → Acknowledge job shifts and what to do

- "Free up the intern time – don't replace the intern"

## → Prompt Governance

- Create trusted and verified prompts
- Record retention. Save prompts and results. How many things have you produced with GPT? Government bodies need to address FOIA.

# AI for Governance: Increase productivity (vendors do this)



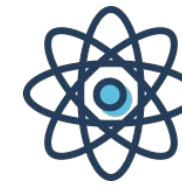
Data Steward



Data Engineer



Data Analyst



Data Scientist



Business User

## Producers

### → Metadata Enrichment

- Detect PII, PHI
- Tagging
- Infer relationships
- Find Anomalies, data quality issues

### → Understand Code

- Explain what code does

## Consumers

### → AI assisted search

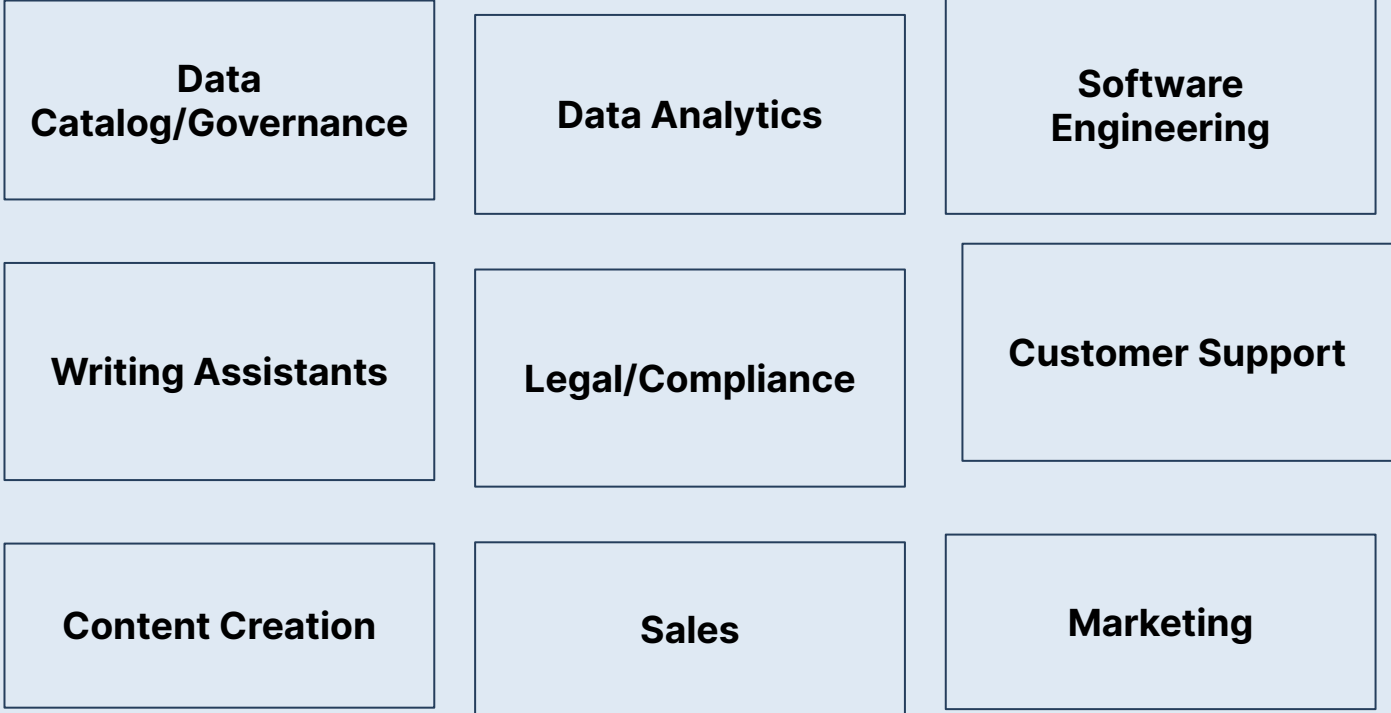
- Suggest filters
- Natural language search

### → Jumpstart data exploration

- Quick ideation based on metadata
- Save questions to help future explorers!

# Generative AI and Large Language Model (LLM) Landscape

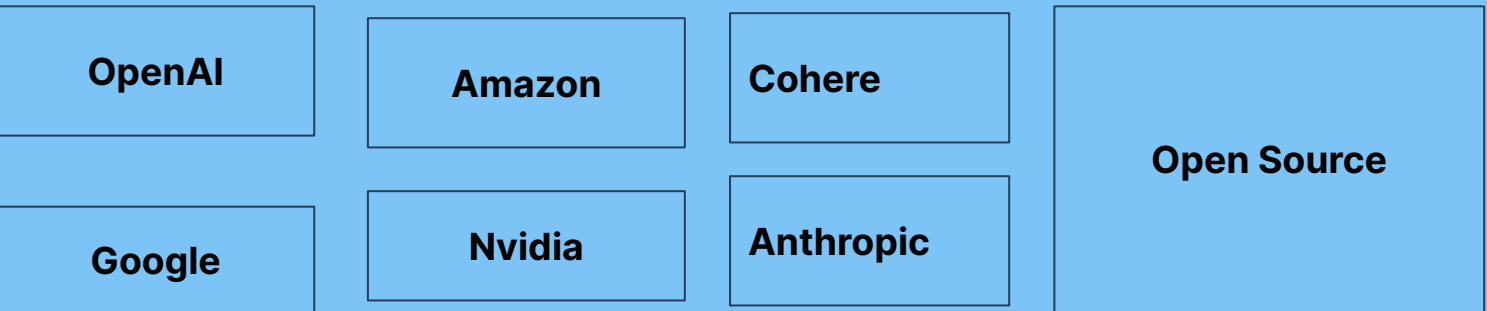
## Applications



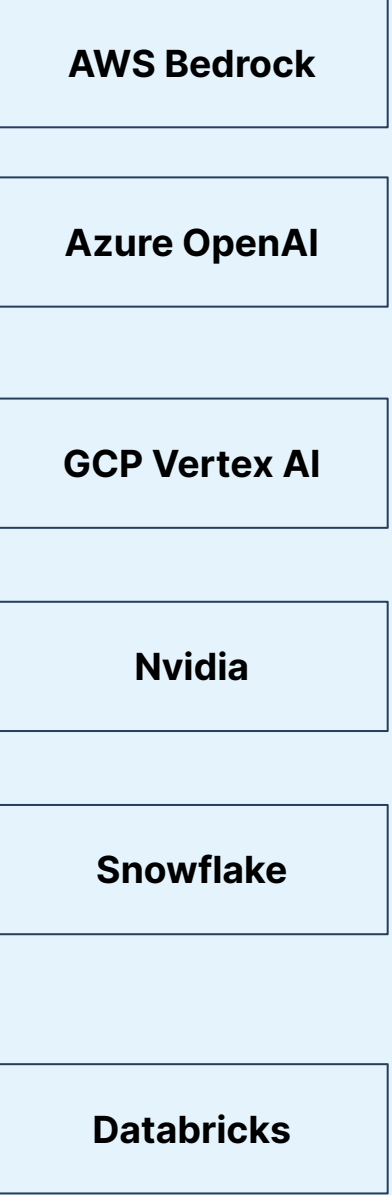
## Tools



## Foundational Models



## Cloud Vendors



3rd Party LLM  
(Option 1)

1st Party LLM  
(Option 2)

**Vendor Managed**

Data Catalog

**Customer Managed**

LLM Partner  
(Option 3)

Customer BYO-LLM  
(Option 4)

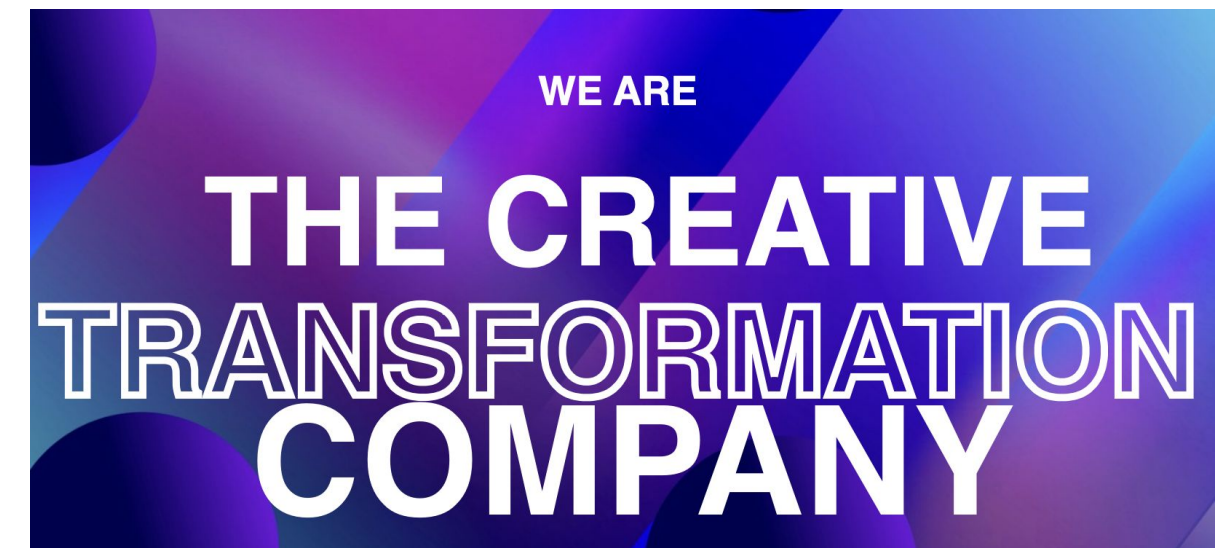
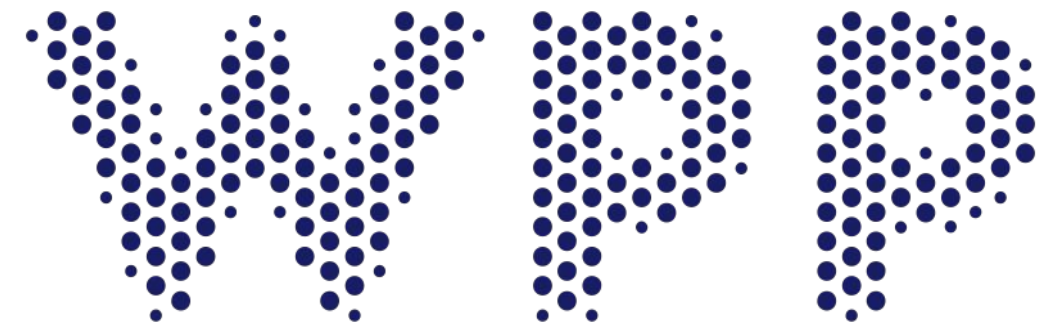
- Consider**
- Legal
  - SOC2 Type 1/2
  - Cost
  - Fine Tuning vs Prompt Engineering

# Generative AI and Data Catalog



**data.world**

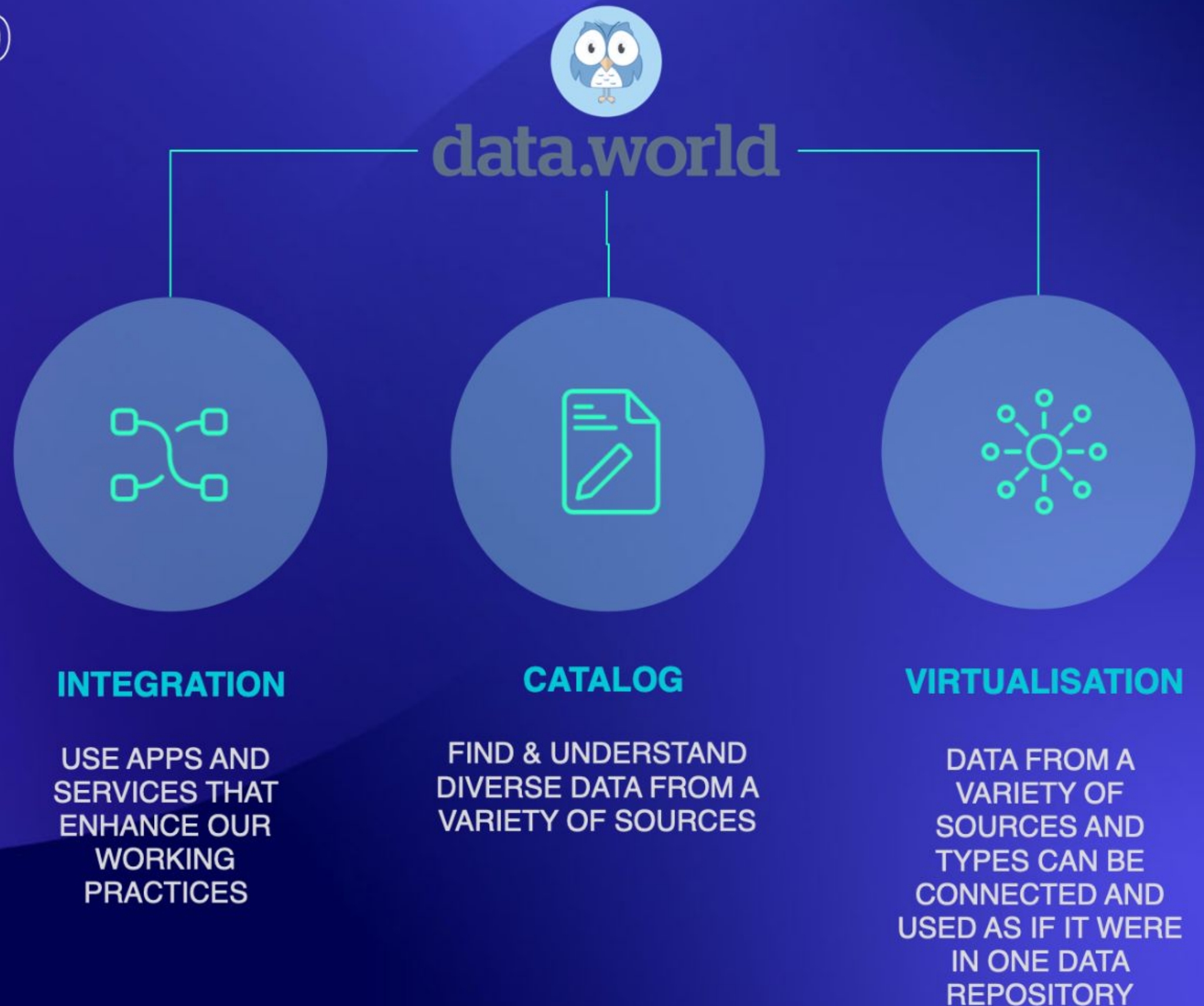
**Co-Innovation**



# WPP OPEN DATA CATLOG

## POWERED BY DATA.WORLD

- A platform that allows you to unlock and share data resources across the WPP network – deliver engaging pitches, client work, and research
- Share data and knowledge – find data, documents, images, videos, and analysis that you can adapt for your own collaborative projects
- Deliver valuable insight from one location – centralise and manage project datasets in one simple user interface



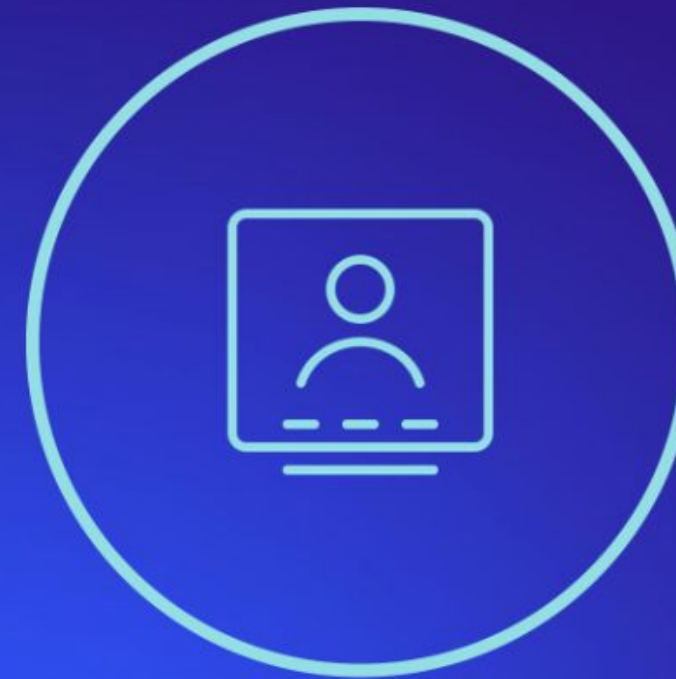


# FOCUS ON TODAY'S PROBLEMS WITH TOMORROW'S CAPABILITIES



## DISCOVERY

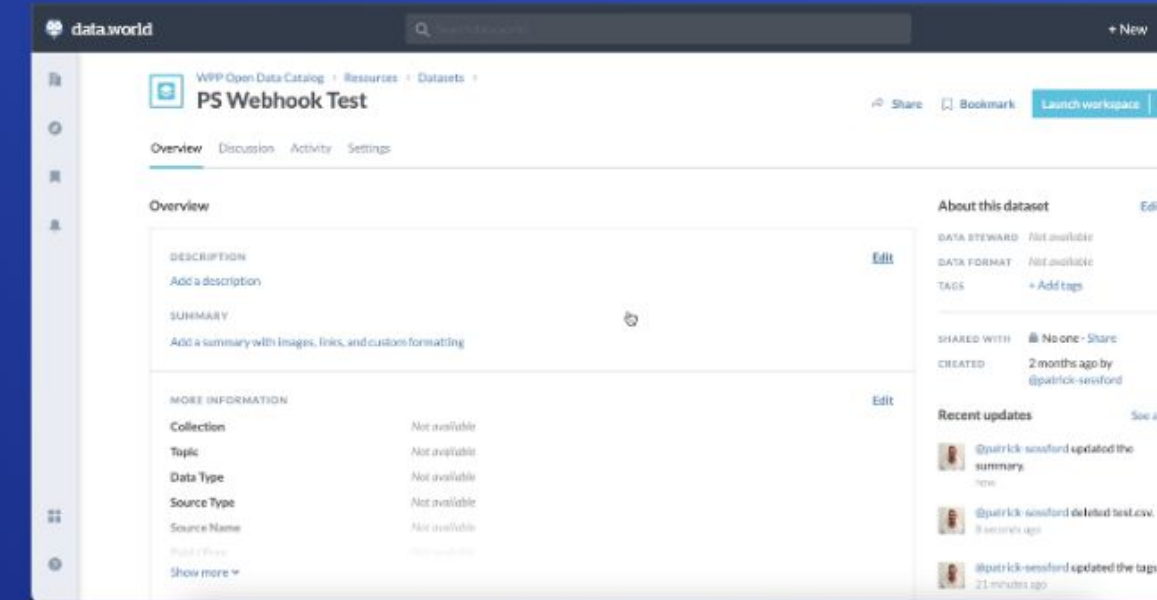
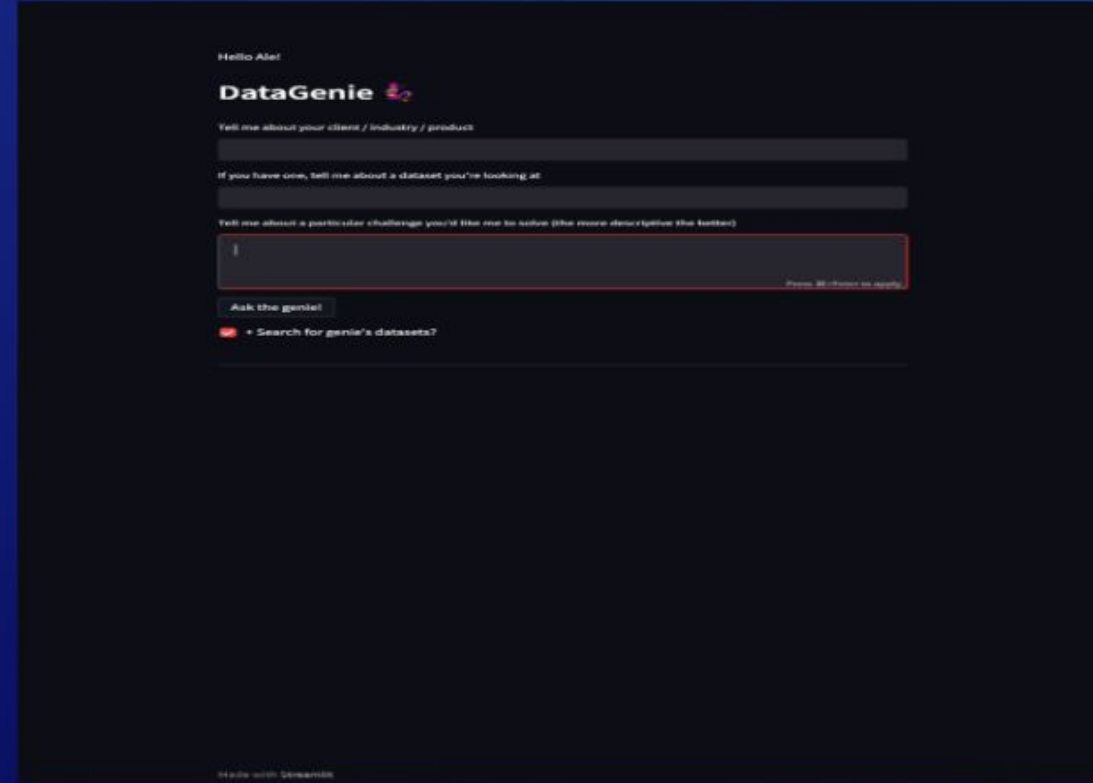
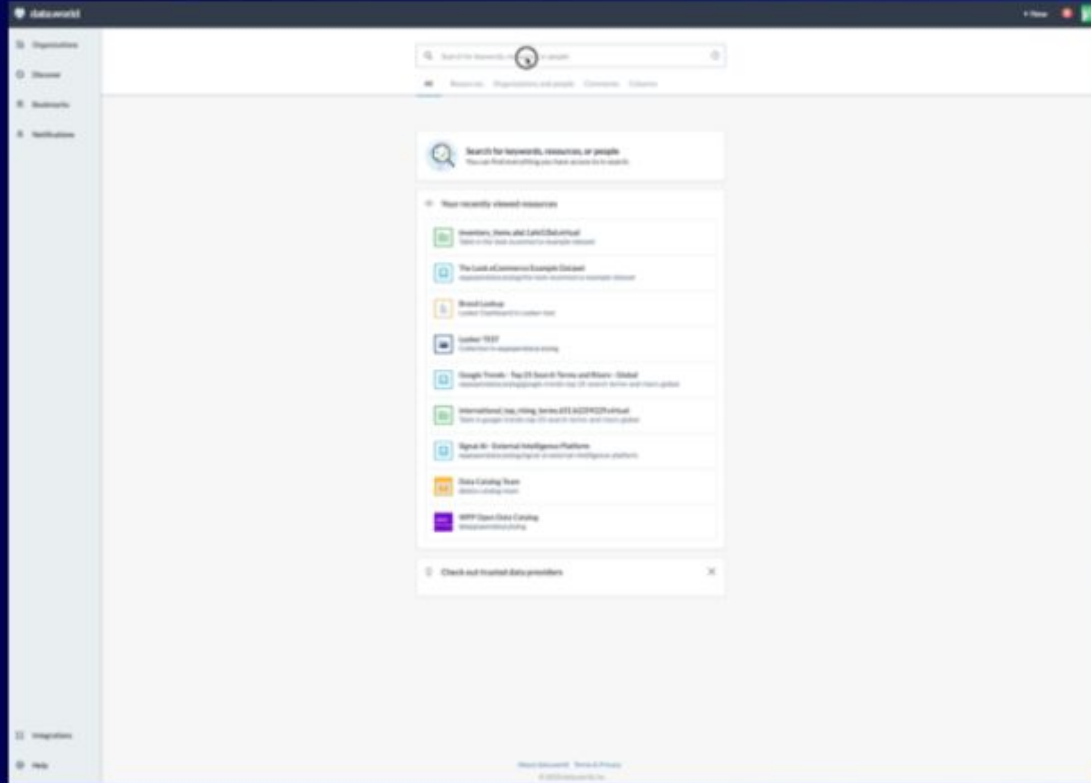
**SIMPLE INTUITIVE  
EXPERIENCES  
POWERED BY  
INTELLIGENCE**



## USE

**PROVOKING THE  
CREATIVE  
APPLICATION OF  
DATA**

# 3 PROTOTYPES TO COUNTER 3 CHALLENGES



## IMPROVING & AUGMENTING THE SEARCH EXPERIENCE

- Referencing multiple data repository's
- Natural language search

## USE CASE INSPIRATION

- Inspire colleagues on how to source and apply data that applies to their needs
- Generate ideas on the application of datasets against specific needs

## ACCELERATED CURATION

- Automatically suggest tags, descriptions and summaries as data is being curated

# Meet Archie Bot

## → Effortless enrichment

- No more catalog ghost-towns
- Aid governance and data ops teams in understanding policies and complex data pipelines.

## → AI-assisted search

- Suggest filters and allow natural language search
- Faster discovery and understanding

## → Jumpstart data exploration

- Quick ideation based on metadata
- Save questions to help future explorers!

## → Lower the code barrier

- Ask your data questions and get answers!

6d what are all of the approved snowflake policies

All Resources Organizations and people Comments Columns

Advanced · Collapse all 4 results Turn off Archie Bot

6d Search with Archie Bot Show query

ACTIVE SEARCH

Status: approved ×

Resource Type: ( Snowflake Row Access Policy OR Snowflake Masking Policy ) ×

Refine your search

6d Type in a follow-up question

Did Archie Bot interpret your search correctly? 👍 👎

NIGHTOWL.PUBLIC.MASK\_CONFIDENTIAL  
PM Sandbox

[Archie Bot suggestion] This policy masks confidential information in the "val" column for all roles except "ACCOUNTADMIN", which has column-level access to...

Snowflake Masking Policy in NIGHTOWL

# Effortless enrichment

**What:** Descriptions for Tables, Columns, and Glossary terms, Collections, SQL queries, Views, dbt Test, dbt Models, Snowflake Masking Policies, Snowflake Access Policies.

**Who:** stewards and governance analysts

## Why

- Enrich the catalog to improve discovery and understanding by adding natural language descriptions.
- Helps all users understand and find the data because it provides more metadata for search!

### Edit Table

Overview Geography Sensitivity Status Migration Status More ▾

Title 100

Click again to get a new suggestion

Description  224

[Archie AI suggestion] The table "ASSET\_ASSETINVENTORY" contains information related to asset inventory, including the ID, end service date, in service date, and serial number.

Did Archie AI suggest a relevant description?  

# Lower the code barrier

**What:** convert SQL query to plain text and plain text to SQL query

**Who:** data analysts and engineers

## Why:

- Get quicker answers about what is happening with the pipeline by providing plain English descriptions of sql for metadata resources like Views, dbt Test, dbt Models, Snowflake Masking Policies, and Snowflake Access Policies.
- Helps users who may not know sql very well
- Helps all future users who might have a similar question because they can find the query later via search!



The screenshot shows the Archie AI interface. At the top, there is a header with the Archie logo, the text "Archie AI Question: What is the total number of asset...", a "PRIVATE" lock icon, and buttons for "New template", "Save", and "Archie". Below the header is a code editor with a SQL query. A dropdown menu is open on the right side of the code editor, showing "Generate SQL query" and "Summarize SQL query".

```
1 -- Generated from prompt: What is the total number of assets assigned to employees?  
2 WITH assets_assigned AS (  
3     SELECT COUNT(DISTINCT orderlines.stockitemid) AS total_assigned  
4     FROM orderlines  
5     JOIN orders  
6     ON orderlines.orderid = orders.orderid  
7     JOIN customers  
8     ON orders.customerid = customers.customerid  
9     WHERE customers.customercategoryid = 1  
10 )  
11 SELECT assets_assigned.total_assigned AS total_number_of_assets_assigned_to_employees  
12 FROM assets_assigned;
```

# AI-assisted search

**What:** use natural language to search and refine your results, get suggested filters

**Who:** all catalog users

**Why:**

- Get quicker answers and find data faster by providing natural language search support, natural language filtering, and suggested filters.
- Create a more expressive query with natural language.

The screenshot displays a search interface with a search bar containing the query "all approved tableau dashboards". Below the search bar are tabs for "All", "Resources", "Organizations and people", "Comments", and "Columns". The "Resources" tab is selected. On the left, a "Filters" sidebar is shown with categories: RESULTS (Include all community results (4)), RESOURCE TYPE (Tableau Dashboard (4)), OWNER (initech (4) is selected), STATUS (approved (4)), TAG (customer (1), kpi (1), order (3), revenue (2), state (1)), COLLECTION (Tableau Catalog (3), Tableau-Analytics-Prod (1), tableau-test (1)), and STEWARD (alex huckabee (1)). The main area shows "4 results" and a "Search with Archie AI" section with active filters: "Resource Type: Tableau Dashboard" and "Status: approved". Below this is a "Refine your search" section with suggested filters: "Collection: Tableau Catalog", "Tech Owner: mo dodge", "Tag: order", "Steward: alex huckabee", and "Popularity: Unpopular". A search bar at the bottom of this section says "Ask Archie a follow up question". The first result is a card for "Executive Overview - Profitability" by Initech, with a description: "This dashboard tracks key profitability metrics across geographic regions for our main product lines." The card includes a table of metrics and a map of the United States.

Size	Order	Profit Margin	Product Order	Category Order	Tag Order	Steward
\$2,901,677	\$950,120	12.4%	\$71.90	\$5,609.11	15.44%	47,010

# Jumpstart data exploration

**What:** Research question generation based on data in a Collection

**Who:** data analysts, data scientists

**Why:**

- Ideate on research questions within the context of a collection,
- When a question is saved as a query to a project, this helps all future users who might have a similar question because they can find the query later via search!

## Generate questions

Below are a list of questions that this Catalog may be able to answer. Select any or all of the questions to save them to a project or dataset where you can generate relevant queries.

- SELECT ALL
- What is the total number of assets assigned to employees?
- How many assets are currently unassigned?
- What is the most common asset category in the inventory?
- How many assets have reached their end of service date?
- What is the percentage of assets currently in service?
- How many departments does the company have?
- What is the total number of locations where the company operates?
- What is the number of studies conducted by the customer?
- How many customers have a specific address type?

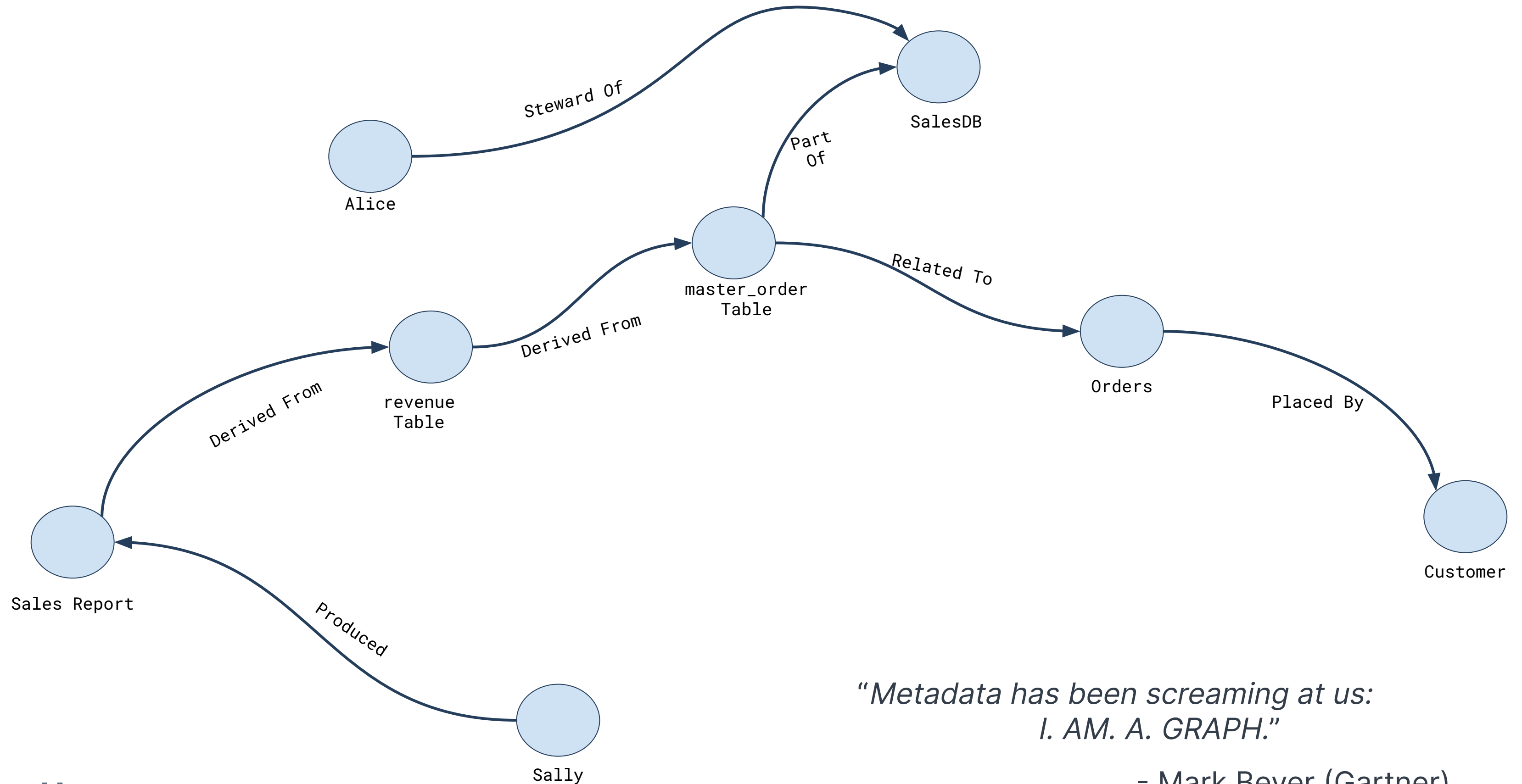
What is the total amount spent by customers on purchases? (aggregate data analysis)

How does this work?

# Knowledge Graphs + AI/LLM



# Knowledge Graph: integrate data and knowledge at scale



*"Metadata has been screaming at us:  
I. AM. A. GRAPH."*

- Mark Beyer (Gartner)

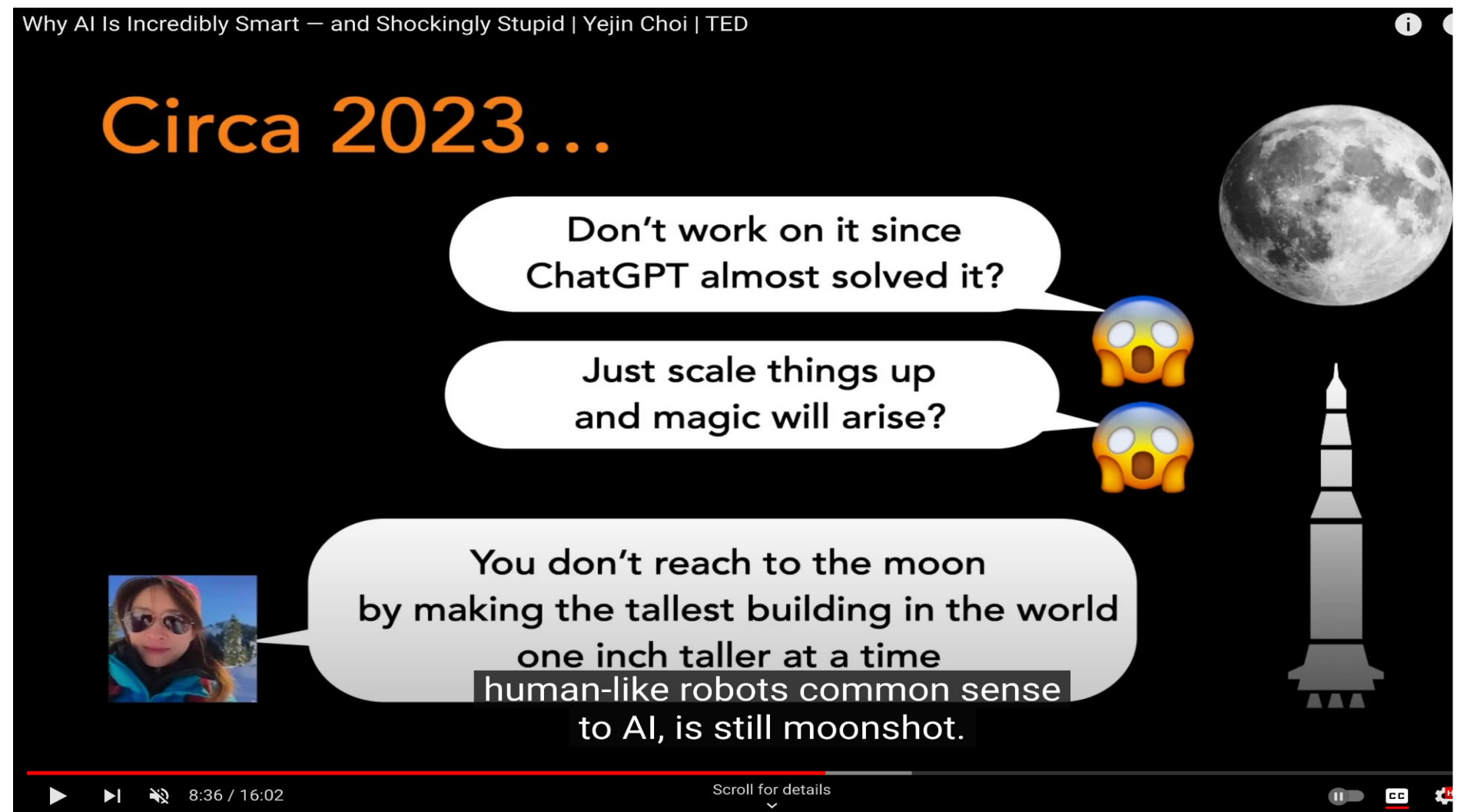
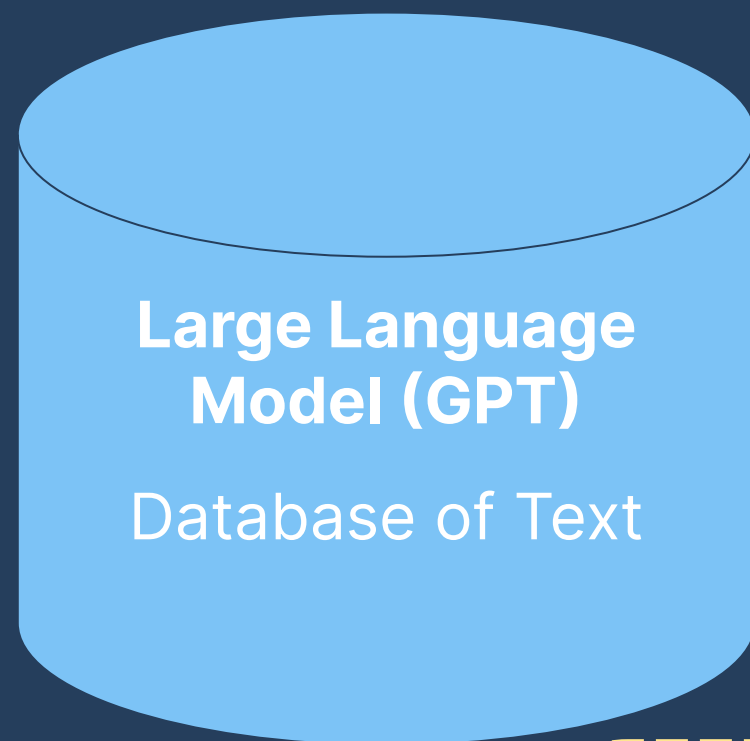
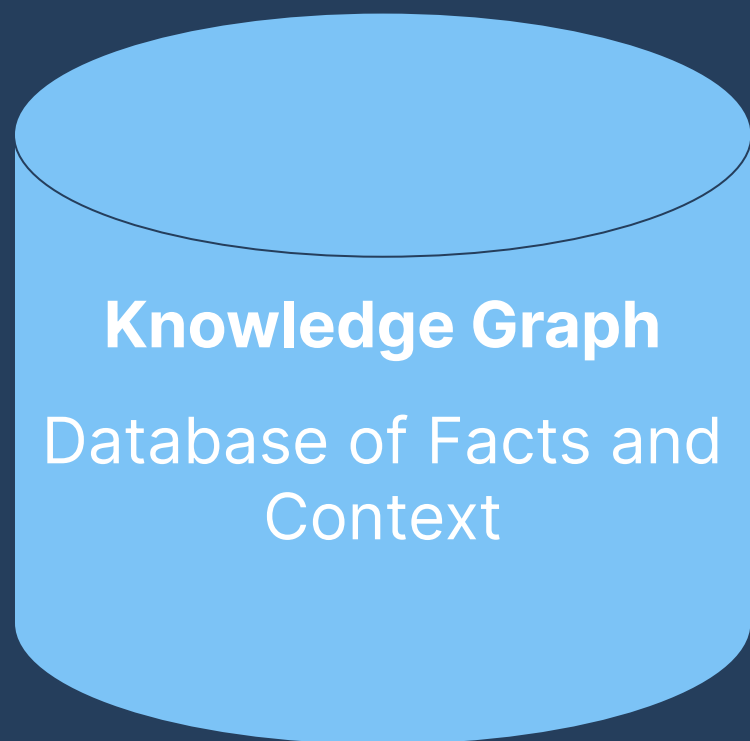
“

**Knowledge graphs provide the perfect complement to LLM-based solutions where high thresholds of accuracy and correctness need to be attained.**

Gartner

AI Design Patterns for Knowledge Graphs and Generative AI - June 9, 2023

# Knowledge Graph and LLM



Why AI Is Incredibly Smart — and Shockingly Stupid  
Yejin Choi | TED April 2023

# Knowledge Graph and LLM



The screenshot shows the data.world interface with a search bar containing the query "my question about our data". The search results are displayed in a list format, including items like "Client", "Account-Based Everything / Revenue", "Order", "Enterprise Resource Planning (ERP)", and "Purchase Order (PO)". A red box highlights the search bar and the first few results. A red arrow points from the text "But not anything over here, or the other details within your catalog" to the search results area.

*But not anything over here, or the other details within your catalog*

# How to get started right now?



## Task Force

- Setup a task force to understand the opportunities and risk ASAP.
- Tie efforts to persona and biz value.



## Survey your Vendors

- Start small. Survey your vendors what they are doing with AI.
- HOW are they are doing it? What have they learned?




## Start Small

- Pilot with a small team. Do a comparison. Does it actually help?
- Measure the productive gain wrt cost.

# Interested in more practical tips? Schedule time with my team through the data.world AI Lab. <https://data.world/ailab/>

**JUNE 22ND: Leapfrog the data trust gap with powerful insights and governance** [Save your seat](#)

 [Product](#) [Customers](#) [Developers](#) [Resources](#) [Company](#) [Sign In](#) [Get a demo](#)

## Collaborate with data.world's AI Lab

Schedule time to explore automating your data management processes with knowledge graphs, third-party AI tools like GPT-4, and data.world.

data.world's knowledge graph architecture easily integrates third-party AI tools into the data catalog, speeding data catalog creation, enrichment, and adoption. Learn more about how current data.world customers are taking advantage of these capabilities and explore the options for integration in your data.world enterprise data catalog.

**In this 30 minute personalized meeting with Dr. Juan Sequeda, Principal Scientist and Head of the AI Lab at data.world, we will:**

- ◆ Share examples of how third-party AI tools are being integrated into the data.world enterprise data catalog
- ◆ Review the Analytics and AI tools you're interested in integrating with data.world
- ◆ Discuss potential approaches to integration with the knowledge graph
- ◆ Answer your questions about opportunities for collaboration to kick-off

First name  Last name

Work phone number

Work email\*

Company name

Job title

Analytics and AI Tool of Interest

Country/Region\*

By clicking "Save your seat" you are agreeing to our [terms of service and privacy policy](#).

[Sign up!](#)

# Follow data.world to learn more about what we're doing with AI

[BLOG POST](#) / [DATA DISCOVERY](#)

## Lowering the barrier to entry for data-driven decision making

With new Generative AI capabilities, far more people can use data.world to discover data and unlock organizational knowledge

6 MIN READ



**Brett Hurt**  
CEO & Co-founder, data.world

ANNOUNCING  
  
Archie Bot

[BLOG POST](#) / [DATA DISCOVERY](#)

## Using generative AI to enrich Snowflake metadata with data.world

6 MIN READ



**Dave Griffith**  
Distinguished Engineer, data.world



**data.world**

**Thank you**

