

# Teamwork Makes the (Open Source) Dream Work

## The Power of Cross-community Collaboration

Kyle Eaton, Growth Lead | Great Expectations

Maggie Hays, Founding Community Product Manager | DataHub



DataHub

+



great  
expectations



# DataHub

## Who We Are



**Maggie Hays**  
Founding Community  
Product Manager,  
DataHub



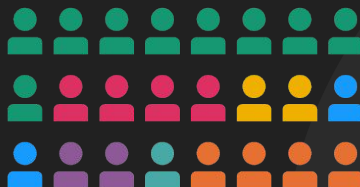
**6,700+ Slack Members**

2.8x YoY Growth

Across 57 Countries & 27 Local Time Zones



### Top Member Roles



- Data Engineer
- Software Engineer
- System Architect
- Engineering Manager
- Product Manager
- Data Analyst
- Other

### Top Member Industries



Software



Ecommerce



Info Tech



Fintech



Healthcare

### DataHub Adoption Stage

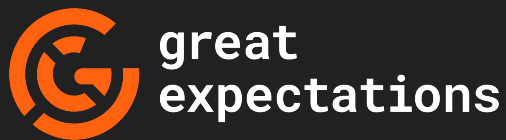


- Evaluating if DataHub is the right fit
- Running a DataHub POC
- Deploying DataHub to production
- Rolling out DataHub to internal stakeholders





**Kyle Eaton**  
Growth Lead,  
Great Expectations



10,000 Members on Slack



~6,000 Organization in the community



~Found in 26 different time zones



358 Contributors



9.5M monthly PyPi downloads



Data Engineer plus, Data Scientist, ML  
Engineer, Data Analyst, Data Practitioner...

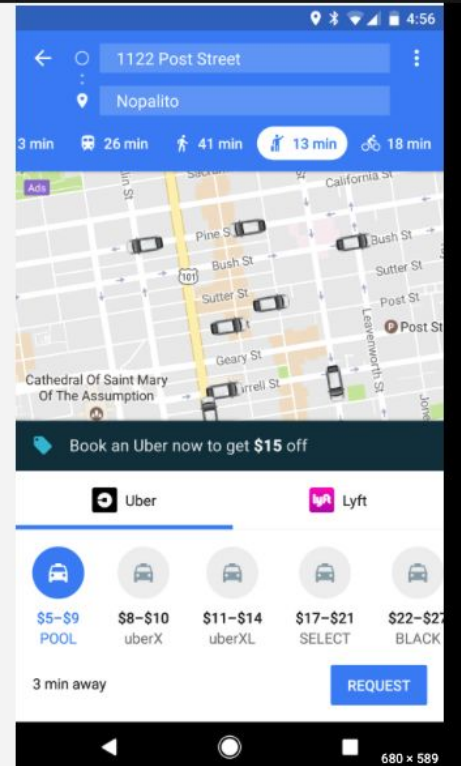
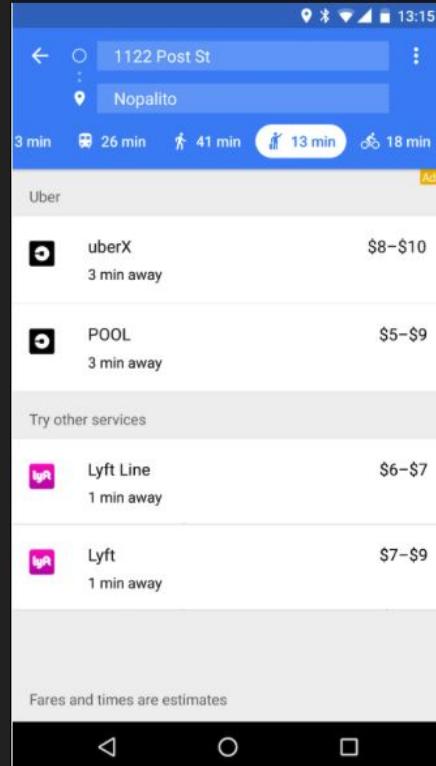




Name some good partnerships!

# Uber

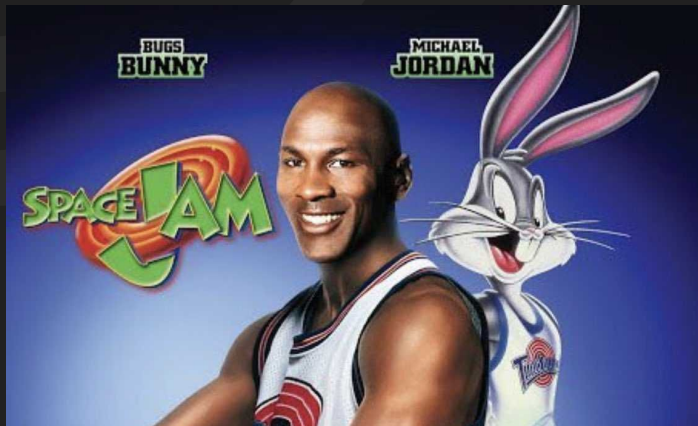
# Google





**APPLE'S  
BUSINESS PARTNERS**



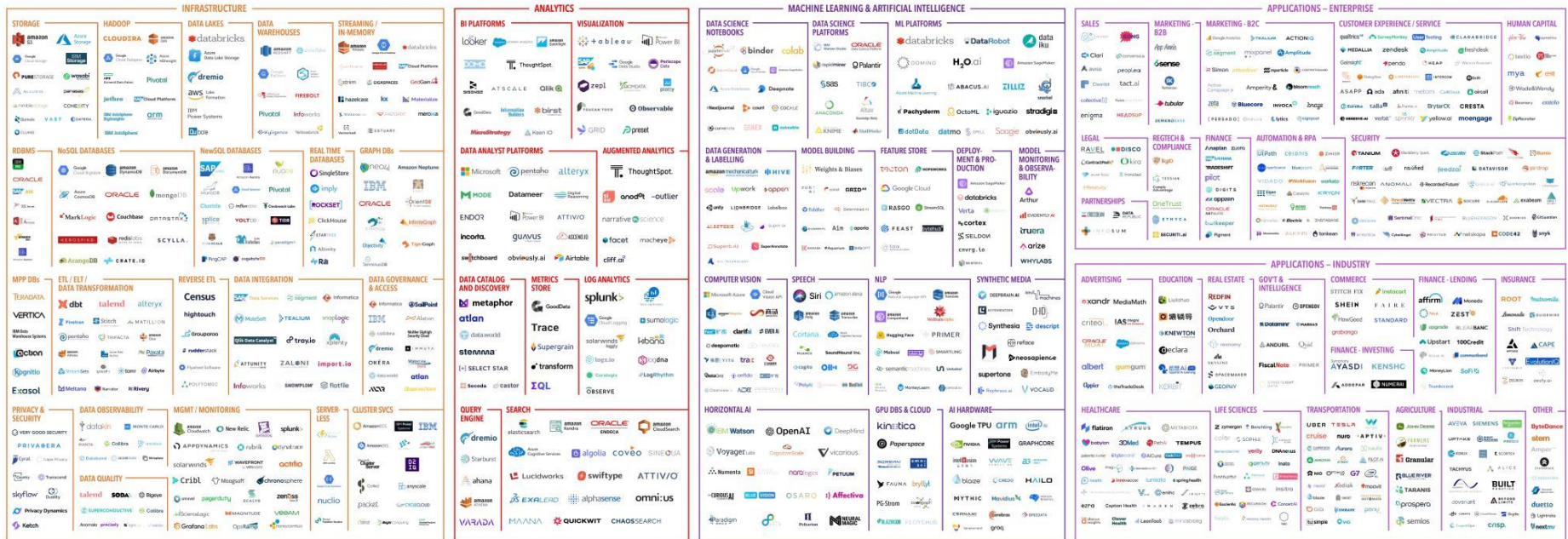








MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021



# Benefits of OSS Partnerships

1. Extra Resources: Discover and solve big complex problems faster.



# Benefits of OSS Partnerships

1. Extra Resources: Discover and solve big complex problems faster.
2. Increase reach: Expand the reach through your collaborators platform.



# Benefits of OSS Partnerships

1. Extra Resources: Discover and solve big complex problems faster.
2. Increase reach: Expand the reach through your collaborators platform.
3. Enhance functionality: By including your collaborators functionality your product is enhanced.



# Benefits of OSS Partnerships

1. Extra Resources: Discover and solve big complex problems faster.
2. Increase reach: Expand the reach through your collaborators platform.
3. Enhance functionality: By including your collaborators functionality your product is enhanced.
4. Increase credibility as a product and a player in the space.







# Community/Product Manager

**Goals:** Grow community, gain credibility, expand product functionality while minimizing risk



# Community/Product Manager

**Goals:** Grow community, gain credibility, expand product functionality while minimizing risk

This is done by...

- Connecting with community members & other communities
- Making the contribution process seamless
- Guiding design and resolving blockers
- Celebrating community achievements loudly!



# Contributing Engineer

**Goals:** Save money by spending time; contribute in order to gain a supported feature in the OSS project; build professional network



# Contributing Engineer

**Goals:** Save money by spending time; contribute in order to gain a supported feature in the OSS project; build professional network

This is done by...

- Contributing to ensure their favorite tools work together
- Collaborate with maintainers
- Building network with other contributing members
- Provide subject matter expertise





# Phases of Collaboration



## DISCOVER

- Engage & listen to community
- Initial outreach to other communities



# Phases of Collaboration



# Phases of Collaboration



- Build solution
- Test with community
- Documentation





# Phases of Collaboration



- Blog posts
- Webinars
- Community shoutouts



# Phases of Collaboration



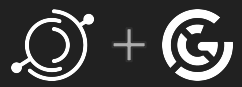
Cross-Community Partnership  
GX & Astronomer (Airflow)

# ASTRONOMER



great  
expectations











# GX & Astronomer (Airflow)

- A user problem is discovered that can be solved through collaboration.
  - The problem: "I want to get my data from A to B and when that happens I want to be confident in the quality of the data."





# GX & Astronomer (Airflow)

- A user problem is discovered that can be solved through collaboration.
  - The problem: "I want to get my data from A to B and when that happens I want to be confident in the quality of the data."
- Core teams meet, plan the work, and engage the community.







# Great Expectations

Official

Data Quality

An Apache Airflow provider for Great Expectations, an open-source data validation framework.

## Helpful Links

[View on GitHub](#)

[Using Providers](#)

[Learn Airflow](#)

VERSION

0.2.5

DOWNLOADS

380,477/month

LAST PUBLISHED

Feb. 17, 2023

## Quick Install

```
pip install airflow-provider-great-expectations==0.2.5
```



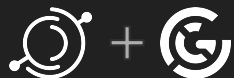
```
astronomer-logo-reverse
```





# GX & Astronomer (Airflow)

- A user problem is discovered that can be solved through collaboration.
  - The problem: "I want to get my data from A to B and when that happens I want to be confident in the quality of the data."
- Core teams meet, plan the work, and engage the community.
- The integration is released and promoted.
  - Documentation, blogs, community events, and more are provided.



Docs Find what you're looking for Learn About Astronomer Try Astro

Home Astro Astro CLI Software **Learn**

Overview  
Airflow concepts >  
Airflow tutorials >  
**Integrations** v  
Amazon Redshift  
Amazon SageMaker  
Apache Kafka/Confluent  
Azure Container Instances  
Azure Data Explorer  
Azure Data Factory  
Databricks  
dbt  
Fivetran  
**Great Expectations**

# Orchestrate Great Expectations with Airflow

**INFO**

You can now find the Great Expectations Provider on the Astronomer Registry, the discovery and distribution hub for Apache Airflow integrations created to aggregate and curate the best bits of the ecosystem.

**Great Expectations** is an open source Python-based data validation framework. You can test your data by expressing what you "expect" from it as simple declarative statements in Python, then run validations using those "expectations" against datasets with **Checkpoints**. Astronomer, with help from Superconductive, maintains an **Airflow provider** that gives users a convenient method for running validations directly from their DAGs.

This guide will walk you through how to use the [GreatExpectationsOperator](#) in an Airflow DAG and the Astronomer environment.

- Assumed knowledge
- Great Expectations concepts
- Setup
- Use Case: Great Expectations operator
- Configuration
- Using the Great Expectations operator
- Operator parameters
- Connections and backends
- Next steps

great expectations CURRENT v Search PRODUCT v COMMUNITY v RESOURCES v COMPANY v Cloud early access

Welcome  
Getting Started (A Tutorial) >  
Step 1: Setup >  
Step 2: Connect to data >  
Step 3: Create Expectations >  
Step 4: Validate data >  
**Integration guides** v  
Using Great Expectations with AWS >  
Deploying Great Expectations in a hosted environment without file system or CLI  
How to instantiate a Data Context on an EMR Spark cluster  
**How to Use Great Expectations with Airflow**

Integration guides > **How to Use Great Expectations with Airflow**

**Version: Current**

# How to Use Great Expectations with Airflow

This guide will help you run a Great Expectations checkpoint in Apache Airflow, which allows you to trigger validation of a data asset using an Expectation Suite directly within an Airflow DAG.

**Prerequisites: This how-to guide assumes you have:**

- Completed the Getting Started Tutorial
- Set up a working deployment of Great Expectations
- Created an Expectation Suite
- Created a checkpoint for that Expectation Suite and a data asset
- Created an Airflow DAG file

Install the [GreatExpectationsOperator](#)  
Using the [GreatExpectationsOperator](#)  
Operator Parameters  
Connections and Backends



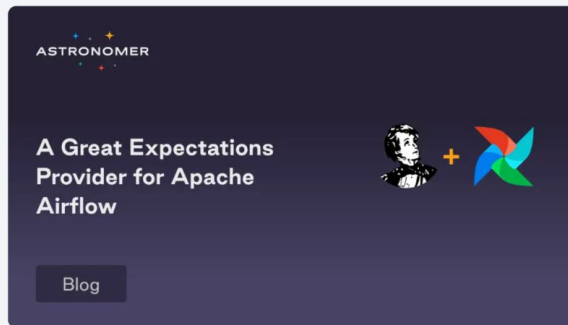
# A Great Expectations Provider for Apache Airflow

We're pleased to announce an official integration that allows users to leverage Great Expectations natively in their DAGs.

NOVEMBER 23, 2020



Pete DeJoy, Staff Product Manager



# An Airflow Operator for Great Expectations

November 30, 2020



SHARE THIS ARTICLE



LIKE OUR BLOGS?

Sign up for emails and get more blogs and news

Join the email list

## A Great Expectations Provider for Apache Airflow

We're pleased to announce an official integration that allows users to leverage Great Expectations natively in their DAGs.

NOVEMBER 23, 2020

ASTRONOMER

A Great Expectations  
Provider for Apache  
Airflow



*I also want to explicitly thank Brian Lavery, Nick Benthem, Bouke Nederstigt, and the Astronomer team (specifically Pete DeJoy) for dedicating their efforts to supporting this project, it's been an absolute joy collaborating with you!*

## An Airflow Operator for Great Expectations

November 30, 2020



SHARE THIS ARTICLE



LIKE OUR BLOGS?

Sign up for emails and  
get more blogs and news

[Join the email list](#)



# Optimize Your Data Pipeline with Apache Airflow and Great Expectations

Apache Application-experience



Shiv Deshmukh  
Thursday, April 21st, 2022  
7 min read

If you build or maintain applications that rely on data, you're familiar with both Apache Airflow and Great Expectations. [Apache Airflow](#) is an open source tool that manages workflows programmatically. It carries standard Python features to define complex DAGs, Acyclic Graphs, or DAGs, then execute the

## An Air

November 30, 2020



TEAMS INDIVIDUALS FEATURES BLOG CONTENT SPONSORSHIP

< VIEW ALL EVENTS

DATA ENGINEERING

# Build a robust data pipeline with Airflow, dbt, and Great Expectations

Published by [O'Reilly Media, Inc.](#)

Hands-on data validation in a modern data stack

<> Interactive

[What you'll learn](#) [Is this live event for you?](#) [Schedule](#)

Data quality has become a much discussed topic in the fields of data engineering and data science, and it's become clear that data validation is crucial to ensuring the reliability of data products produced by an organization's data pipelines. Apache Airflow and dbt (data build tool), among the prominent open source tools in the data engineering ecosystem, and while dbt offers data testing capabilities, another open source data tool, Great Expectations, enhances them with data validation and can add layers of robustness.

Join expert Sam Bail to explore the "dAG stack" and learn how to combine the functions of open source tools to build, test, validate, document, and orchestrate an entire pipeline, end-to-end from scratch.

### Hands-on learning with interactive scenarios

All exercises and labs are provided as [O'Reilly interactive scenarios](#)—complete development environments that are preconfigured with everything you need. There's nothing to install or configure; just click a link and get started!

Join the email list



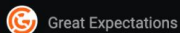
## Building a Robust Data Pipeline with the "dAG Stack": dbt, Airflow, and Great Expectations

BLOGS FROM ODSC SPEAKERS CONFERENCES MODELING AIRFLOW EAST 2021 GREAT EXPECTATIONS posted by ODSC Community

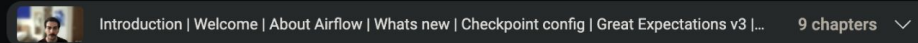


## How to use Great Expectations in an Airflow DAG to perform data quality checks.

4.2K views · 1 year ago



In this video, Benjamin Lampel goes over how to use Great Expectations in an Airflow DAG to perform data quality checks.

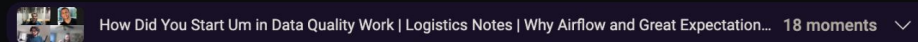


## Advanced Data Quality Use Cases with Airflow and Great Expectations

3.3K views · 1 year ago



For this webinar, Benji Lampel (Enterprise Platform Architect @ Astronomer) and Tal Gluck (Software Engineer ...



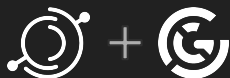
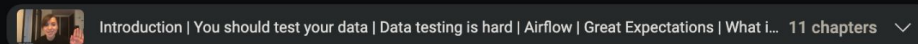
## Building a robust data pipeline with dbt, Airflow, and Great Expectations

26K views · 2 years ago



How do dbt and Great Expectations complement each other? In this video, Sam Bail of Superconductive will outline a convenient ...

CC






# GX & Astronomer (Airflow)

- A user problem is discovered that can be solved through collaboration.
  - The problem: "I want to get my data from A to B and when that happens I want to be confident in the quality of the data."
- Core teams meet, plan the work, and engage the community.
- The integration is released and promoted.
  - Documentation, blogs, community events, and more are provided.
- The community starts to enhance the offering, and the momentum continues to build.





 Published in Apache Airflow



Benji Lampel

Jan 19 · 4 min read · [Listen](#)



## Community Story: Airflow + Great Expectations at Factory Pal



### Who are we/what do we do?

We're Benji Lampel from Astronomer and Tino Pietrassyk from FactoryPal, here to tell you how FactoryPal has used [Great Expectations](#) and [Airflow](#) to ensure their factory optimization recommendations are the right ones.



Cross-Community Partnership  
GX & Astronomer (Airflow)

# ASTRONOMER



great  
expectations



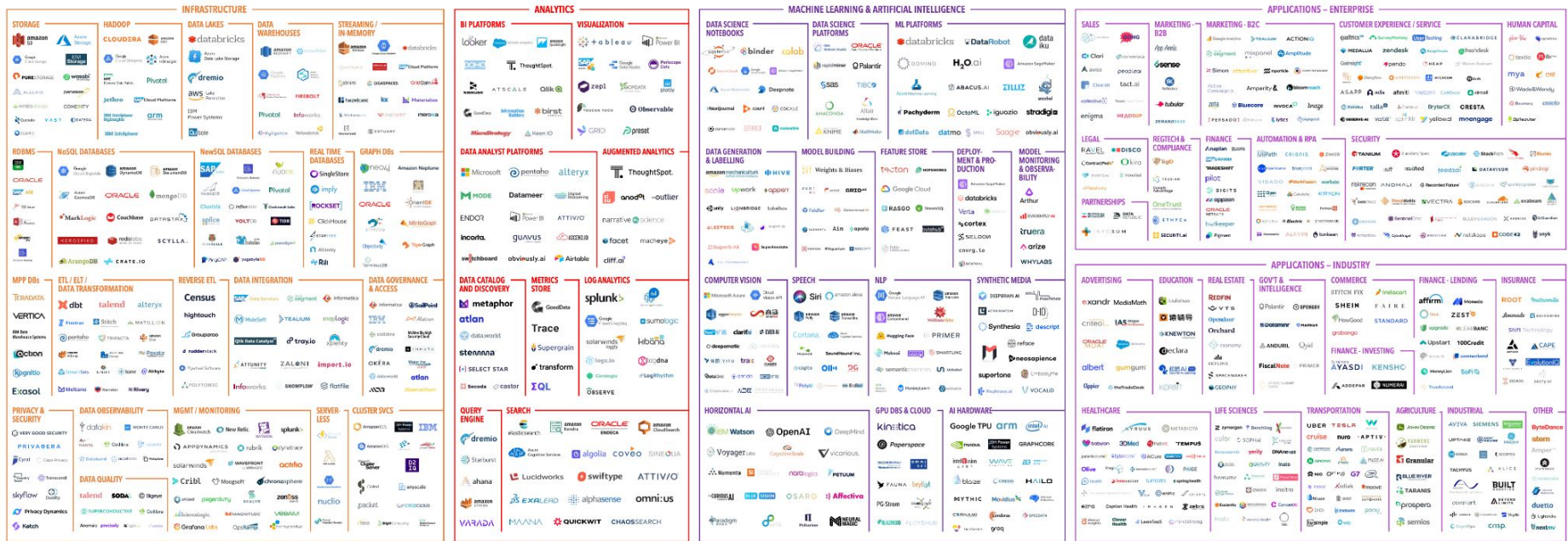


DataHub

Gradio



MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021



# Search across your entire Data Stack

The screenshot shows the Acryl DataHub search interface. The browser address bar displays the URL `longtailcompanions.acryl.io/search?page=1&query=pet`. The search bar contains the text "pet". The interface is divided into a left sidebar with filters and a main content area showing search results.

**Filter**

- Type**
  - Charts (28)
  - Datasets (19)
  - Dashboards (9)
  - Glossary Terms (5)
  - Tasks (4)
  - + More**
- Sub Type**
  - Table (7)
  - View (4)
  - Explore (4)
  - Source (3)
  - Incremental (1)
- Platform**
  - Looker (43)
  - Snowflake (6)
  - dbt (5)
  - Airflow (4)
  - AWS S3 (1)
  - + More**
- Domain**
  - Pet Adoptions (7)

**Showing 1 - 10 of 66 results**

**pet\_profiles**  
this table contains profile details of all pets  
Marketing Sensitive Confidential Tier1  
A R P A M  
Queried 1000+ times in the past month

**pet\_details**  
business critical  
A A  
Queried 1000+ times in the past month

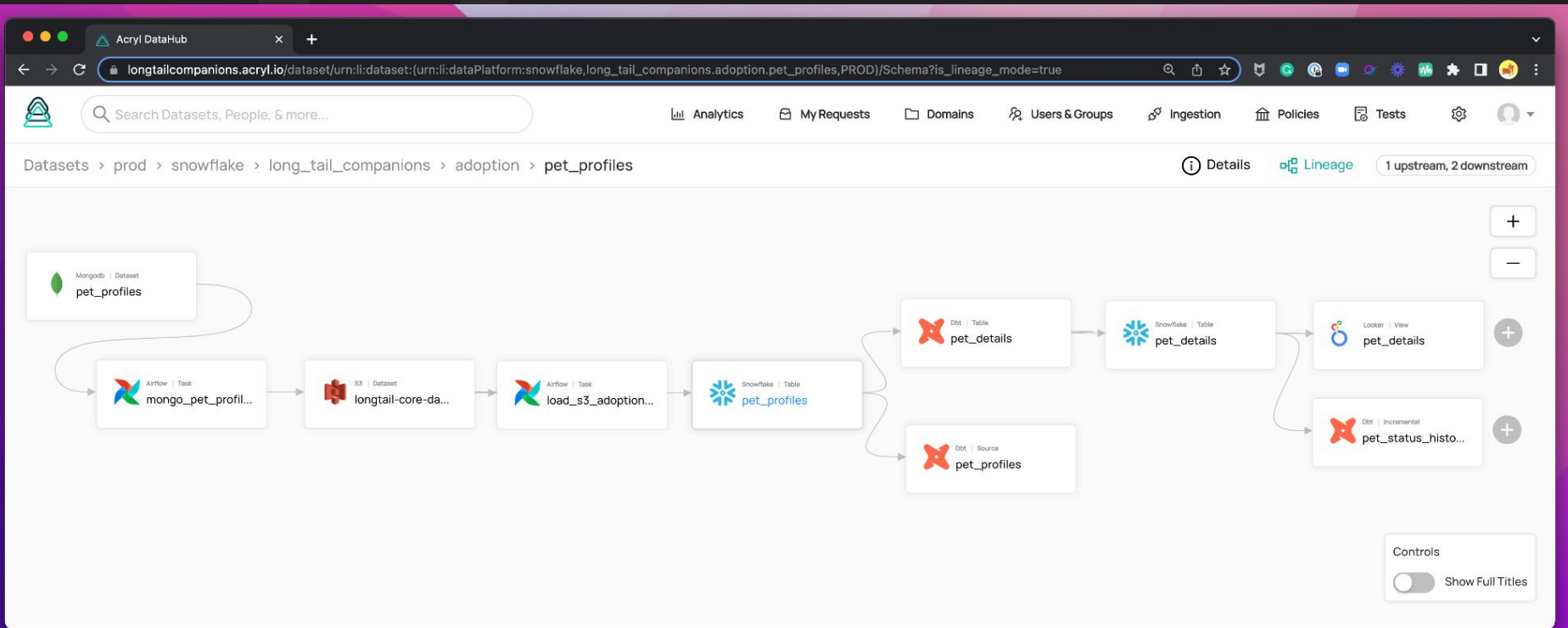
**mongo\_pet\_profile\_etl**  
Data Task Airflow

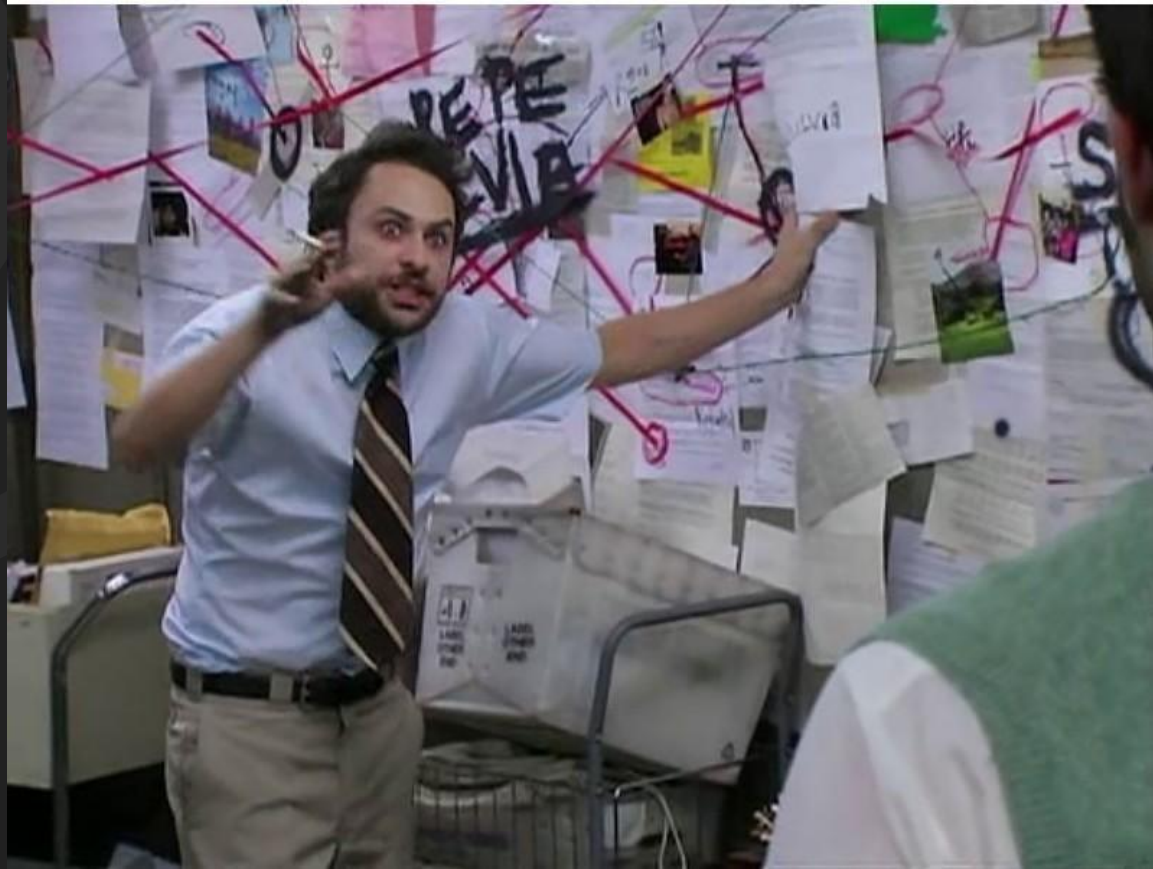
**load\_s3\_adoption\_pet\_profiles**  
Data Task Airflow

**postgres\_adoptions\_pets**  
Data Task Airflow



# Understand the end-to-end journey of data







# DataHub & Grab

- The Grab Team escalated a problem with the user experience
  - “Documenting and enriching assets in DataHub is intimidating for non-Markdown users”







Search Datasets, People, & more...

Select a View

Analytics

Ingestion

Govern



Schema Snowflake > LONG\_TAIL\_COMPANIONS

Share

## PRODUCT\_MANAGEMENT

41 entities

Entities **Documentation** Properties

Back

Save

**B I**

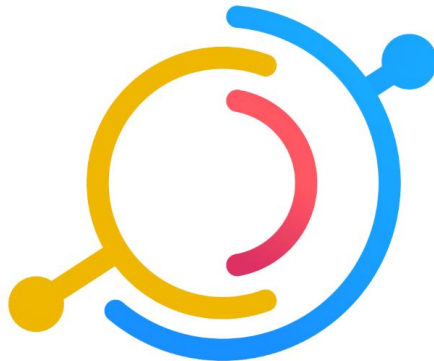
## Title of docs

This is some documentation with a "not-so-great" editor for "most" people.



### Title of docs

This is some documentation with a **not-so-great** editor for *most* people.



Last synchronized last week

#### About

Title of docs • This is some documentation with a not-so...

+ Add Link

#### Tags

No tags added yet. Tag entities to help make them more discoverable and call out their most important attributes.

+ Add Tags

#### Glossary Terms

Testing Term + Add Terms

#### Owners

No owners added yet. Adding owners helps you keep track of who is responsible for this data.

+ Add Owners

#### Domain

No domain set. Group related entities based on your organizational structure using by adding them to a Domain.

Set Domain



# DataHub & Grab

- The Grab Team escalated a problem with the user experience
  - “Documenting and enriching assets in DataHub is intimidating for non-Markdown users”
- Core teams meet, scope work and engage the community
  - Open-source Text Editor Framework
  - Support tagging/associating DataHub entities and DataHub Users
  - Upload files/images into documentation
  - Move away from markdown-first editors to remove friction for users not well-versed in .md formatting





**Polly** APP 1:02 PM

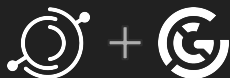
**Maggie** & the Core DataHub Team would like you to complete this polly

### 1min Poll! Would you be impacted if we removed Markdown Support for Entity documentation?

We are exploring different text editors to provide a richer user experience for managing documentation within DataHub. The goal is to store and render rich text which will likely result in deprecating Markdown support. Our migration strategy will ensure there is no interruption to rendering existing Markdown documentation within the UI, but we need to understand if there are common API-driven use cases within the Community for us to address as well.

Non-Anonymous | Results Will Not Be Shared | Oct 18, 1:00 PM - Oct 28, 1:00 PM

[Start Polly](#)





# DataHub & Grab

- The Grab Team escalated a problem with the user experience
  - “Documenting and enriching assets in DataHub is intimidating for non-Markdown users”
- Core teams meet, scope work and engage the community
- DataHub Maintainers are available for consultation/ongoing guidance



remirror





Search Datasets, People, & more... Select a View Analytics Ingestion Govern

Schema Showflake > LONG\_TAIL\_COMPANIONS

# PRODUCT\_MANAGEMENT

41 entities

Entities Documentation Properties

Share Last synchronized last week

About  
Title of docs • This is some documentation with a not-so...

Back

B I S H U L P C

## Title of docs

This is some documentation with a "not-so-great" editor for 'most' people.

![[https://raw.githubusercontent.com/datahub-project/datahub/566b7b089289129de3258486fca83f95b732964f/datahub-web-react/src/images/datahub-logo-color-stable.svg]]

## Title of docs

This is some documentation with a "not-so-great" editor for 'most' people.

Search Datasets, People, & more... Select a View Analytics Govern

Datasets > prod > snowflake > dbt\_test\_db > dbt\_hsheth Details Lineage 0 upstream, 1 downstream

Table Snowflake > DBT\_TEST\_DB > DBT\_HSHETH Share

# Pets

Schema Documentation Lineage Properties Queries Stats Validation

Back Save


Normal B I U S L P C

## Pets

this table contains records of all **pets** that have been considered for adoption regardless of outcome

Used by chart Weekly Adoption Projection

```
select * from pets;
```



|      | Dogs   | Cats  | Birds  |
|------|--|---|--|
| Pros | <ul style="list-style-type: none"><li>cute</li><li>fun</li><li>energetic</li></ul> | <ul style="list-style-type: none"><li>cute</li><li>low maintenance</li><li>cuddly</li></ul> | <ul style="list-style-type: none"><li>cute</li><li>can fly</li></ul> |

Last synchronized 3 months ago

About  
Pets • this table contains records of all pets that have b...

+ Add Link

Owners  
harshal x Analytics Engineering x  
+ Add Owners

Tags  
pii x + Add Tags

Glossary Terms  
Breed x + Add Terms

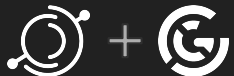
Domain  
No domain set. Group related entities based on your organizational structure using by adding them to a Domain.  
Set Domain



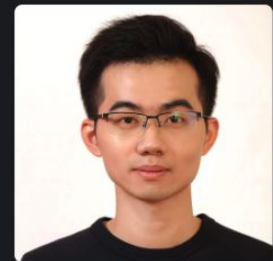


# DataHub & Grab

- The Grab Team escalated a problem with the user experience
  - “Documenting and enriching assets in DataHub is intimidating for non-Markdown users”
- Core teams meet, scope work and engage the community
- DataHub Maintainers are available for consultation/ongoing guidance
- Communicate outcomes via live events & release updates



Amanda Ng



Harvey Li

1. Extra Resources: Discover and solve big complex problems faster.
2. Increase reach: Expand the reach through your collaborators platform.
3. Enhance functionality: By including your collaborators functionality your product is enhanced.
4. Increase credibility as a product and a player in the space.



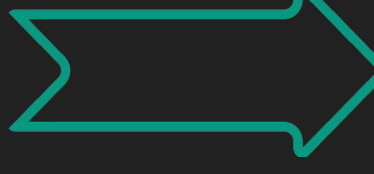
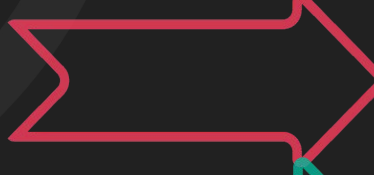
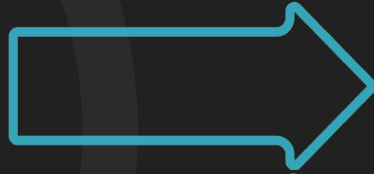
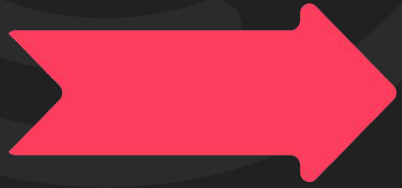


**PARTNERSHIPS GOOD**

imgflip.com







# Learnings from other OSS Communities

TODO: compile survey questions; send to folks at Meltano/Dagster/Astronomer/Trino/Shipyard/Prefect

## Questions for Community Managers/Product Managers

- What's been your most successful partnership?
- What's the one thing you do in order to make a partnership successful
- What best practices & anti-patterns/pitfalls to avoid
- What advice would you give to someone that's early in their partnership journey?
- What's the best way to enable people to make contributions?
- What's the one thing you must do to effectively engage with a contributor?

## Questions for Contributors

- What incentivizes you to contribute? What keeps you coming back?



# Contributing Consultants

**Goals:** Implement OSS for their customers; contribute to support customers' evolving needs

This is done by...

- Contributing to ensure their favorite tools work together
- Collaborate with maintainers
- Building network with other contributing members
- Provide subject matter expertise



# Contributing Consultants Example

- Consultants are on the ground floor with users of your product and will eventually understand their end to end problems. This gives them a broader vision that a team developing a single tool.
- In order to make their lives easier they will contribute to OSS platforms to ensure their desired technology stack works well together
- Consultants are also motivated by promotion of their services. This can result in some excellent content promoting your tool along with their work. It's great to create relationships with these groups to ensure your on the same page and can coordinate on future contributions to the project.

