DATA+AI
SUMMIT 2022

# Connecting the dots with DataHub

## LakeHouse and beyond


Shirshanka Das
CEO and Co-Founder
Acryl Data


ORGANIZED BY databricks

# Hello!

Shirshanka Das

CEO and Co-Founder, Acryl Data

Founder, DataHub

tweets at @shirshanka

# About Acryl Data

## Company

Founded early 2021 by data engineers from LinkedIn, Airbnb

## What we do

Bring clarity to complex data ecosystems by driving forward the open source DataHub project

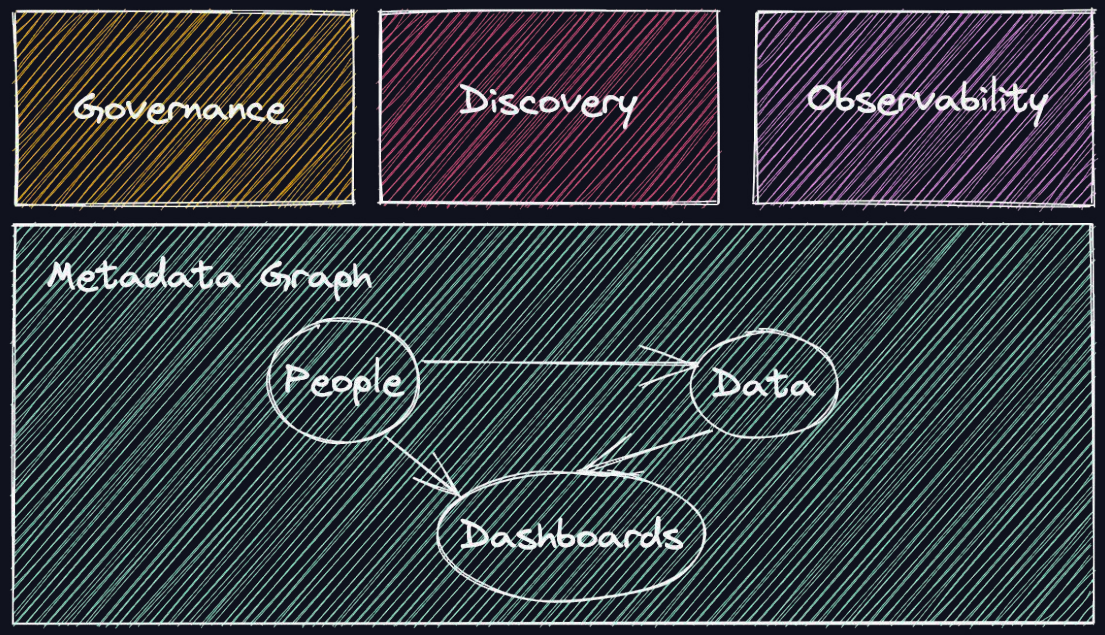
## Team

14 FTE, 3 interns, 5+ puppers

Learn more at acryldata.io

# What is DataHub?

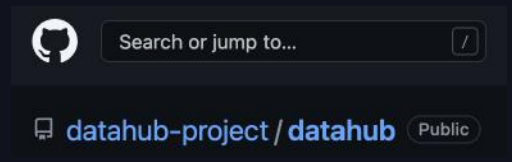
DataHub is an open source metadata platform that enables Data Discovery, Data Observability, and Federated Governance on top of a high-fidelity Metadata Graph.




Learn more datahubproject.io

# DataHub: #1 Open Source Metadata Platform

Integrations




Adopters

udemy
PELOTON
in
zynga
Optum
Expedia
ThoughtWorks
Klarna.
Viasat
MOLOCO
Adevinta

# DataHub Community

slack.datahubproject.io

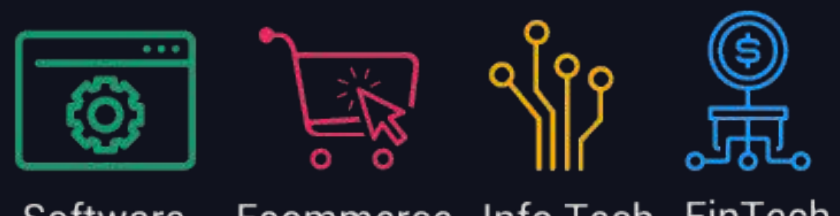
## 3,400+ Slack Members

10x YoY Growth

Across 56 Countries & 27 Local Time Zones

### Top Member Roles


### Top Member Industries



# DataHub Community

**950+**
Weekly Active Slack Members

**240+**
OSS Contributors

**10.5k+**
Monthly Slack Messages

**200+**
Monthly Commits

**130+**
Monthly Town Hall Attendees

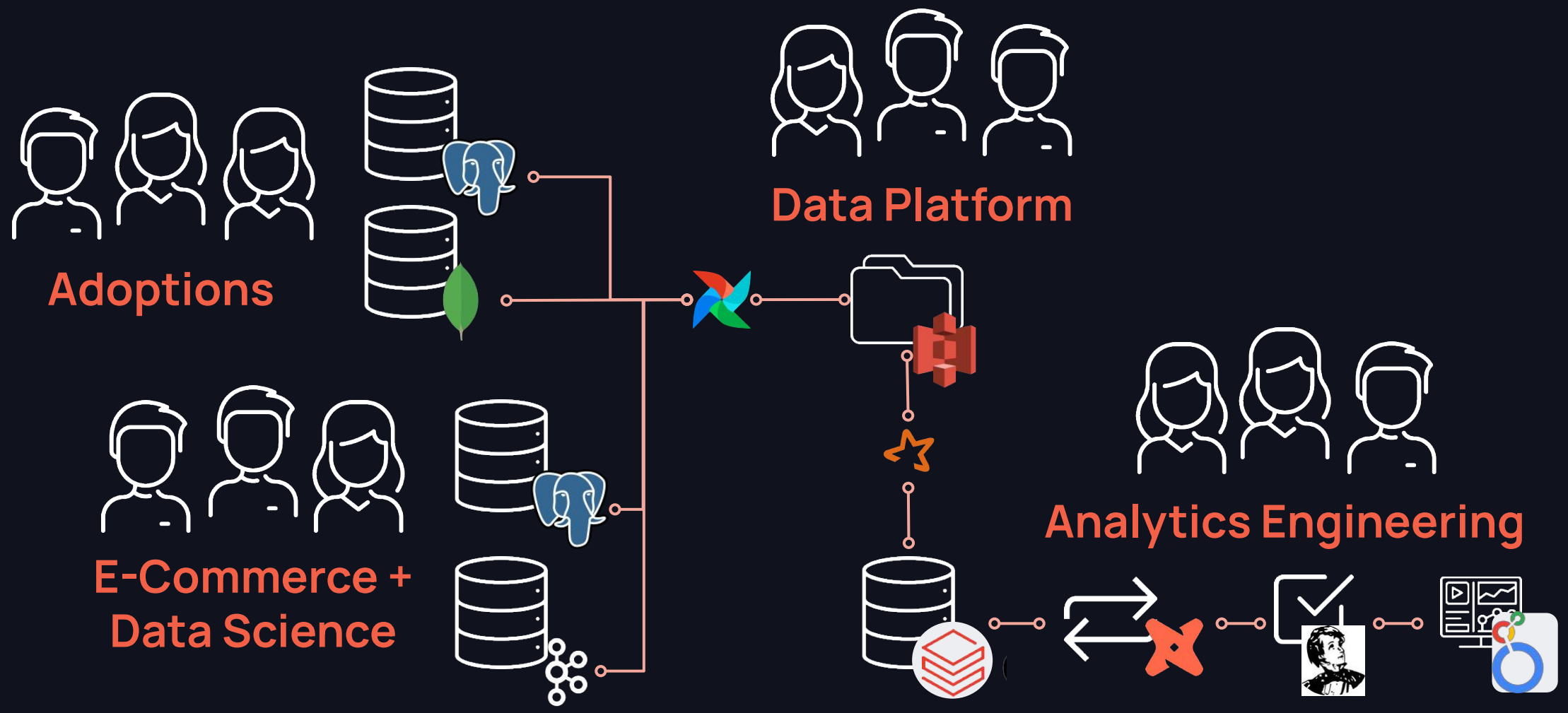
**5.6k+**
GitHub Stars

# What can you do with DataHub?

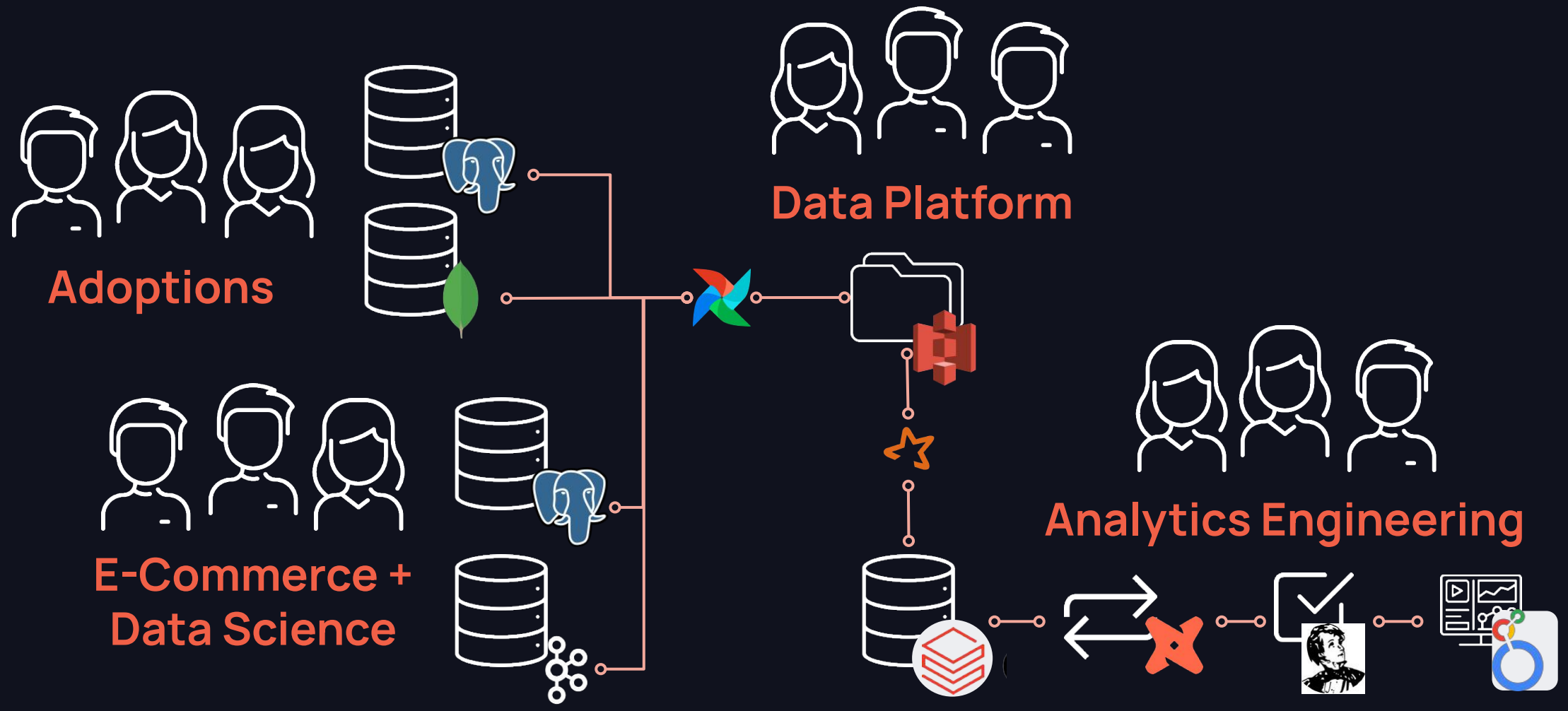
# Example: Long Tail Companions

Where Every Pet is Exceptional

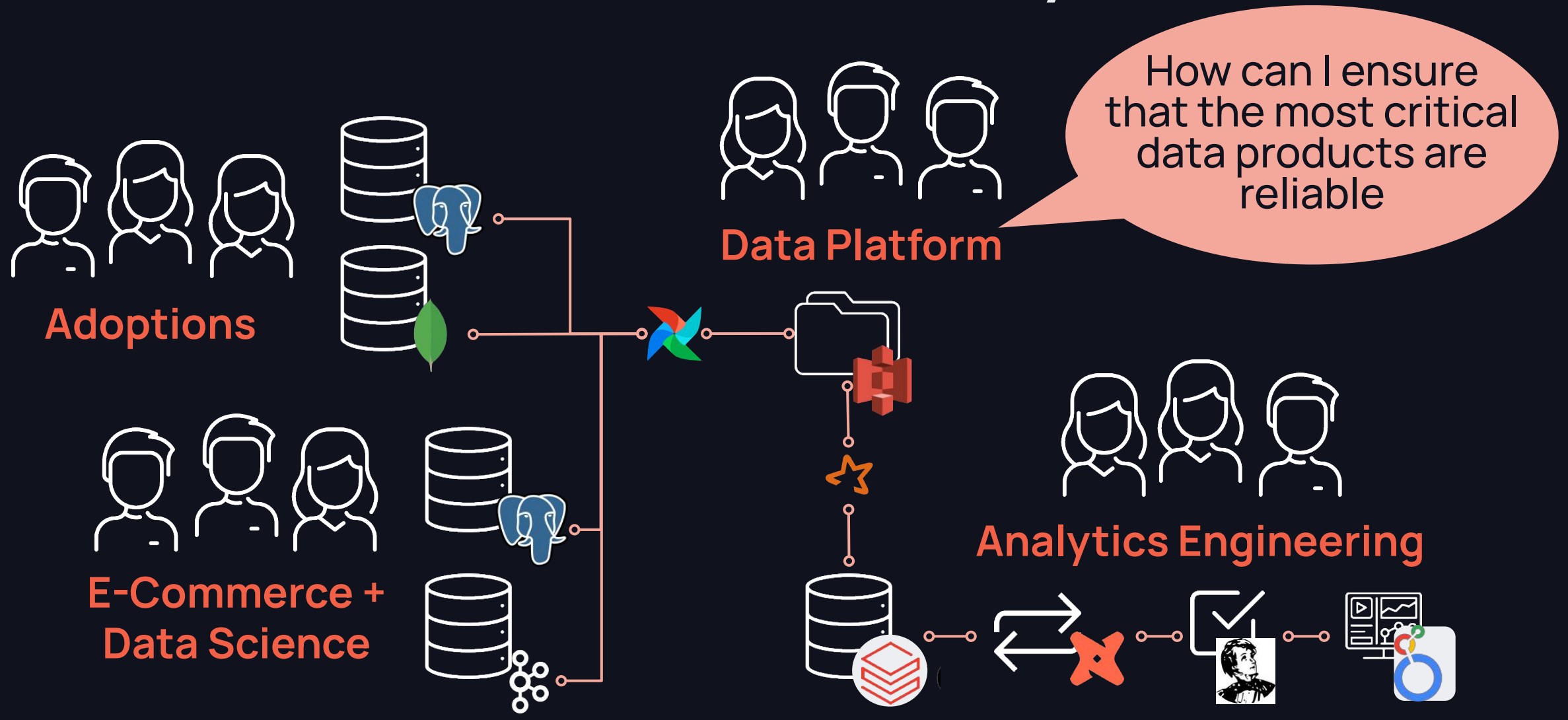
# Long Tail Companions' Fragmented Data Stack



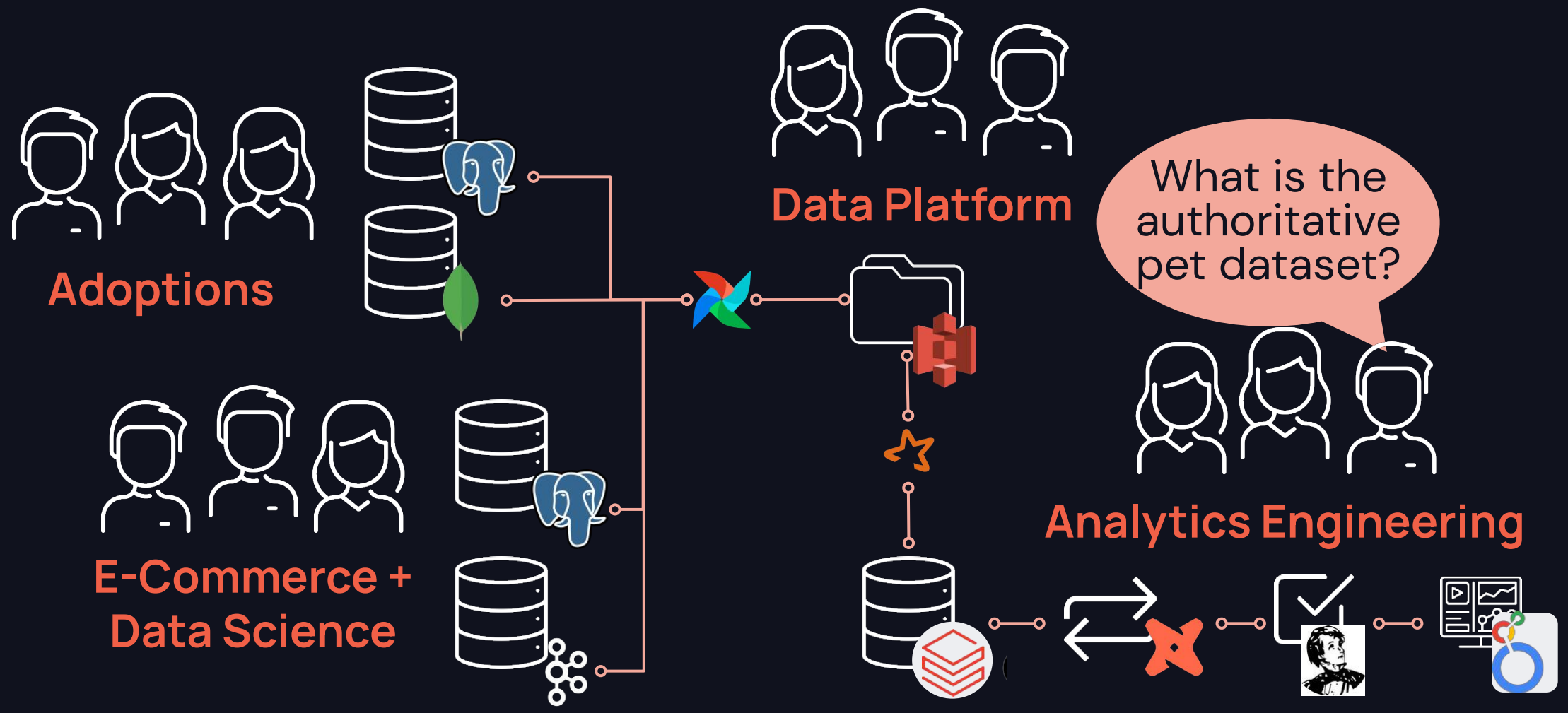
# LTC Data Practitioners Routinely Ask:



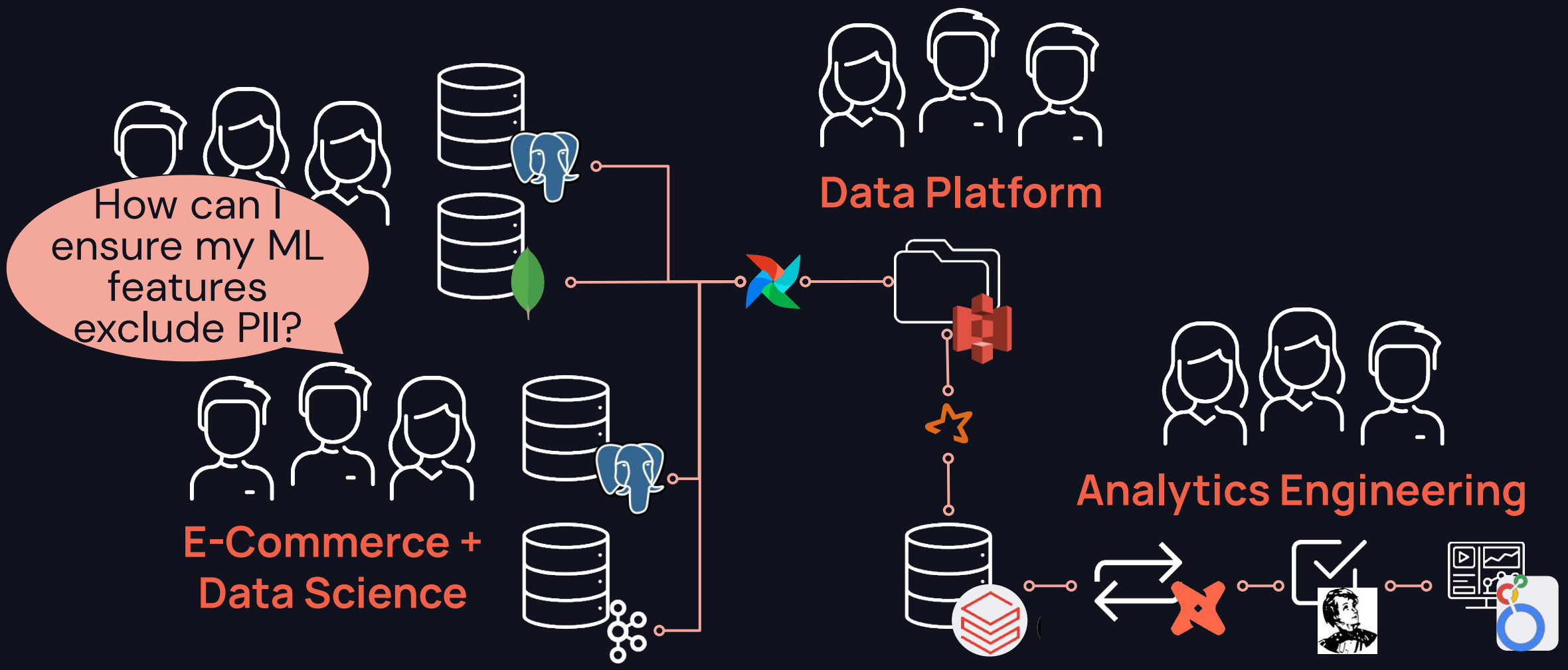
# LTC Data Practitioners Routinely Ask:



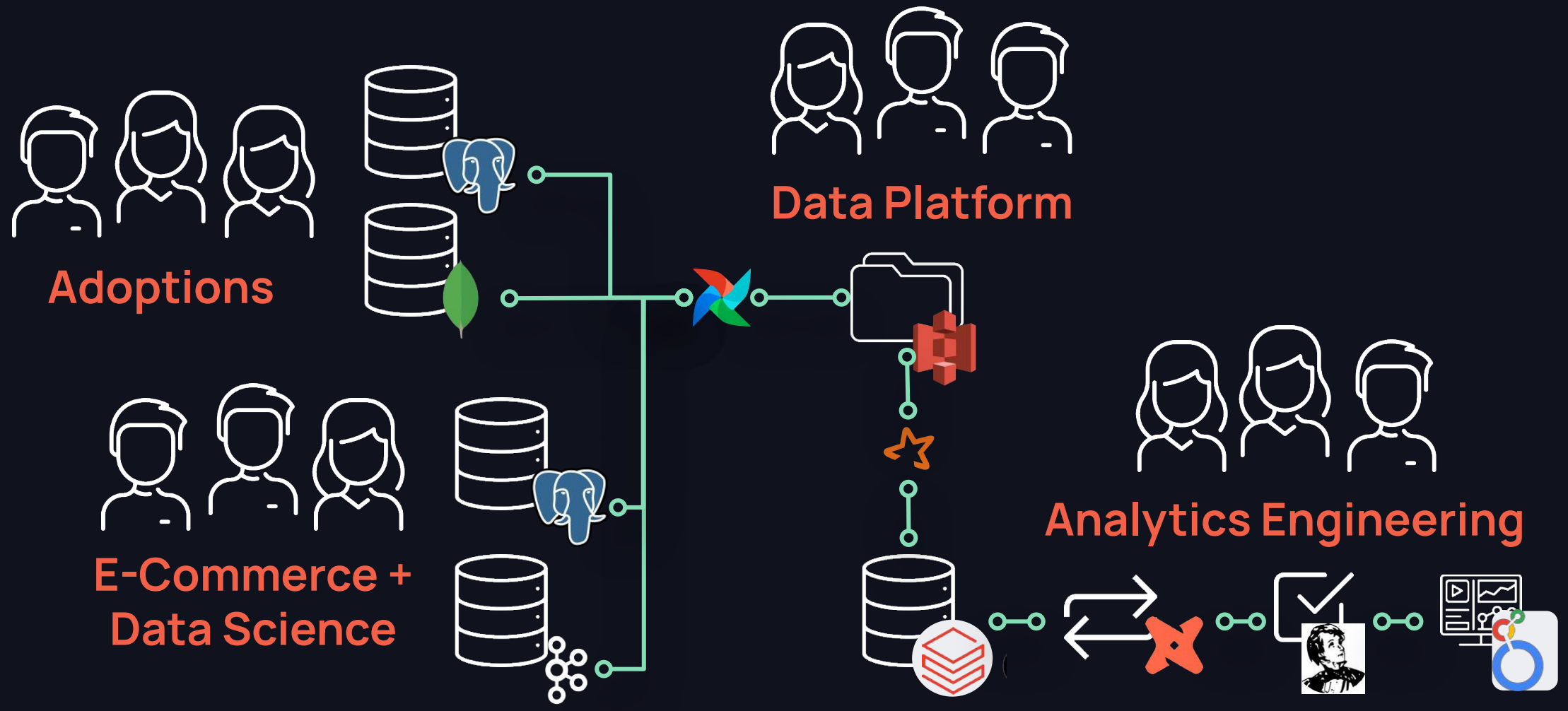
# LTC Data Practitioners Routinely Ask:



# LTC Data Practitioners Routinely Ask:



# The Key to the Solution: Metadata!



# 3 Must-Haves for Metadata Management

Metadata 360

Shift Left

Active Metadata

# What Does this Look Like in Practice?

# Metadata 360

## Integrations: DataBricks Hive

DataHub

Docs Demo Blog Feature Requests Roadmap

- ClickHouse
- CSV
- Data lake files
- dbt
- Delta Lake
- Druid
- Elastic Search
- Feast
- File Based Lineage
- File
- Glue
- SAP HANA
- Hive

```
# ------------------------------------------------------------
# Recipe (Databricks)
# Ensure that databricks-dbapi is installed. If not, use ```pip install databricks-dbapi``` to install.
# Use the ```http_path``` from your Databricks cluster in the following recipe.
# See (https://docs.databricks.com/integrations/bi/jdbc-odbc-bi.html#get-server-hostname-port-http-path-and-j
# ------------------------------------------------------------

source:
  type: hive
  config:
    host_port: <databricks workspace URL>:443
    username: token
    password: <api token>
    scheme: 'databricks+pyhive'

    options:
      connect_args:
        http_path: 'sql/protocolv1/o/xxxyyyzzzaaasa/1234-567890-hello123'

sink:
  # sink configs
```

# Metadata 360

## Integrations: Delta Lake (new!)

DataHub

Docs Demo Blog Featur

Sources

- Athena
- Azure AD
- BigQuery
- Business Glossary
- ClickHouse
- CSV
- Data lake files
- dbt
- Delta Lake
- Druid
- Elastic Search
- Feast
- File Based Lineage

# Delta Lake

## Module `delta-lake`

support status incubating

### Important Capabilities

| Capability | Status | Notes |
|---|---|---|
| Extract Tags | ✅ | Can extract S3 object/bucket tags if enabled |

This plugin extracts:

- Column types and schema associated with each delta table
- Custom properties: number_of_files, partition_columns, table_creation_time, location, version etc.

```yaml
source:
  type: delta-lake
  config:
    env: "PROD"
    platform_instance: "my-delta-lake"
    base_path: "/path/to/data/folder"

sink:
  # sink configs
```

```
pip install acryl-datahub[delta-lake]
```

# Metadata 360

## Integrations: Delta Lake (new!)

Search Datasets, People, & more...

Datasets › prod › delta-lake › long_tail_companions › adoption › pet_profiles

Dataset | Delta Lake

**long_tail_companions/adoption/pet_profiles**

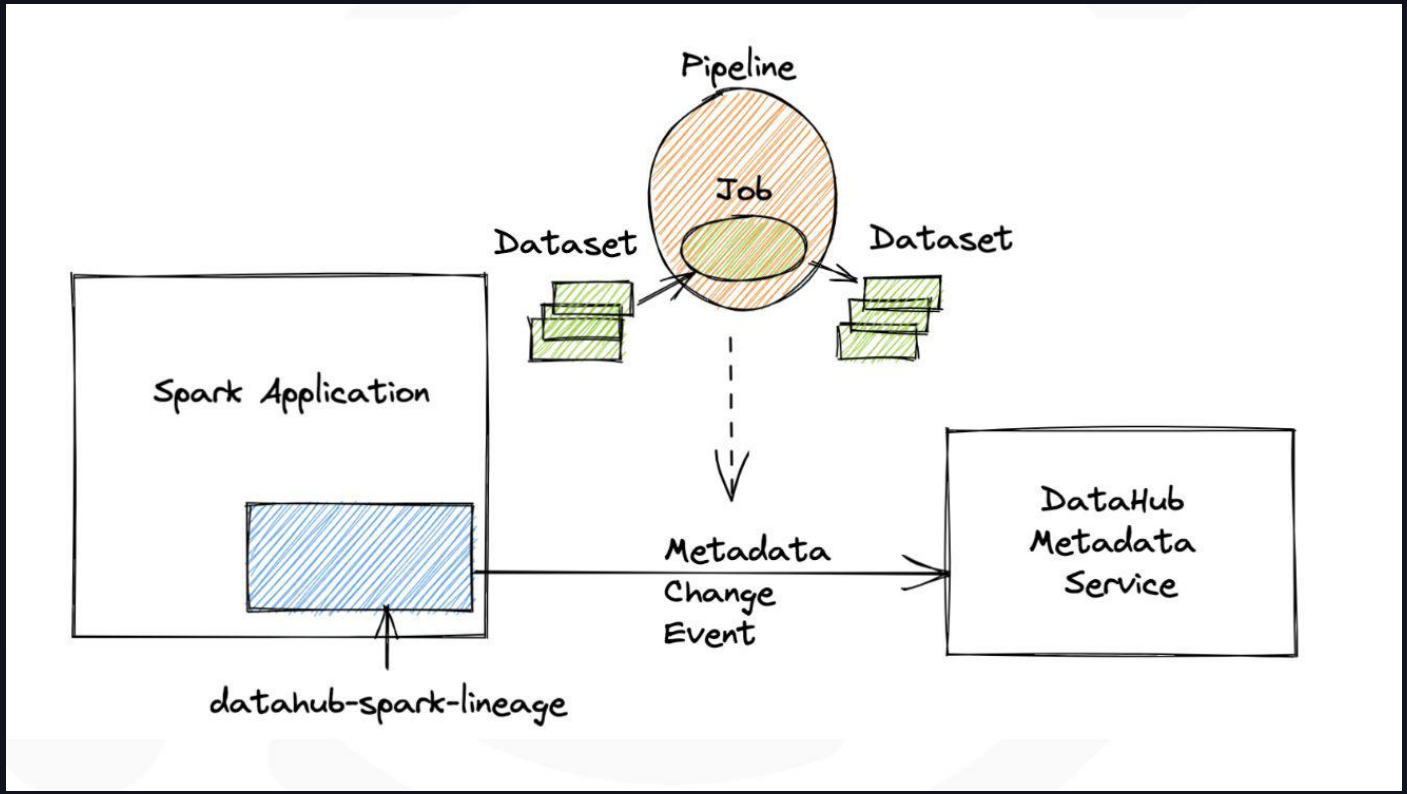
Schema | Documentation | Properties | Lineage | Queries | Stats | Validation

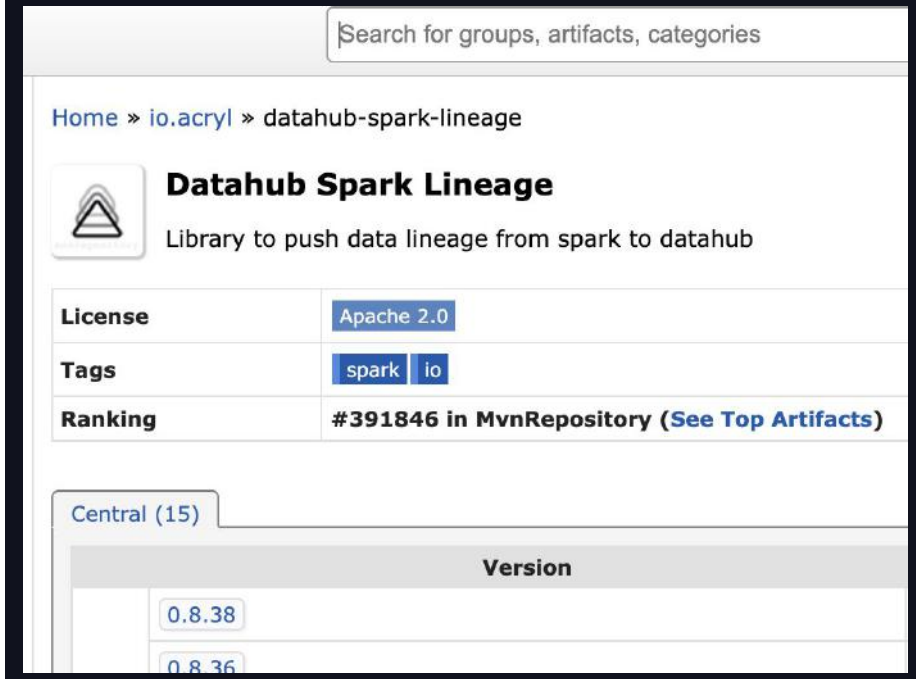
Reported at 29/06/2022, 11:21 GMT+5:30 | Normal | Blame | 0.0.0 - 14 seconds ago

| Field | Description | Tags | Terms |
|---|---|---|---|
| profile_id String | Unique identifier of pet profile | | |
| species String | Species of pet, either feline or canine | | |
| breed String | Breed of pet as determined during intake | | |
| sex String | Sex of pet | | |
| color String | Color of pet, as determined druing intake | | |
| coat_type String | Coat type of pet, as determined druing intake | | |
| name String | Name assigned to pet during intake | | |

# Metadata 360

## Integrations: Push-based Integration with Spark

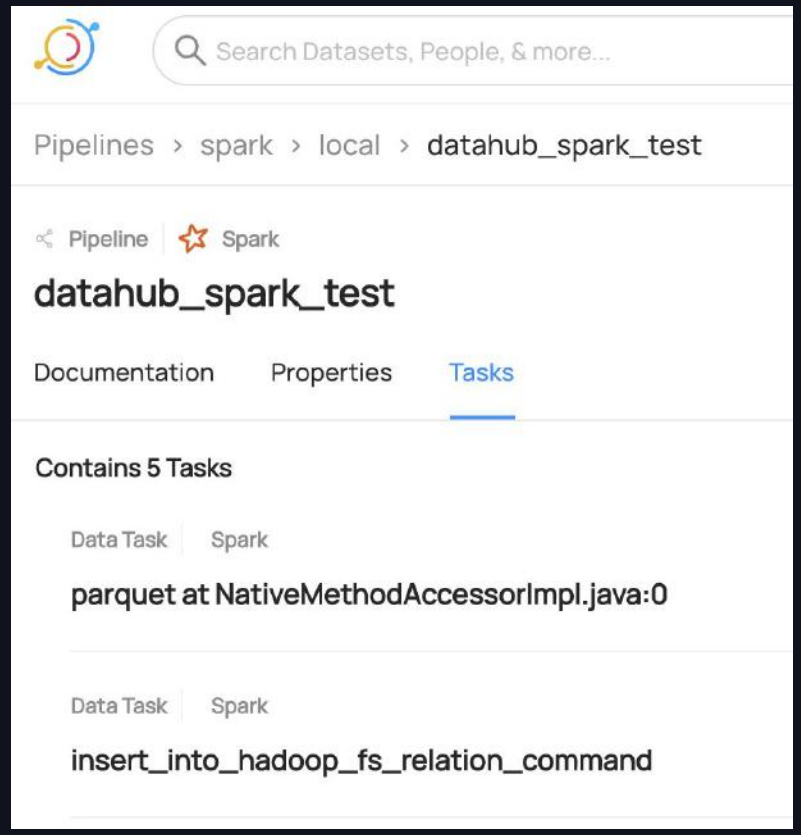


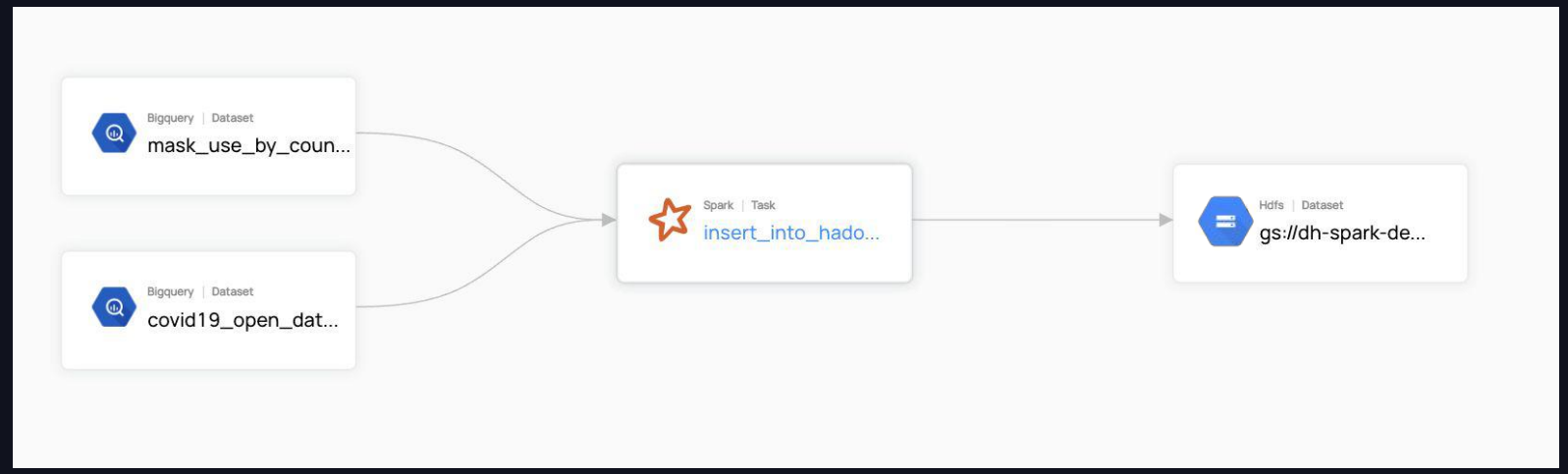




```
.config("spark.jars.packages", ",".join(jar_packages))
.config("spark.extraListeners", "datahub.spark.DatahubSparkListener")
.config("spark.datahub.rest.server", "http://datahub-gms:8080")
```

# Metadata 360

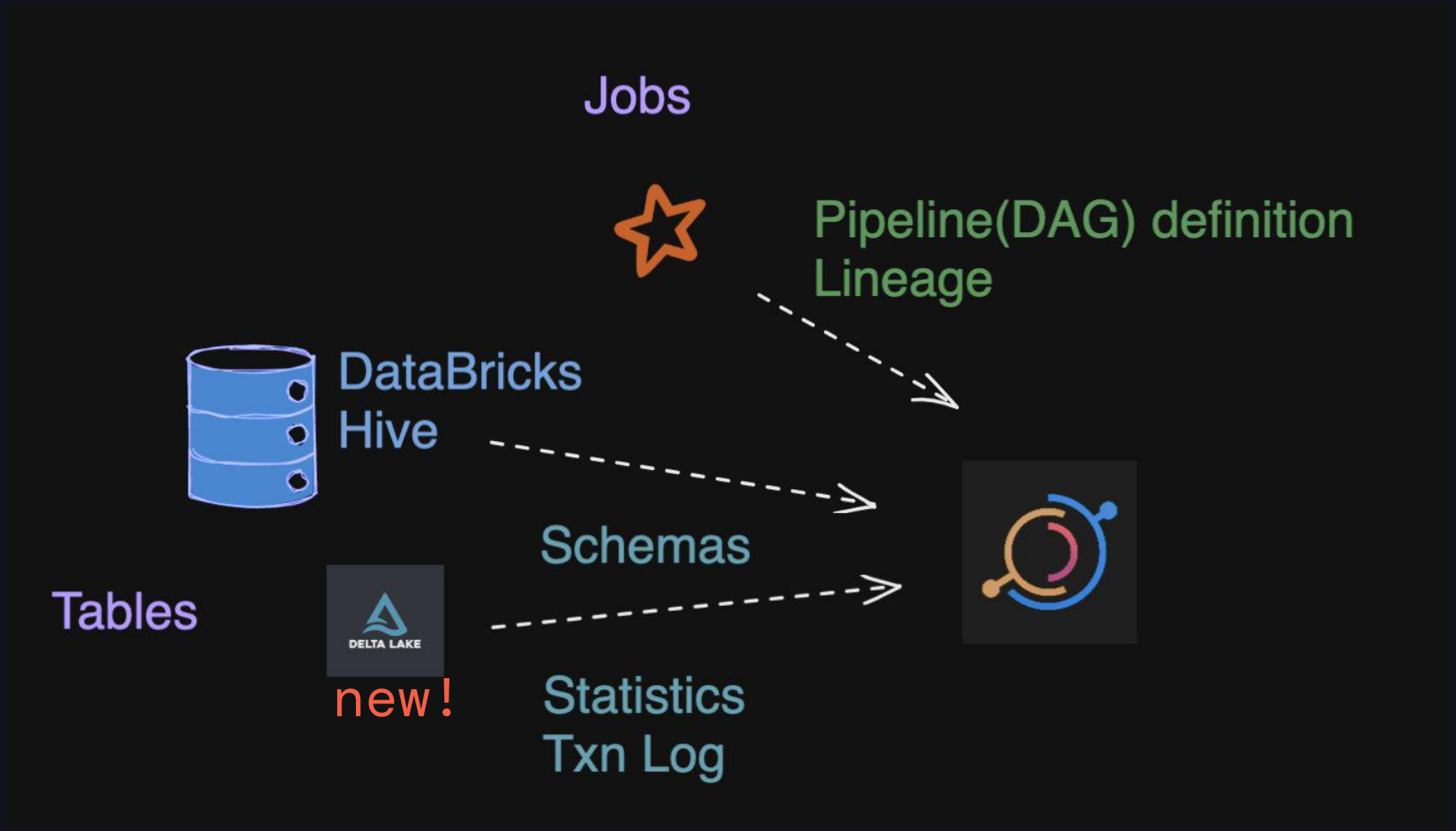
## Integrations: Push-based Integration with Spark



# Metadata 360

## Integrations Recap



# Metadata 360

## Searching for pet_profile

Acryl DataHub

longtailcompanions.acryl.io/search?page=1&query=pet_profile

DataHub PRs | acryl github | https://github.com... | Business Chats | AWS Management... | datahub-apps | trigger-happy | acryldata/datahub... | GitHub Assigned | PR- acryldata/dat... | Use Lineage for I...

pet_profile

Analytics | My Requests | Ingestion | Tests | Govern

Filter

Showing 1 - 10 of 27 results

Type
- Datasets (15)
- Charts (7)
- Tasks (2)
- Dashboards (2)
- Glossary Terms (1)

Tag

Glossary Term
- DaysInStatus (2)
- Confidential (1)

Domain
- Pet Adoptions (5)
- Finance (1)
- Marketing (1)

Owned By
- maggie@acryl.io (9)
- Adoption (4)
- Shannon Lovett (3)
- Admin (2)
- Avigdor Bramhall (1)
- + More

Sub Type

Platform
- Looker (14)

Data Task | Airflow
mongo_pet_profile_etl

Data Task | Airflow
load_s3_adoption_pet_profiles

Glossary Term
ReturnRate
The percentage of adopted animals that were returned to the shelter within 60 days of adoption • with adoption_date as ( • select • profile_id • , asofdate as adoption_date • from • analytics.petstatushistory • where • status = 'adopted' • ) ...

Table | DataBricks › long_tail_companions › adoption
pet_profiles
this table contains profile details of all pets
Marketing | Confidential | Tier1

Source | Dbt
pet_profiles
Pet Adoptions

Dataset | MongoDB
pet_profiles
Production data in Mongo storing the initial profile created when a pet is first entered into LTC system.
Pet Adoptions | prod | business critical

Dataset | AWS S3

# Metadata 360

## Dive into technical metadata

Search Datasets, People, & more...

Analytics | My Requests | Ingestion | Tests | Govern

Datasets › prod › databricks › long_tail_companions › adoption › pet_profiles

Details | Lineage | 1 upstream, 2 downstream

Table | DataBricks › long_tail_companions › adoption

pet_profiles ✓

Schema | Documentation | Properties | Lineage | Queries | Stats | Validation | Incidents

View Technical Schema

Reported at 3/20/2022, 09:23 PM PDT | Normal | Blame | 0.0.0 - 4 months ago

| Field | Description | Tags | Terms |
|---|---|---|---|
| profile_id String | Unique identifier of pet profile | | Sensitive |
| species String | Species of pet, either feline or canine | | Confidential |
| breed String | Breed of pet as determined during intake | (edited) | Breed |
| sex String | Sex of pet | | Confidential |
| color String | Color of pet, as determined druing intake | | Email |

**About**

this table contains profile details of all pets

+ Add Link

**Stats** More stats ›

Rows 43,517

Columns 13

Last Updated 6/7/2022, 03:01 AM PDT

**Tags**

Tier1 | + Add Tags

**Glossary Terms**

Confidential | Sensitive | + Add Terms

# Metadata 360

## Dive into technical metadata

Search Datasets, People, & more...

Analytics | My Requests | Ingestion | Tests | Govern

Datasets › prod › databricks › long_tail_companions › adoption › pet_profiles

Details | Lineage | 1 upstream, 2 downstream

Table | DataBricks › long_tail_companions › adoption

**pet_profiles** ✓

Schema | Documentation | Properties | Lineage | Queries | Stats | Validation | Incidents

Reported at 3/20/2022, 09:23 PM PDT | Normal | Blame | 0.0.0 - 4 months ago

| Field | Description | Tags | Terms |
|---|---|---|---|
| profile_id String | Unique identifier of pet profile | | Sensitive |
| species String | Species of pet, either feline or canine | | Confidential |
| breed String | Breed of pet as determined during intake | (edited) | Breed |
| sex String | Sex of pet | | Confidential |
| color String | Color of pet, as determined druing intake | | Email |

About

**Understand Usage**

Stats | More stats ›

Rows: 43,517

Columns: 13

Monthly Queries: 2710

Top Users: B M +26

Last Updated: 5/23/2022, 07:12 PM CDT

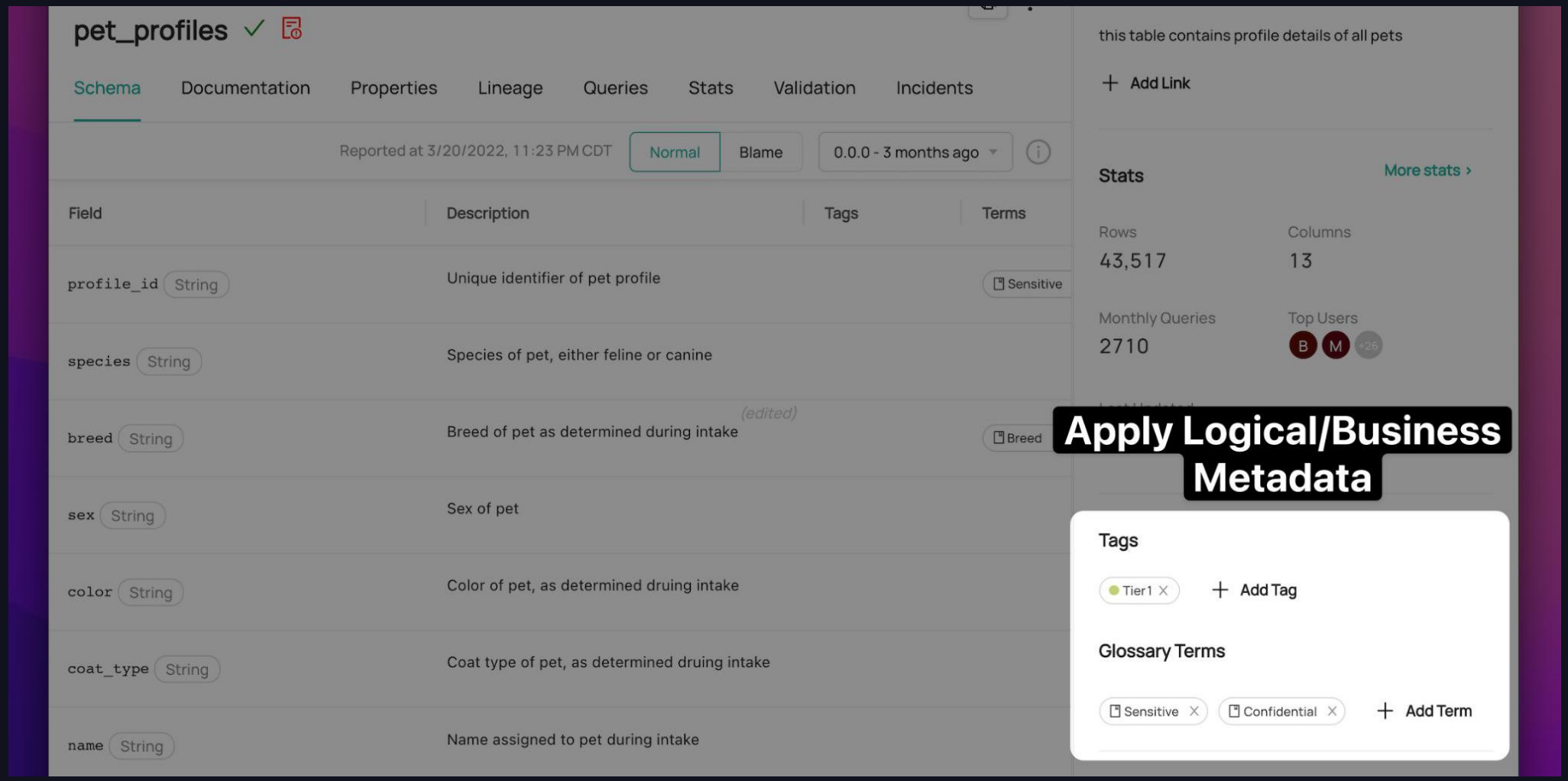
Tags

Tier1 | + Add Tags

Glossary Terms

# Metadata 360

## Understand Business Context



# Metadata 360

## Identify the right humans

| | | |
|---|---|---|
| profile_id String | Unique identifier of pet profile | Sensitive |
| species String | Species of pet, either feline or canine | |
| breed String | Breed of pet as determined during intake (edited) | Breed |
| sex String | Sex of pet | |
| color String | Color of pet, as determined druing intake | |
| coat_type String | Coat type of pet, as determined druing intake | |
| name String | Name assigned to pet during intake | |
| age_m Number | Approximate age of pet - months | |
| age_y Number | Approximate age of pet - years | |

Monthly Queries
2710

Top Users
B M +26

Last Updated
5/23/2022, 07:12 PM CDT

Tags
Tier1 ×
+ Add Tag

Glossary Terms

**Assign Ownership**

Owners
A Adoption ×
R Ricca MacAnespie ×
P Prentiss Matussov ×
A Avigdor Bramhall ×

# Metadata 360

## Live and Historical Views

Search ... | Analytics | My Requests | Domains | Users & Groups | Ingestion | Policies | Tests

Datasets › prod › snowflake › long_tail_companions › adoption › pet_profiles | Details | Lineage | 1 upstream, 2 downstream

Table | Snowflake › long_tail_companions › adoption

pet_profiles

**View Operational Detail**

Schema | Documentation | Properties | Lineage | Queries | Stats | Validation | Incidents

Reported at 3/20/2022, 11:23 PM CDT | Normal | Blame | 0.0.0 - 3 months ago

| Field | Description | Tags | Terms |
|---|---|---|---|
| profile_id String | Unique identifier of pet profile | | Sensitive |
| species String | Species of pet, either feline or canine | | |
| breed String | (edited) Breed of pet as determined during intake | | Breed |

About

this table contains profile details of all pets

+ Add Link

Stats | More stats ›

Rows: 43,517

Columns: 13

Monthly Queries: 2710

Top Users: B M +26

Last Updated: 5/23/2022, 07:12 PM CDT

# Metadata 360

## Stats

Acryl DataHub

longtailcompanions.acryl.io/dataset/urn:li:dataset:(urn:li:dataPlatform:snowflake,long_tail_companions.adoption.pet_profiles,PROD)/Stats?is_lineage_mode=false

Showing profile from 3/17/2022 at 3:53:00 AM

### Table Stats

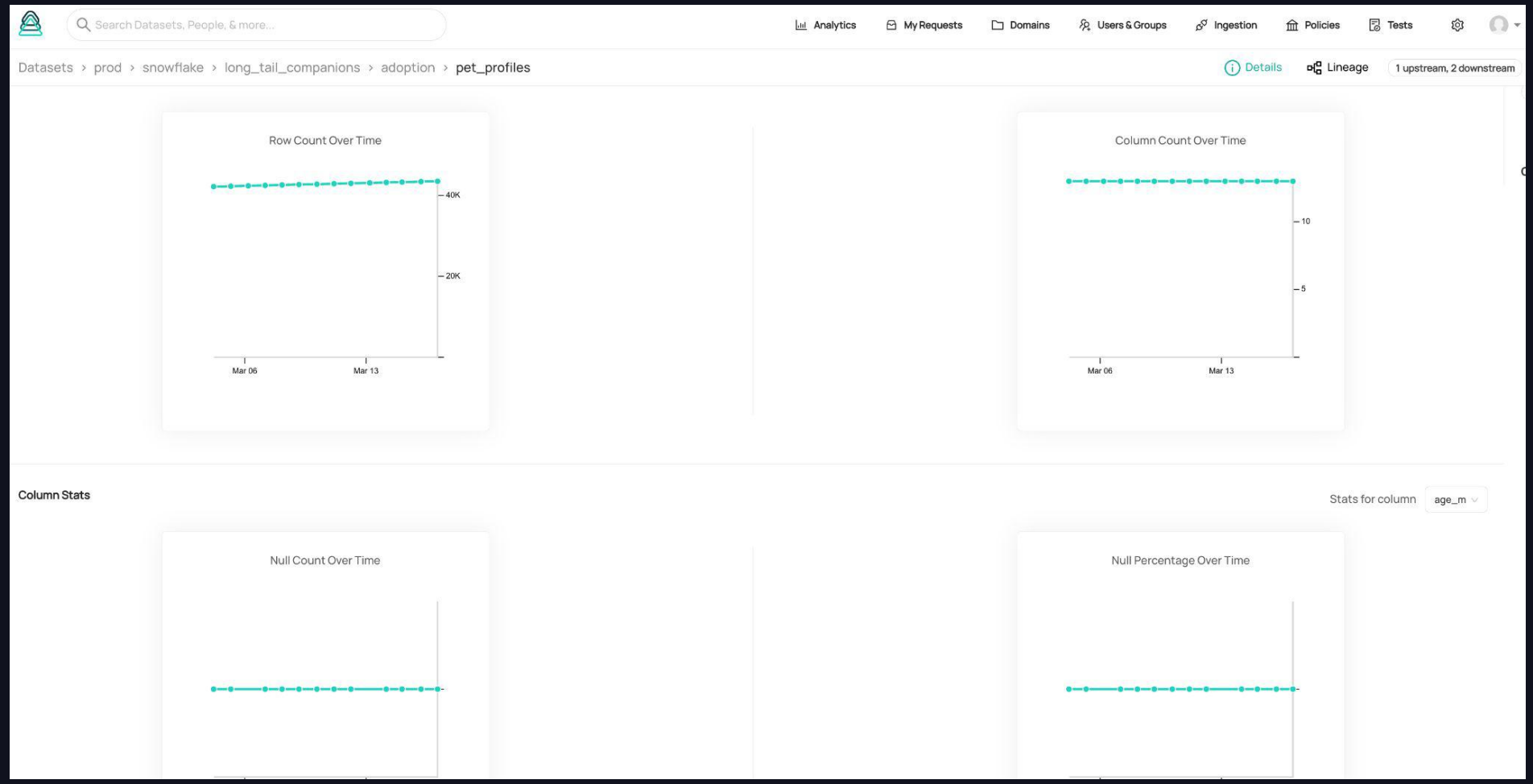
| Rows | Columns |
|---|---|
| 43.5K | 13 |

### Column Stats

| Name | Min | Max | Mean | Median | Null Count | Null % | Distinct Count | Distinct % | Std. Dev | Sample Values |
|---|---|---|---|---|---|---|---|---|---|---|
| age_m | 0 | 11 | 5.49 | 5 | 0 | 0.00% | 12 | 0.35% | 3.50 | 10, 1, 3 |
| age_y | 0 | 16 | 7.12 | 7 | 0 | 0.00% | 17 | 0.04% | 4.46 | 10, 2, 3 |
| breed | unknown | unknown | unknown | unknown | 0 | 0.00% | 630 | 0.15% | unknown | Russian Blue, "Segugio Maremmano", "Black Mouth Cur" |
| coat_type | unknown | unknown | unknown | unknown | 273 | 0.63% | 218 | 0.50% | unknown | Short, Short, Long |
| | | | | | | | | | | "2021-07-14 22:10:32" |

# Metadata 360

## Stats over time



# Metadata 360

## Queries

Acryl DataHub

longtailcompanions.acryl.io/dataset/urn:li:dataset:(urn:li:dataPlatform:snowflake,long_tail_companions.a...

Search Dat... | Analytics | My Requests | Domains | Users & Groups | Ingestion | Policies | Tests

Datasets › prod › snowflake › long_tail_companions › adoption › pet_profiles

Details | Lineage | 1 upstream, 2 downstream

Table | Snowflake › long_tail_companions › adoption

**pet_profiles**

Schema | Documentation | Properties | Lineage | Queries | Stats | Validation | Incidents

```
select
human_profiles.first_name
, human_profiles.last_name
, human_profiles.email
, human_profiles.age
, pet_profiles.name
, pet_profiles.species
, pet_profiles.breed
, adoptions.status
from
  humans
left join
  human_profiles
  on humans.profile_id = human_profiles.profile_id
left join
  adoptions
  on humans.pk = adoptions.human_fk
left join
  pets
  on adoptions.pet_fk = pets.pk
left join
  pet_profiles
  on pets.profile_id = pet_profiles.profile_id
```

```
select
   pet_profiles.name
   , popular_dog_names.dogname is not null as is_popular_dog_name
```

**About**

this table contains profile details of all pets

+ Add Link

**Stats** | More stats ›

Rows: 43,517

Columns: 13

Monthly Queries: 2710

Top Users: B M +25

Last Updated: 5/23/2022, 07:12 PM CDT

**Tags**

Tier1 | + Add Tag

**Glossary Terms**

Sensitive | Confidential | + Add Term

**Owners**

Adoption | Ricca MacAnespie | Prentiss Matussov | Avigdor Bramhall

# Metadata 360

## Data Quality assertions

Acryl DataHub

longtailcompanions.acryl.io/dataset/urn:li:dataset:(urn:li:dataPlatform:snowflake,long_tail_companions.adoption....

Analytics | My Requests | Domains | Users & Groups | Ingestion | Policies | Tests

Datasets > prod > snowflake > long_tail_companions > adoption > pet_profiles

Details | Lineage | 1 upstream, 2 downstream

Table | Snowflake > long_tail_companions > adoption

**pet_profiles**

Schema | Documentation | Properties | Lineage | Queries | Stats | Validation | Incidents

Assertions (8) | Tests (3)

All assertions have passed

8 successful assertions, 0 failed assertions

- Passed — Column **species** values are **not null**
- Passed — Column **age_m** values are **not null**
- Passed — Column **age_y** values are **not null**
- Passed — Column **spay_neutered** values are **not null**
- Passed — Column **sex** values are in ["F", "M"]
- Passed — Column **sex** values are **not null**

**About**

this table contains profile details of all pe...

+ Add Link

**Stats** More stats >

Rows
43,517

Columns
13

Monthly Queries
2710

Top Users
B M +26

Last Updated
5/23/2022, 07:12 PM CDT

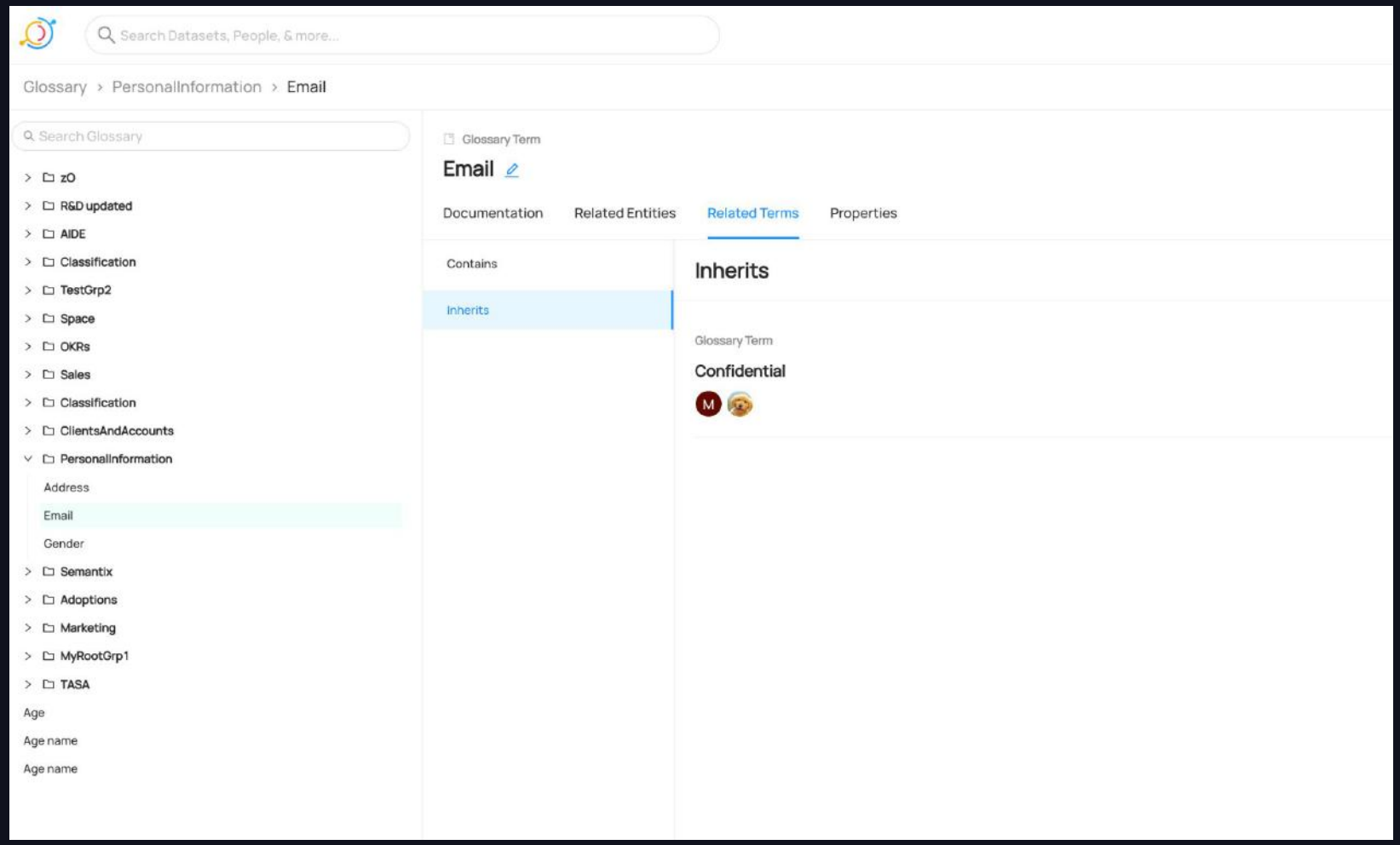
**Tags**

Tier1 X + Add Tag

**Glossary Terms**

# Metadata 360

## Business Context using Business Glossary



# Metadata 360

## Relate assets to KPIs

Acryl DataHub

longtailcompanions.acryl.io/glossary/urn:li:glossaryTerm:Adoption.ReturnRate/Documentation?is_lineage_mode=false

Search... | Analytics | My Requests | Domains | Users & Groups | Ingestion | Policies | Tests

Glossary Terms › Adoption › ReturnRate

Details | Lineage | 0 upstream, 0 downstream

Glossary Term

**ReturnRate**

Related Entities | Documentation | Related Terms | Properties

Edit | + Add Link

The percentage of adopted animals that were returned to the shelter within 60 days of adoption

```
with adoption_date as (
select
  profile_id
  , as_of_date as adoption_date
from
  analytics.pet_status_history
where
  status = 'adopted'
)


, return_date as (
select
  profile_id
  , as_of_date as return_date
from
  analytics.pet_status_history
where
  status = 'returned to facility'
)
```

**About**

The percentage of adopted animals that were returned to the ...

+ Add Link

**Owners**

Luigi Bonsale | Nonie Radband | Melina Eliez | Analytics | Shannon Lovett

+ Add Owner

# Metadata 360

Combine technical and business metadata

Who will provide this accurately?

What is the authoritative **pet** dataset?

**BUSINESS METADATA**

Social Tags, Glossary Terms, Domains, Data Product Definition, Ownership, KPIs, …

**TECHNICAL METADATA**

Tables, Schemas, Comments and Descriptions, Table and Column-level statistics, Query history, Usage statistics, Owners, Lineage, …

# Shift Left

Declare & collect metadata at the source

schema.yml

```yaml
version: 2

models:
  - name: pet_details
    description: Table with all pet-related details
    meta:
      owner: "shannon@longtail.com"
      model_maturity: prod
      contains_pii: false
      business_critical: true
      domain: "Pet Adoptions"
```

# Shift Left

Acryl DataHub

longtailcompanions.acryl.io/dataset/urn:li:dataset:(urn:li:dataPlatform...

Analytics | My Requests | Domains | Users & Groups | Ingestion | Policies | Tests

Datasets › prod › dbt › long_tail_companions › analytics › pet_details

Details | Lineage | 3 upstream, 1 downstream

Table | Dbt

pet_details

Schema | View Definition | Documentation | Properties | Lineage | Queries

Edit | Add Link

Table with all pet-related details

True | SQL

**Tags**

business_critical | prod_model

Add Tag

**Glossary Terms**

DaysInStatus

Add Term

**Owners**

Shannon Lovett

Add Owner

**Domain**

Pet Adoptions

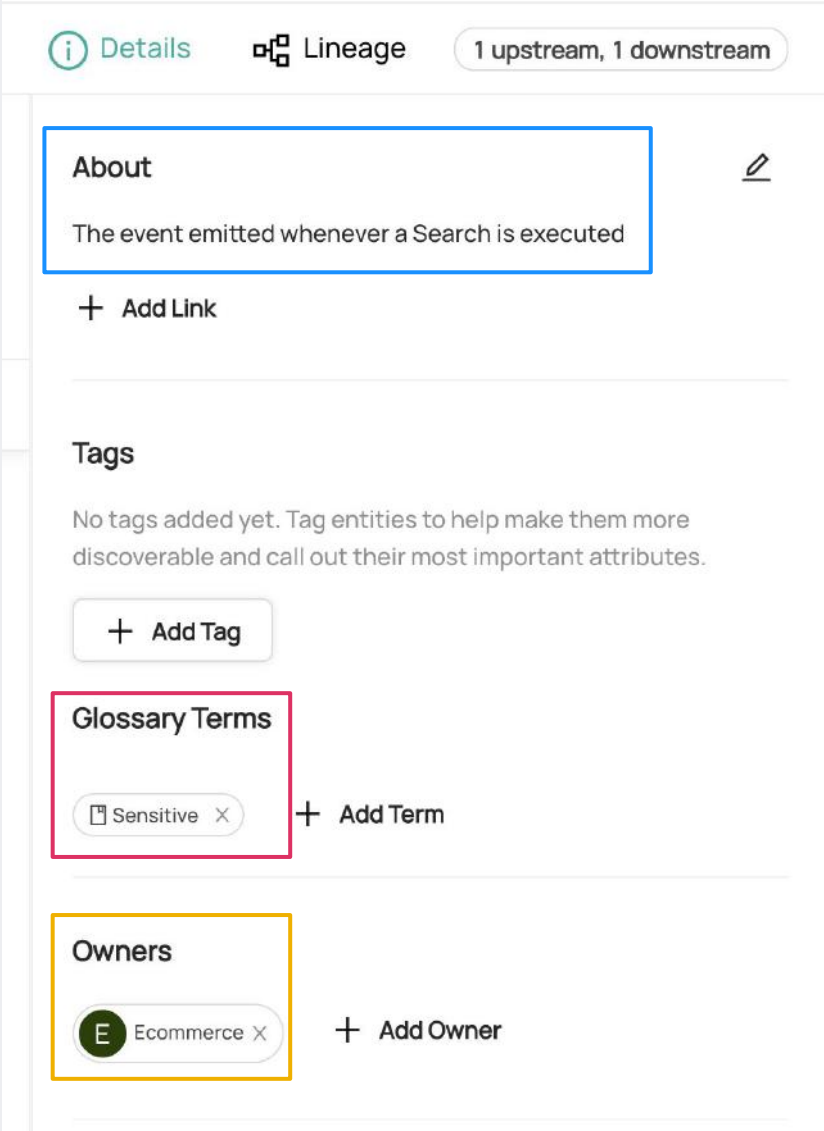
schema.yml

```yaml
version: 2

models:
  - name: pet_details
    description: Table with all pet-related details
    meta:
      owner: "shannon@longtail.com"
      model_maturity: prod
      contains_pii: false
      business_critical: true
      domain: "Pet Adoptions"
```

# Shift Left

## Declare & collect metadata at the source

```
syntax = "proto3";
package ecommerce;
import "protobuf/meta/meta.proto";
import "common/context.proto";


message SearchResult {
  int32 item_id=1;
}

/**
  The event emitted whenever a Search is executed
**/
message SearchEvent {
  option(meta.msg.classification) = "Classification.Sensitive";
  option(meta.msg.team) = "Ecommerce";

  common.EventContext context = 4;
  // the search identifier
  int32 search_id = 1;

  repeated SearchResult result_array = 3;
}
```



# Shift Left

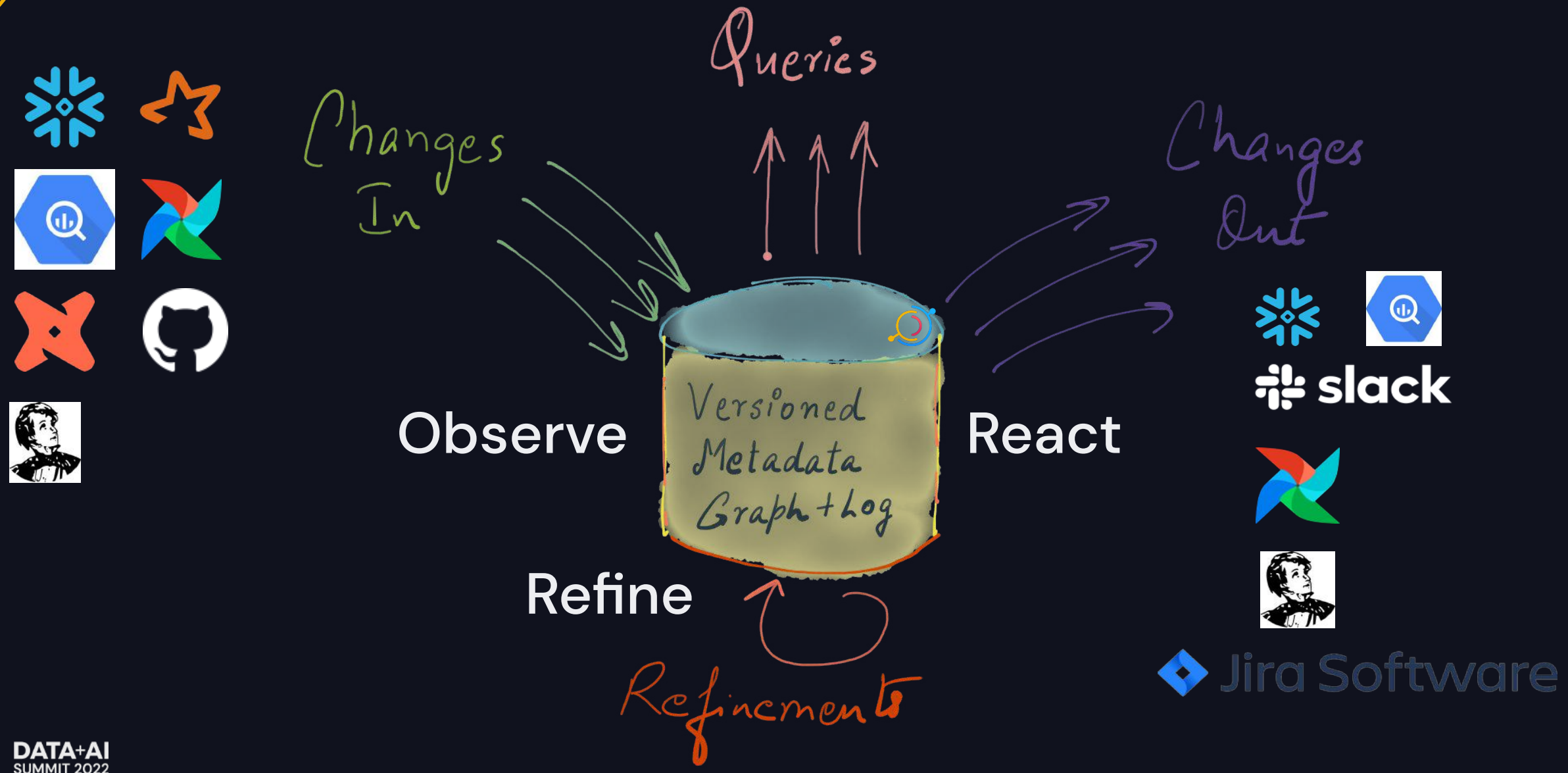
Declare metadata at source => high quality metadata

What is the authoritative **pet_profile** dataset?

# Active Metadata

Inject metadata into the operational plane

# DataHub's Architecture



# Active Metadata

DataHub Actions let you react to changes in real-time

```yaml
# hello_world.yaml
name: "hello_world"
source:
  type: "kafka"
  config:
    connection:
      bootstrap: ${KAFKA_BOOTSTRAP_SERVER:-localhost:9092}
      schema_registry_url: ${SCHEMA_REGISTRY_URL:-http://localhost:8081}
filter:
  event_type: "EntityChangeEvent_v1"
  event:
    category: "TAG"
    operation: "ADD"
    modifier: "urn:li:tag:pii"
action:
  type: "hello_world"
```

Copy

```
datahub actions -c hello_world.yaml
```

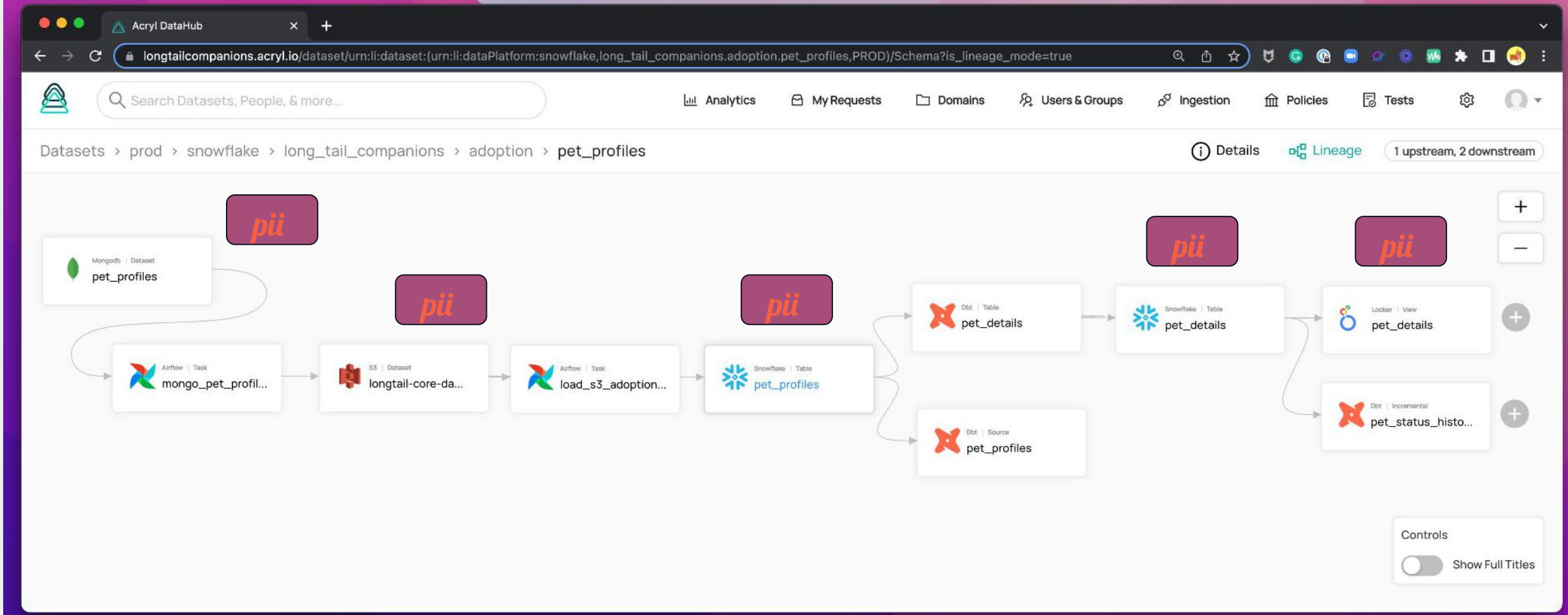
# Active Metadata

## DataHub Actions let you react to changes in real-time

```python

# A basic example of a DataHub action that prints all
# events received to the console.
class HelloWorldAction(Action):
    @classmethod
    def create(cls, config_dict: dict, ctx: PipelineContext) -> "Action":
        action_config = HelloWorldConfig.parse_obj(config_dict or {})
        return cls(action_config, ctx)

    def __init__(self, config: HelloWorldConfig, ctx: PipelineContext):
        self.config = config

    def act(self, event: EventEnvelope) -> None:
        print("Hello world! Received event:")
        message = json.dumps(json.loads(event.as_json()), indent=4)
        if self.config.to_upper:
            print(message.upper())
        else:
            print(message)

    def close(self) -> None:
        pass
```

# Active Metadata

## Tag Propagation using real-time actions

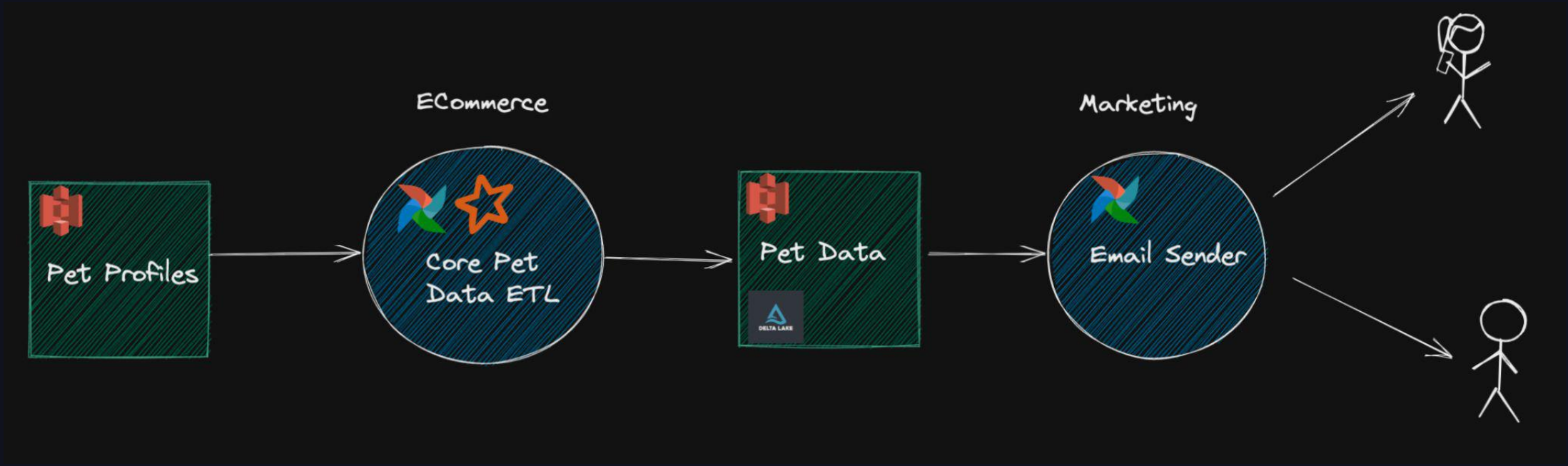


# Active Metadata

Prevent bad data
from causing more damage

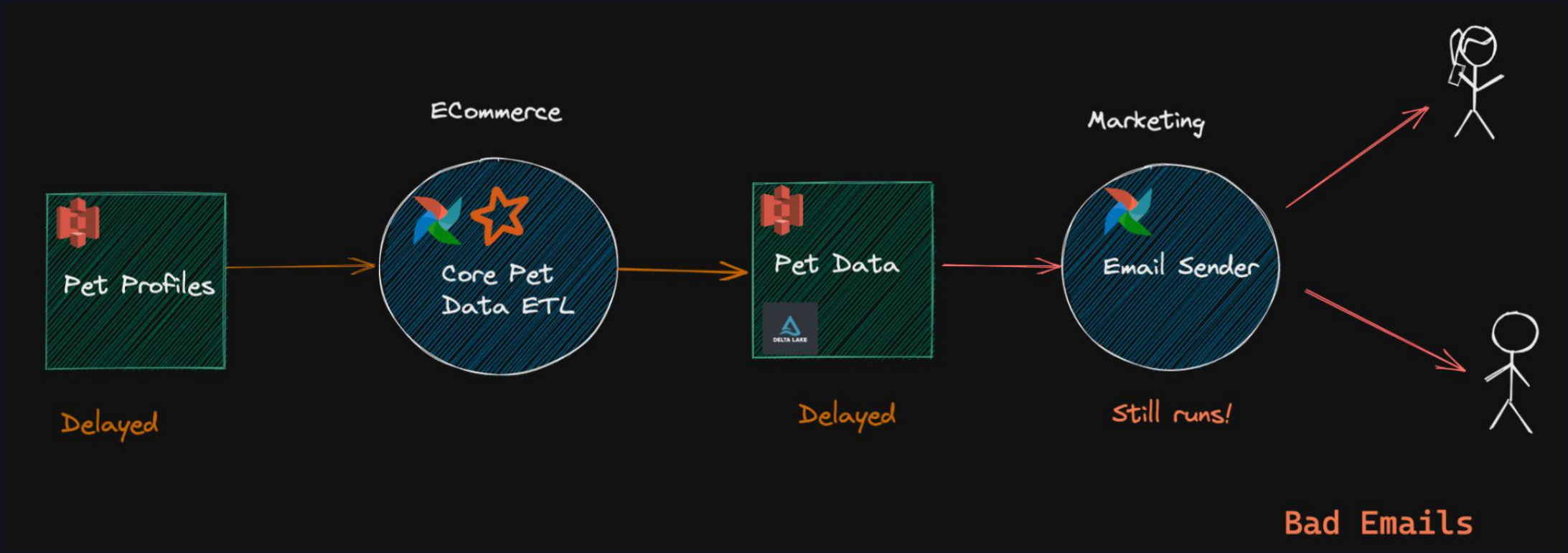
# A Typical Scenario

## Data Standardization followed by Marketing Emails



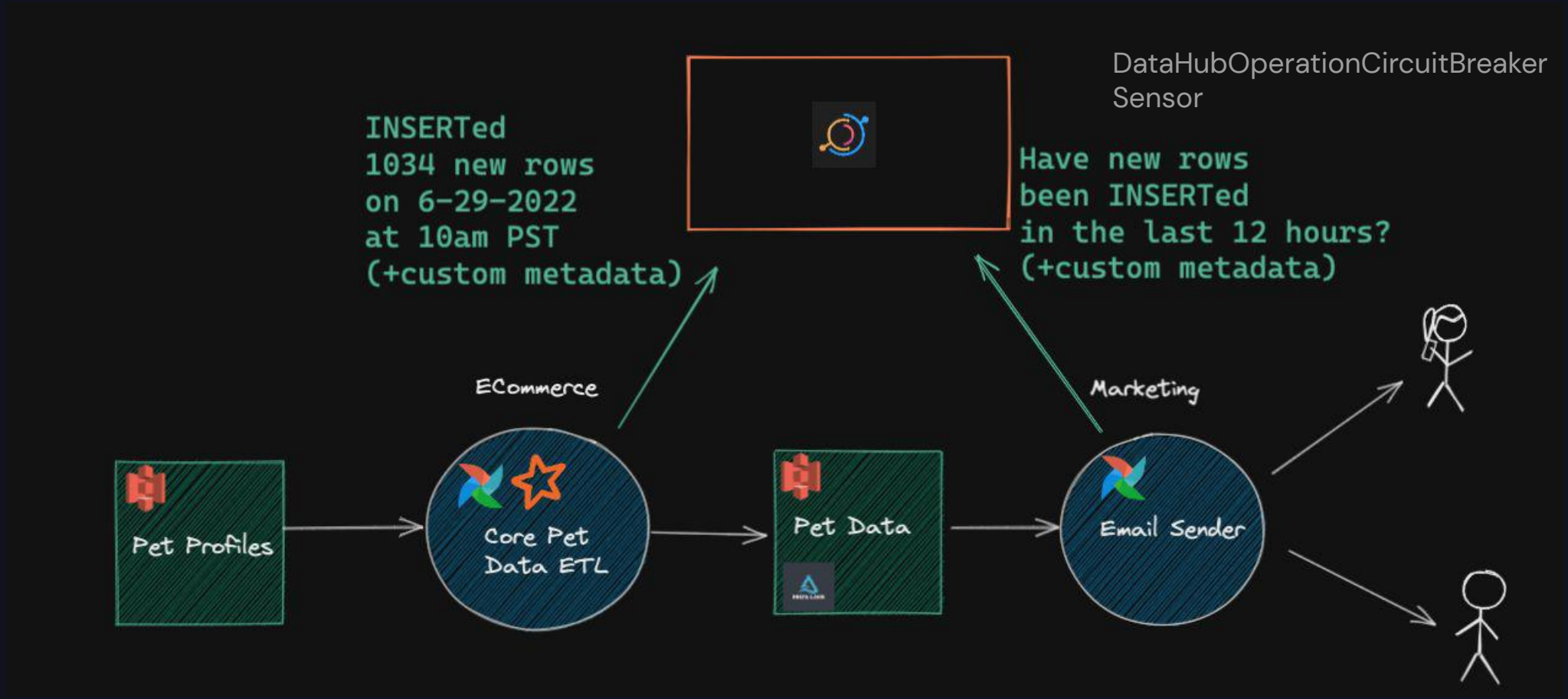
# A Problem: Delayed Data

One day…



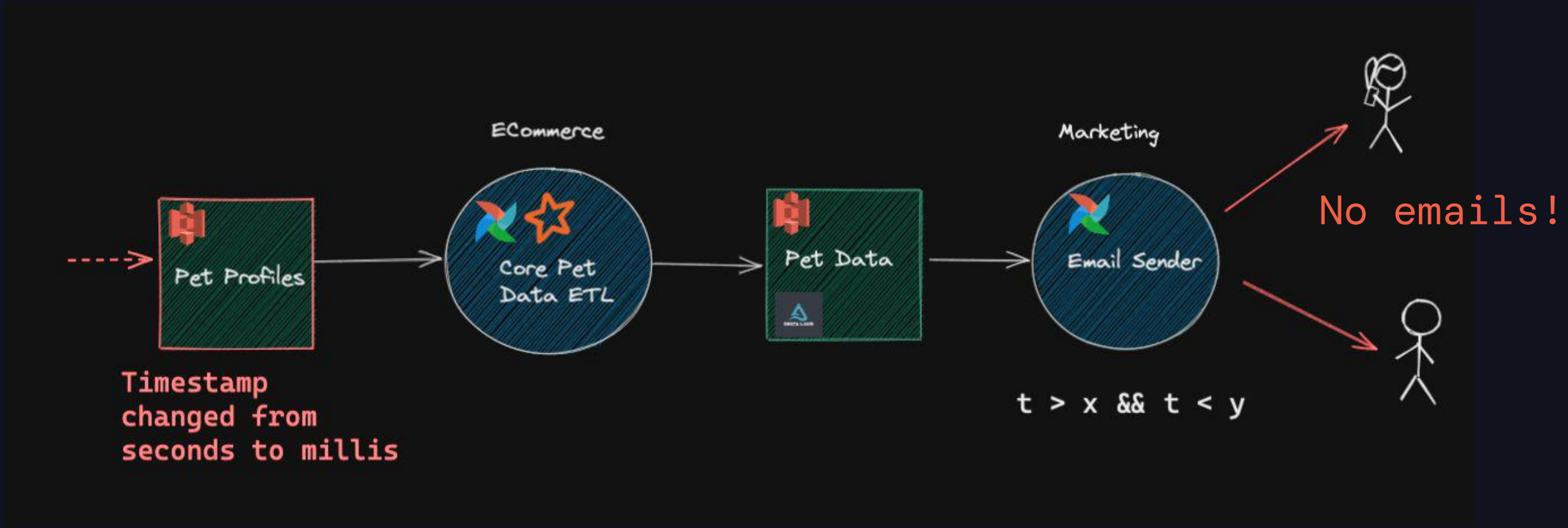
# A Problem: Delayed Data

## Step 2: Verify



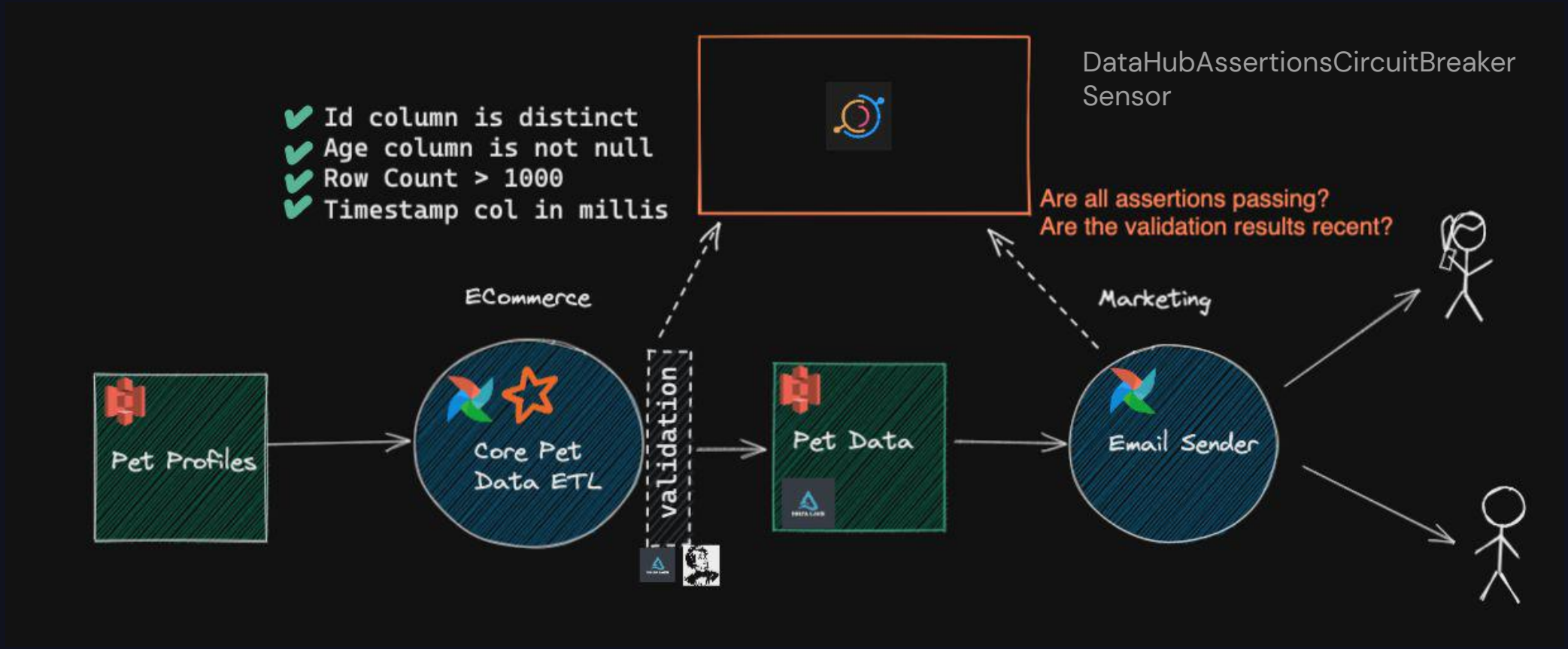
# Another Problem: Broken Data

A few months later…



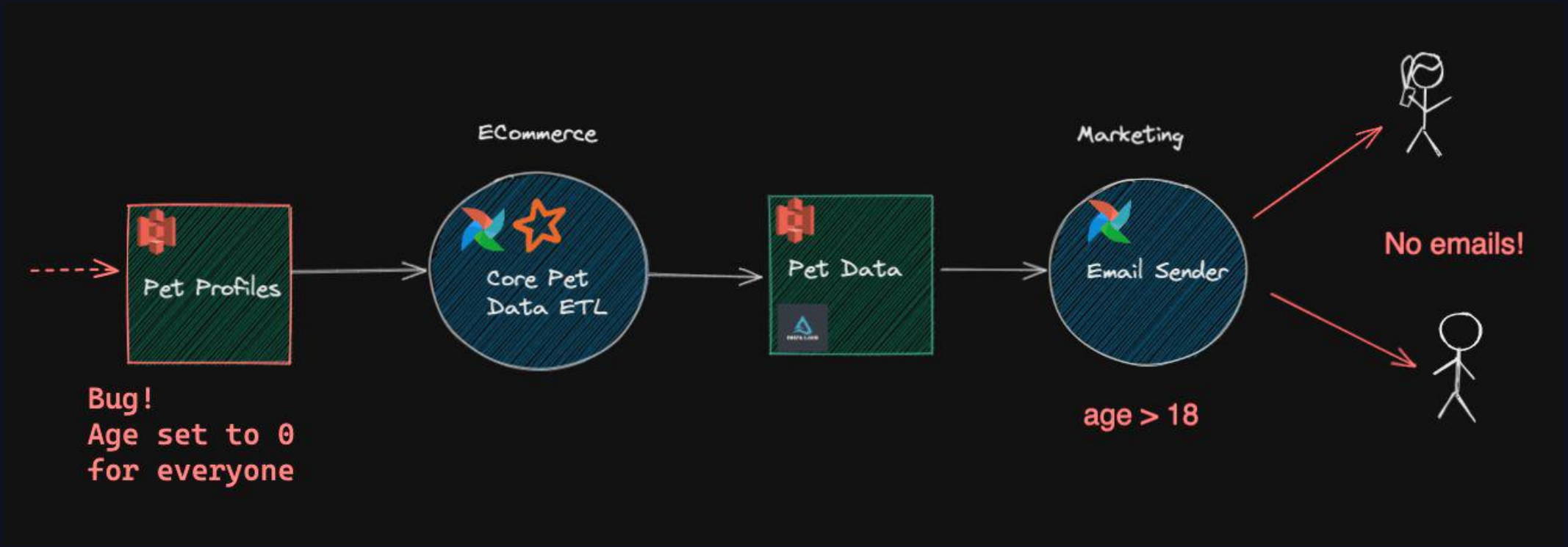
# DataHub Assertions

## Step 2: Verify



# Another Problem: Broken Data

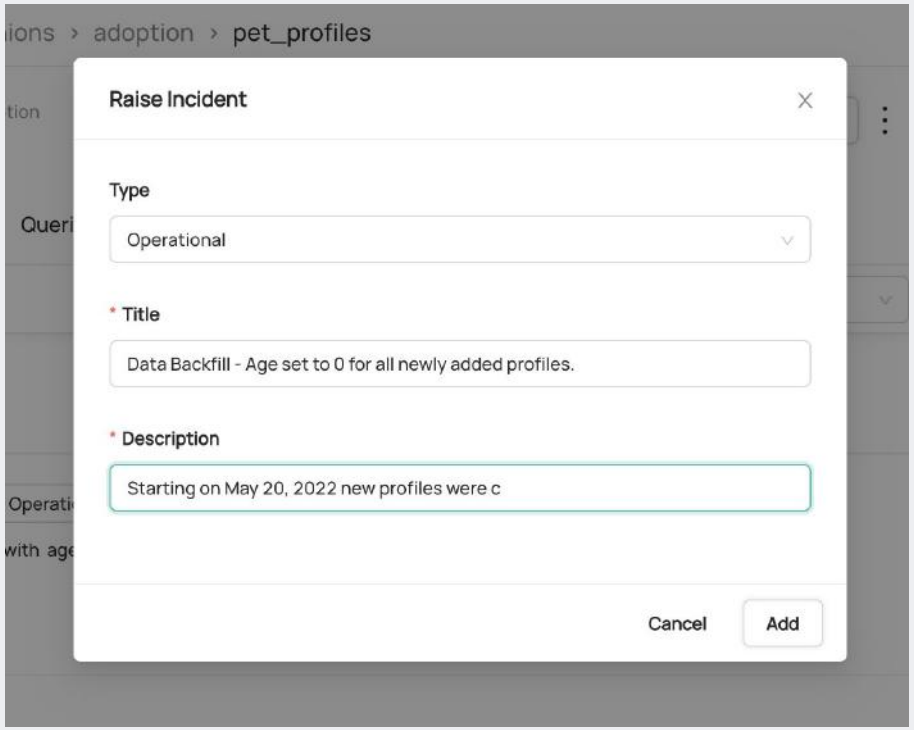
A few weeks later…

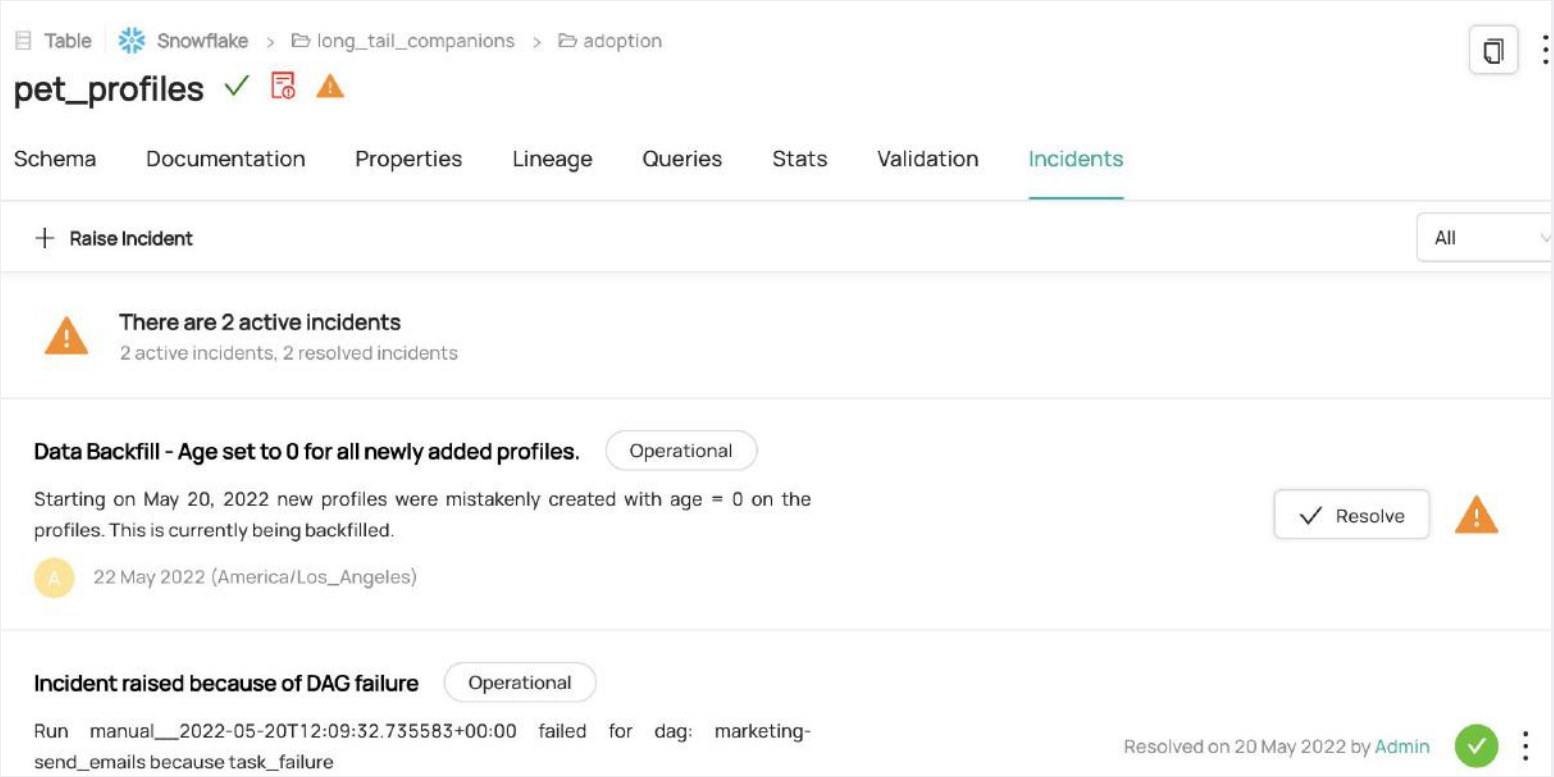



*Tests can't catch everything*

# DataHub Incidents

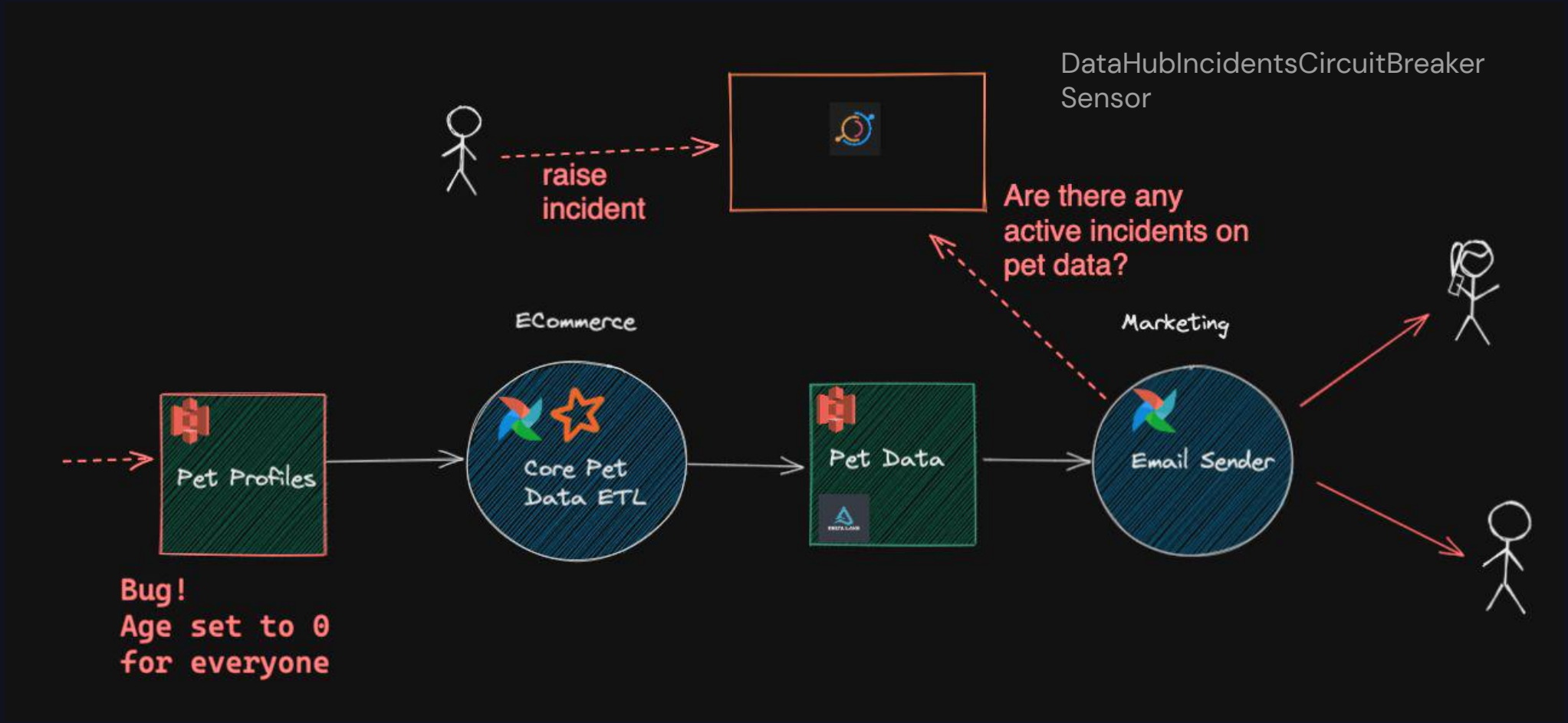
## Step 1: Raise Incident

# DataHub Incidents

## Step 2: Verify



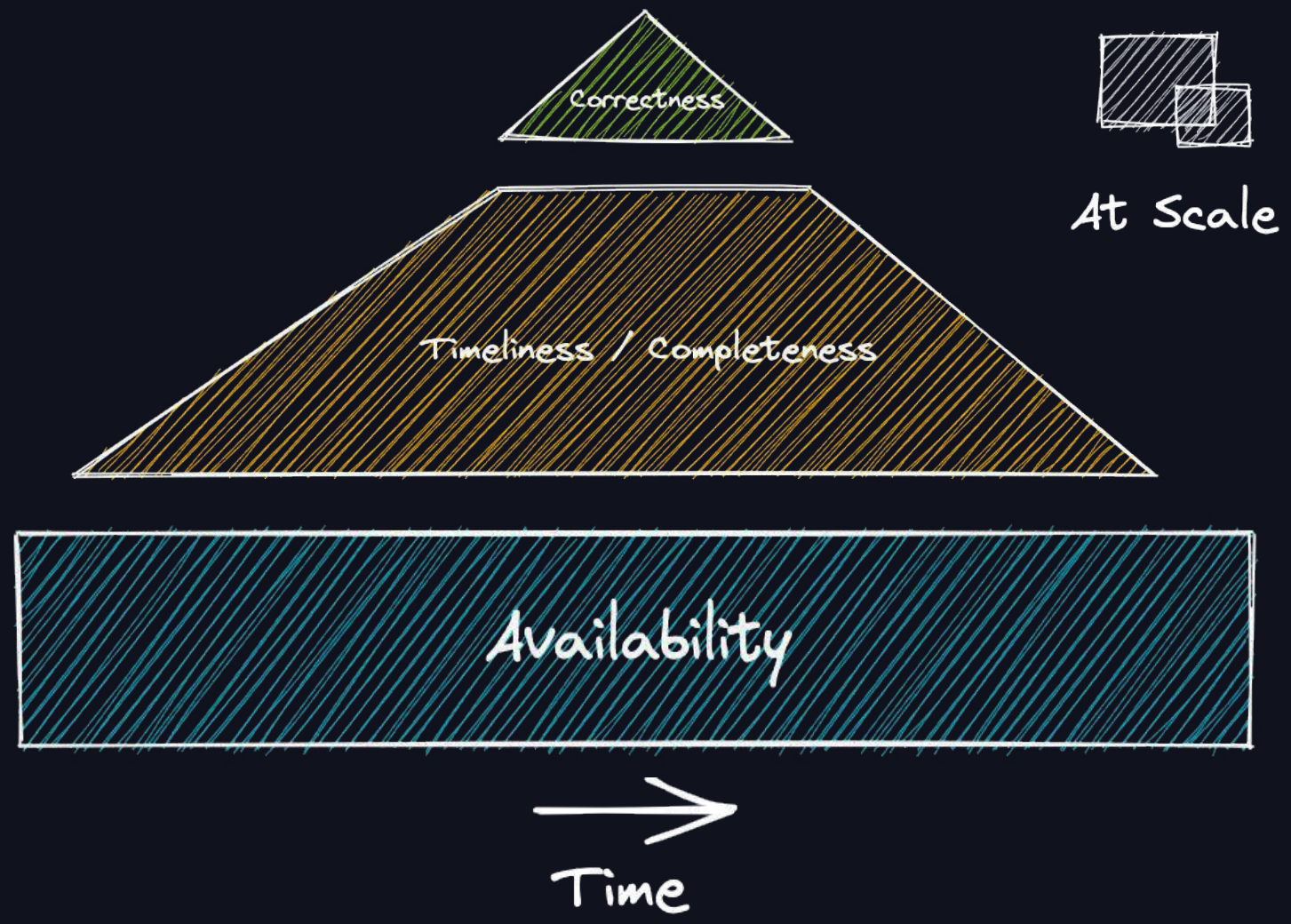
# The Reliability Hierarchy of Needs

## Toolkit

DataHub Incidents

DataHub Assertions

DataHub Operations

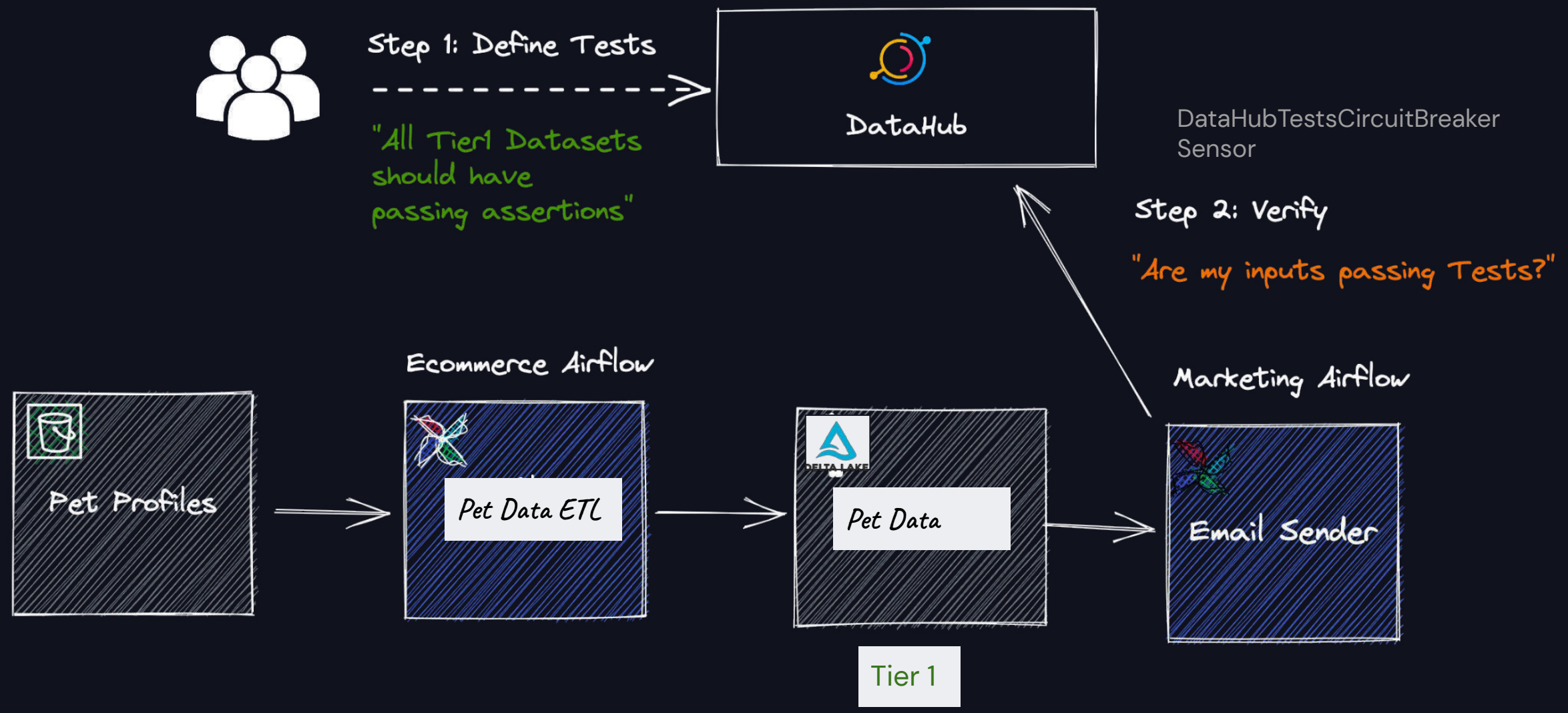





At Scale

? How to handle 100s of DAGs? 1000s of Datasets?

# DataHub Tests



# DataHub Tests

## Central policy definition, distributed enforcement

* Name

Give your test a name.

All Tier 1 Datasets must have passing Assertions

* Category

The category of your test.

Governance

Description

An optional description to help keep track of your test.

All Tier 1 Datasets MUST have Assertions defined and passing.

Define your Test

For more information about how to configure a Test, check out the DataHub Tests Guide.

```
on:
    dataset:
        -
            field: tags
            condition: EQUALS
            values:
                - 'urn:li:tag:Tier1'
rules:
    -
        field: assertions
        conditon: EXISTS
    -
        field: assertions.runEvents.result.type
        conditon: NOT_EQUALS
        value: FAILURE
```

Manage Tests

DataHub Tests allows you to continuously evaluate a set of conditions on the assets comprising your Metadata Graph.

+ Create new test

| Name | Category | Description | Results |
|---|---|---|---|
| All Tier 1 Datasets must have passing Assertions | Governance | All Tier 1 Datasets MUST have Assertions defined and passing. | 1 passing, 0 failing |
| All Datasets must have > 0 Glossary Terms | Governance | For this test, all Datasets must have a Glossary Term assigned to them. | 1 passing, 0 failing |
| All Datasets must have Domain set | Governance | All datasets must have a domain set. | 0 passing, 1 failing |
| All Datasets on Snowflake must have > 2 Owners | Governance | Each Dataset on Snowflake MUST have > 2 Owners assigned to it. | 0 passing, 0 failing |

Some tests are failing

2 passing tests, 1 failing tests

Test Results

Failing — All Datasets must have Domain set
Governance | All datasets must have a domain set.

Passing — All Datasets must have > 0 Glossary Terms
Governance | For this test, all Datasets must have a Glossary Term assigned to them.

Passing — All Tier 1 Datasets must have passing Assertions
Governance | All Tier 1 Datasets MUST have Assertions defined and passing.

# DataHub Tests Circuit Breaker

## Step 1: Define Task policy in *airflow_local_settings.py*

```python
def metadata_test_pre_execute(context) -> None:
    hook: DatahubRestHook = DatahubRestHook("datahub_longtail")
    host, password, timeout_sec = hook._get_config()

    config: MetadataTestCircuitBreakerConfig = MetadataTestCircuitBreakerConfig(
        datahub_host=host,
        datahub_token=password,
        timeout=timeout_sec,
    )
    cb = MetadataTestCircuitBreaker(config)
    print(f"context: {context}")
    ti = context["ti"]
    inlets = get_inlets_from_task(ti.task, context)
    for inlet in inlets:
        print(f"Urn: {inlet.urn}")
        if cb.is_circuit_breaker_active(inlet.urn):
            print(f"Circuit Breaker is active for {inlet.urn}")
            raise Exception(f"Metadata Test Circuit Breaker is active for {inlet.urn}")
        else:
            print(f"Metadata Test Circuit breaker is closed for {inlet.urn}")
    return

def task_policy(task: BaseOperator):
    print("Applying task policy")
    task.pre_execute = metadata_test_pre_execute
```

Set up Datahub Connection

Create a Metadata Test Circuit Breaker

Check if all the metadata tests pass for all the inlets of the task

Define a task policy which get applied to every task in every dag

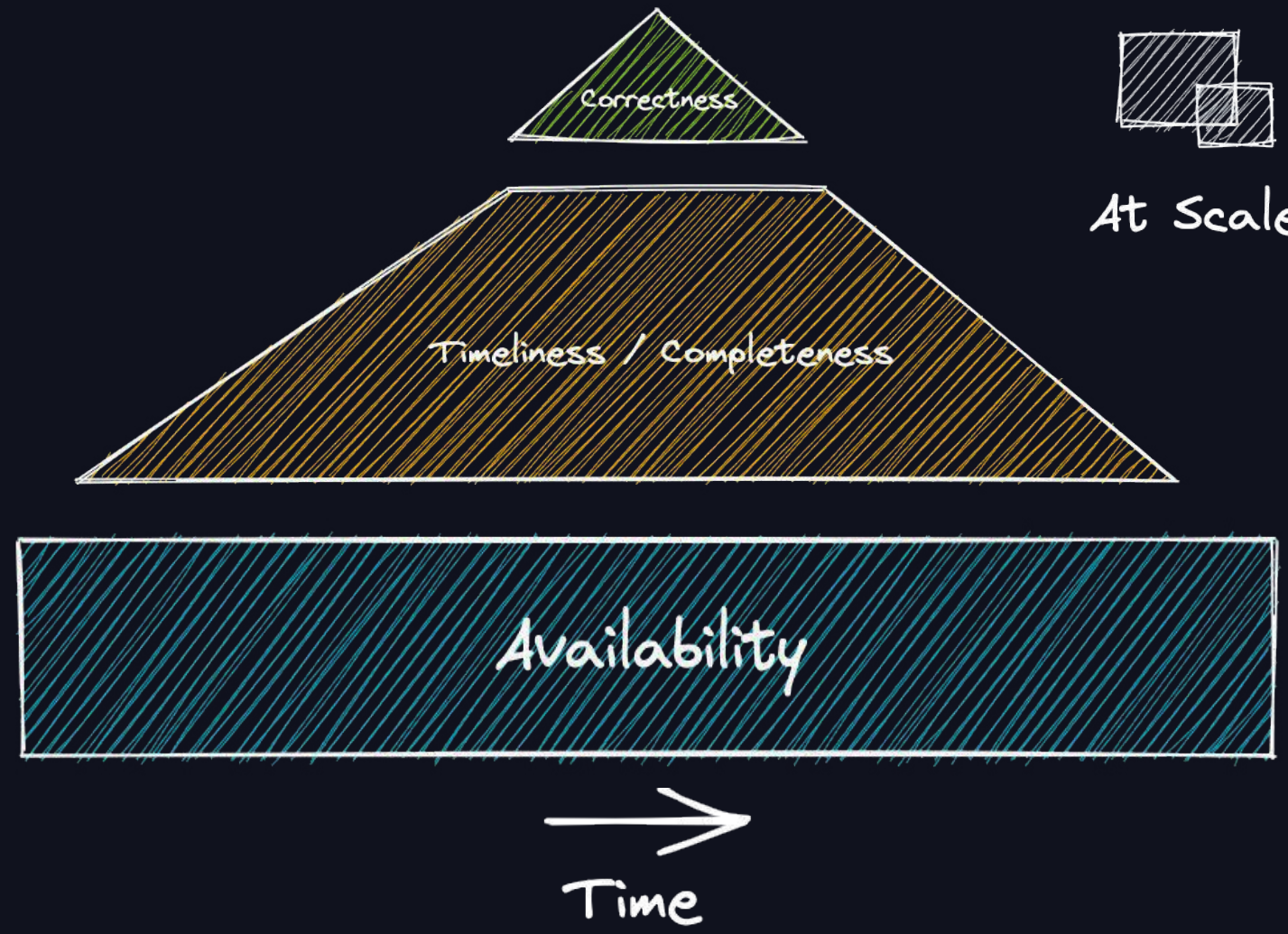
# Realizing Reliability

## Preventative Metadata : The Data Reliability Toolkit

DataHub Incidents

DataHub Assertions

DataHub Operations





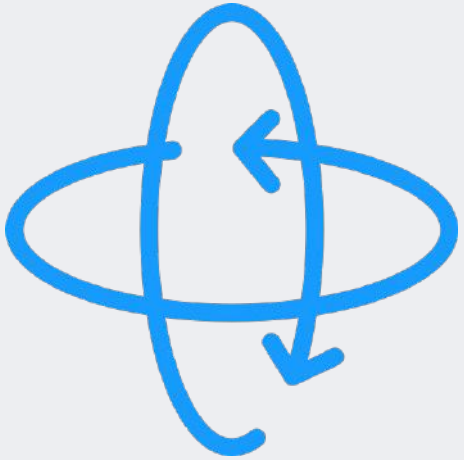

DataHub Tests

At Scale

# Active Metadata

Inject metadata into the operational plane


# 3 Must-Haves for Connecting the Dots



## Metadata 360

Combine *technical* and *business* metadata

## Shift Left

Declare & collect metadata at the source

## Active Metadata

Inject metadata into the operational plane

slack.datahubproject.io

# Now you're ready to connect the dots!

```
❯ pip install acryl-datahub
❯ datahub docker quickstart
```

https://www.acryldata.io/sign-up

DATA+AI
SUMMIT 2022

# Thank you

**Shirshanka Das**
CEO and Co-Founder, Acryl Data