

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

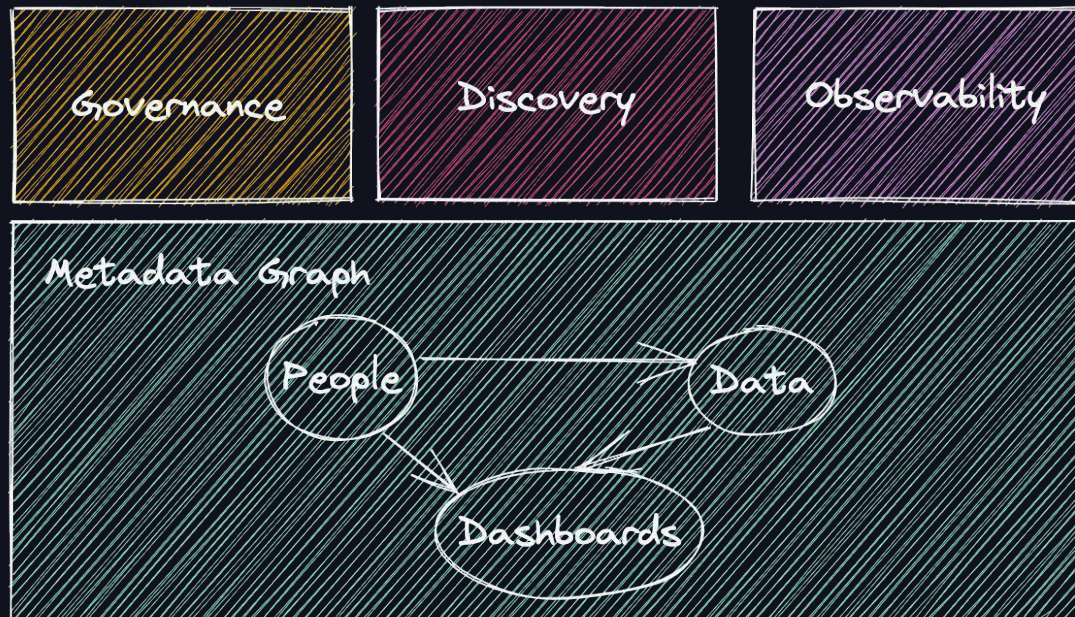


Prezi



What is DataHub?

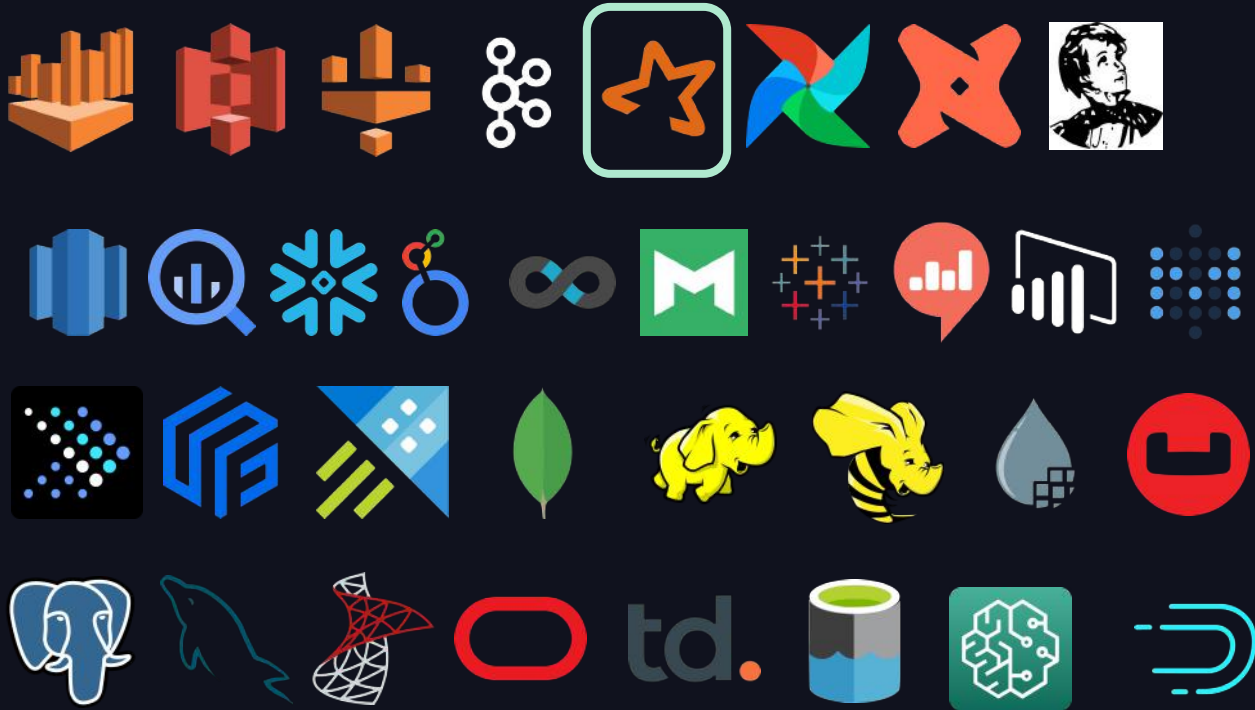
DataHub is an open source metadata platform that enables Data Discovery, Data Observability, and Federated Governance on top of a high-fidelity Metadata Graph.



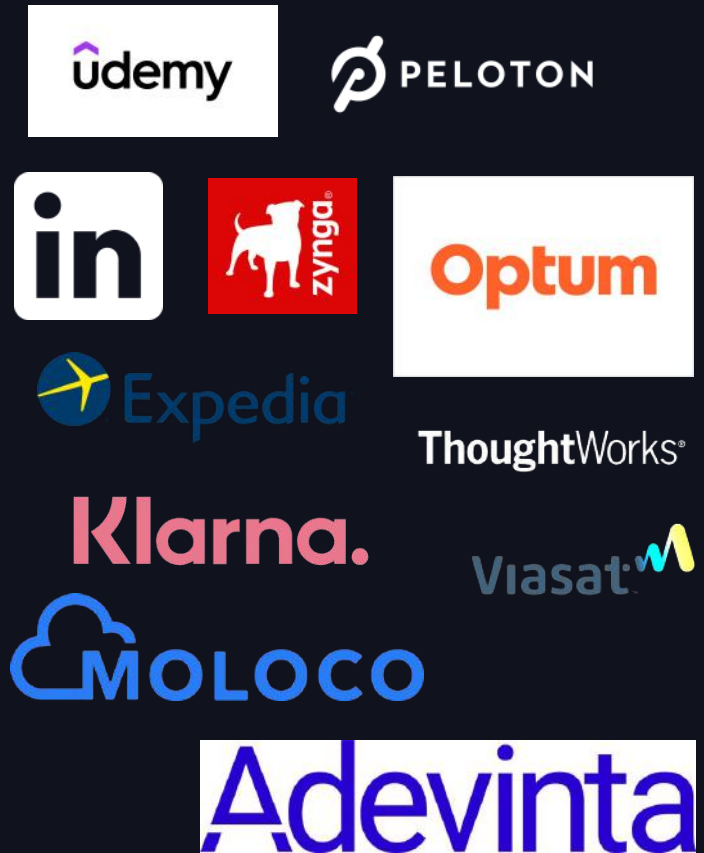
Learn more datahubproject.io

DataHub: #1 Open Source Metadata Platform

Integrations



Adopters



DataHub Community



slack.datahubproject.io



3,400+ Slack Members

10x YoY Growth

Across 56 Countries & 27 Local Time Zones

Top Member Roles



- Data Engineer
- Software Engineer
- System Architect
- Data Team Lead
- Eng Manager
- Product Manager
- Data Scientist

Top Member Industries



Software



Ecommerce



Info Tech



FinTech

DataHub Community



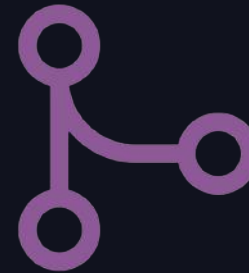
950+
Weekly Active Slack
Members



240+
OSS Contributors



10.5k+
Monthly Slack
Messages



200+
Monthly Commits



130+
Monthly Town Hall
Attendees



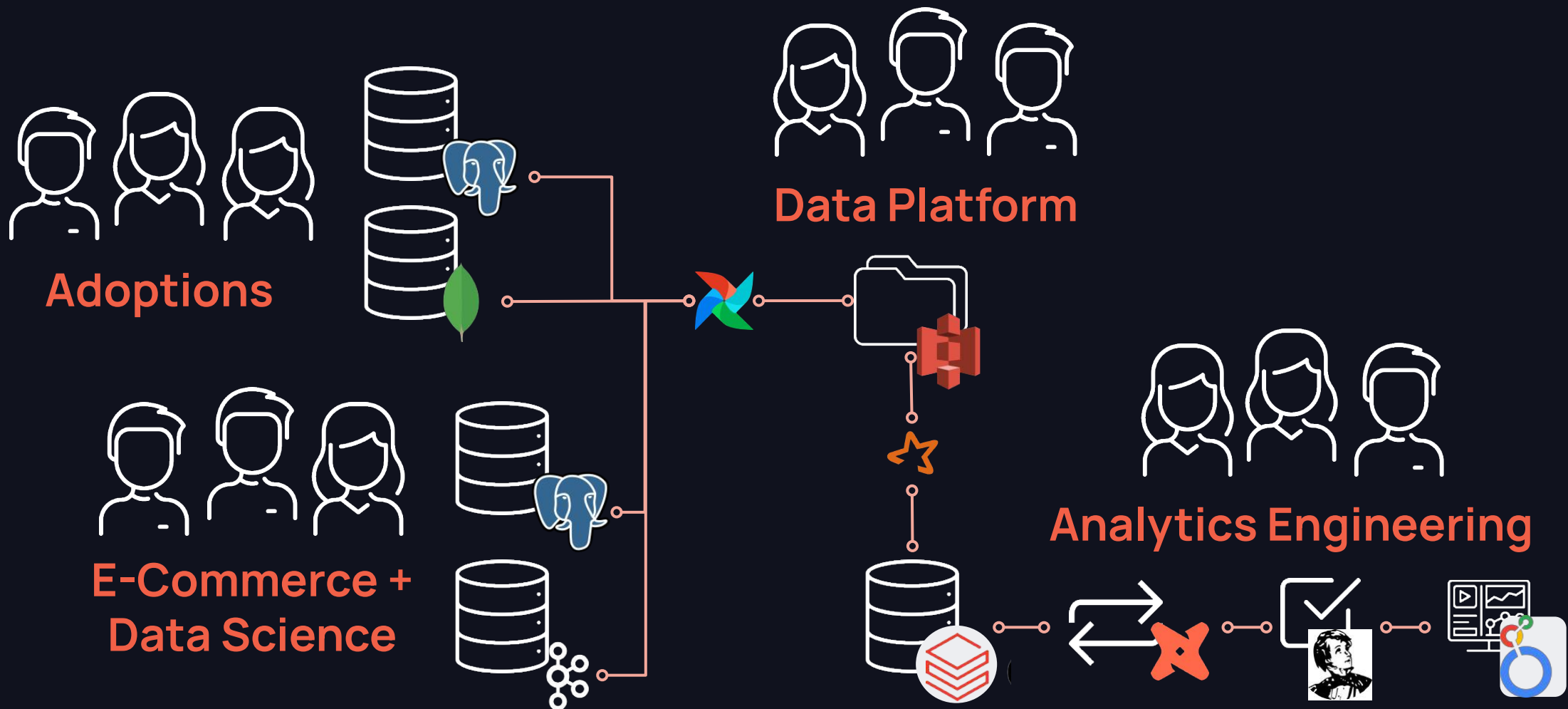
5.6k+
GitHub Stars

What can you do
with DataHub?

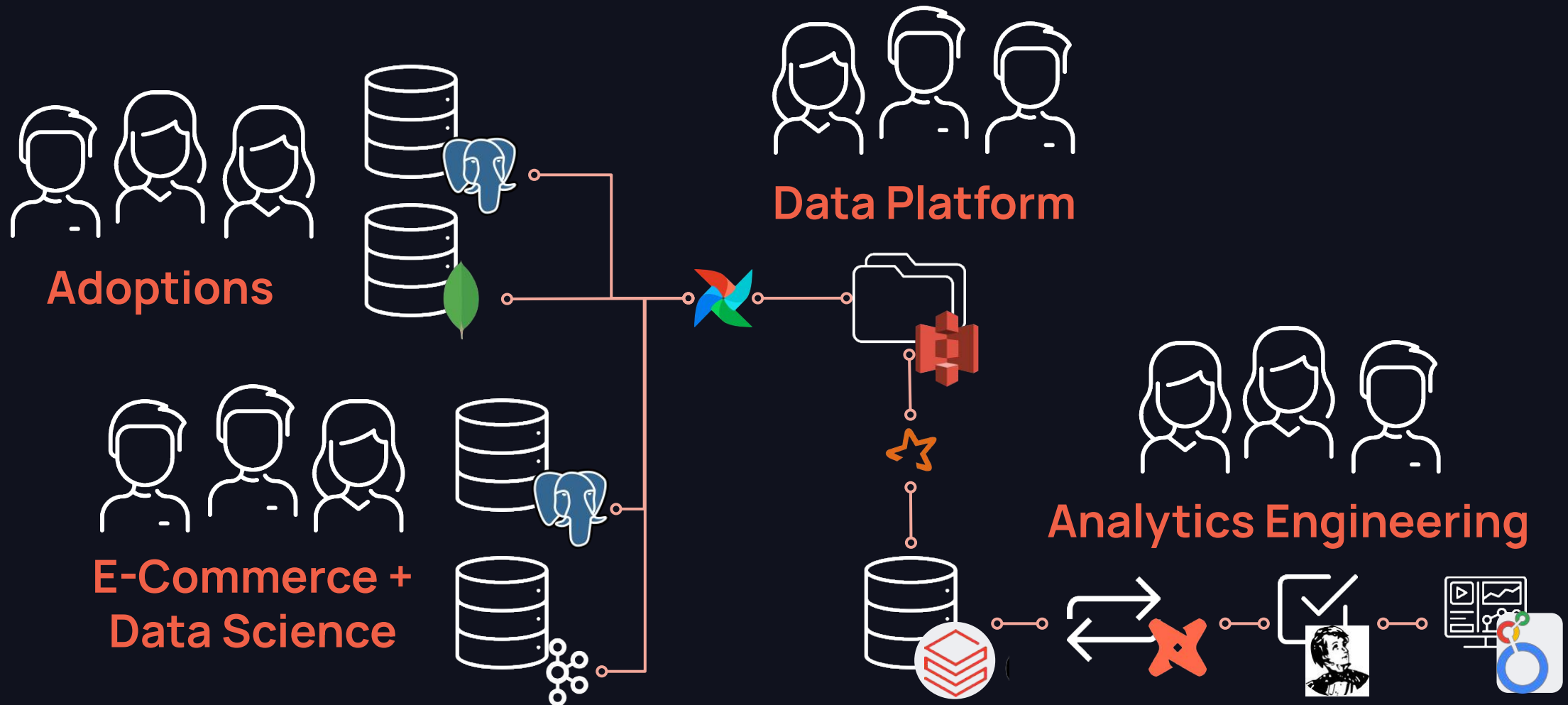
Example: Long Tail Companions

Where Every Pet is Exceptional

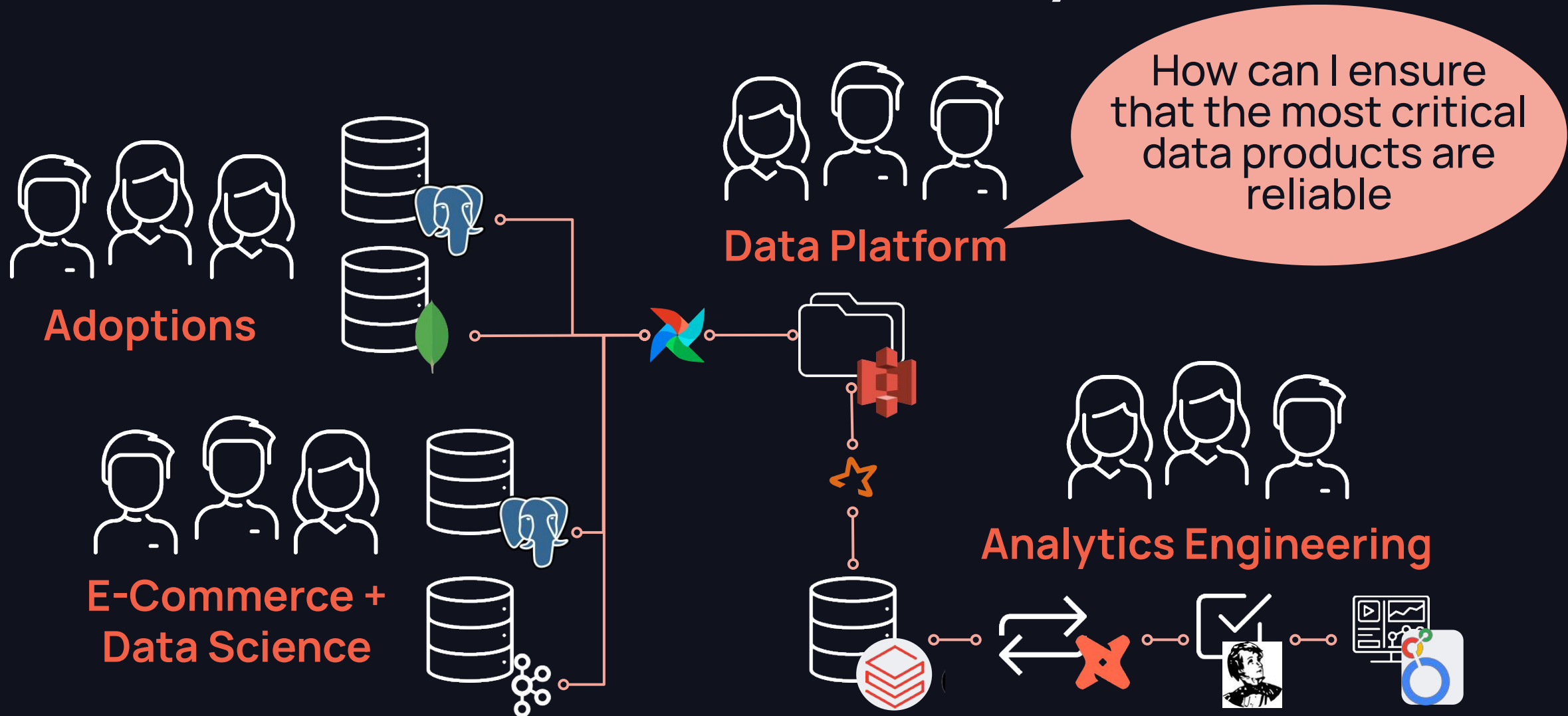
Long Tail Companions' Fragmented Data Stack



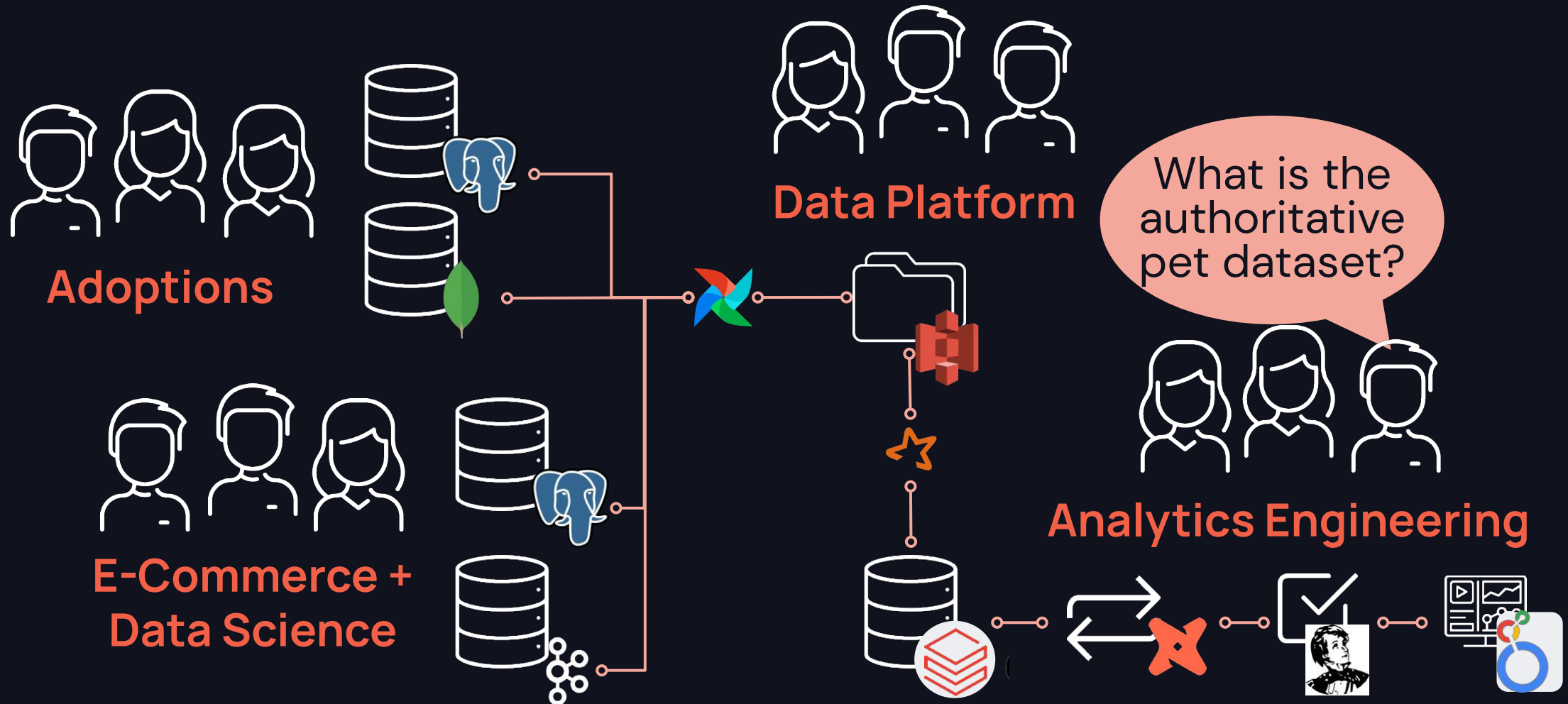
LTC Data Practitioners Routinely Ask:



LTC Data Practitioners Routinely Ask:



LTC Data Practitioners Routinely Ask:



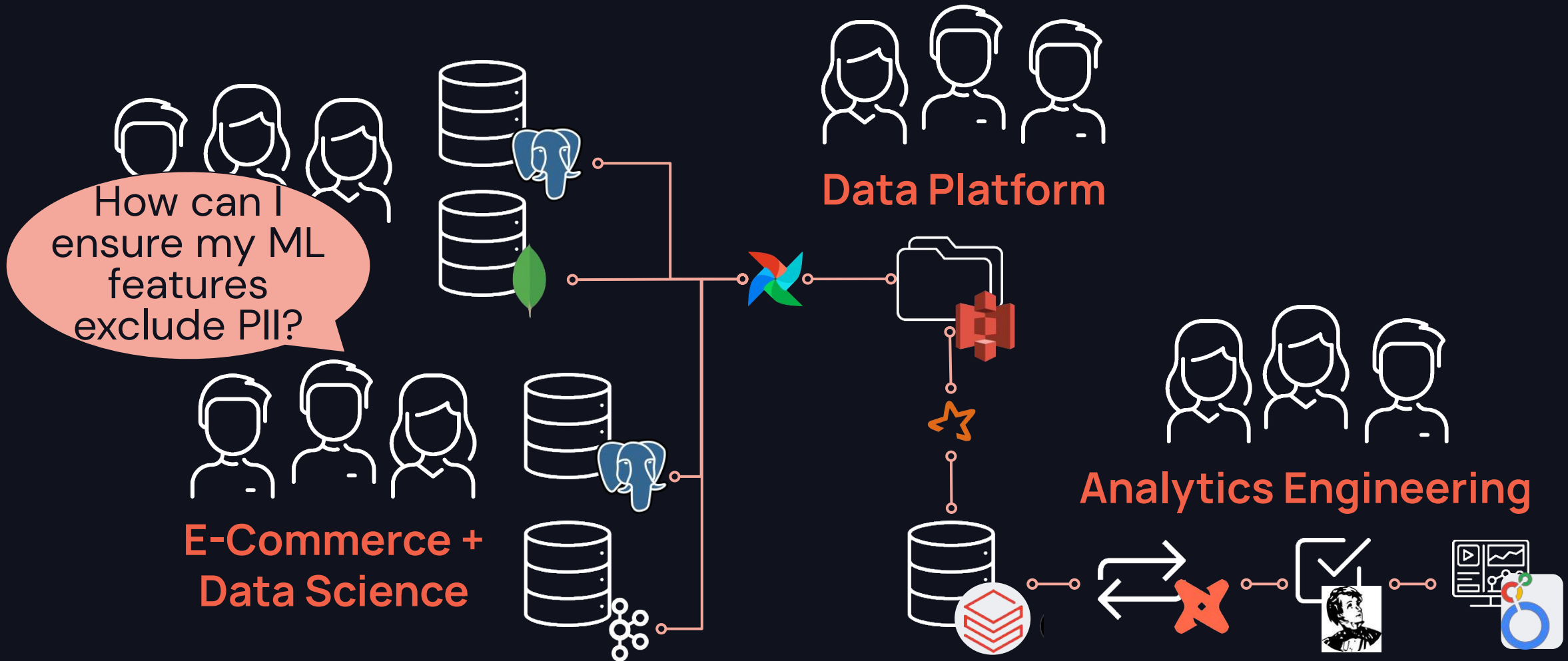
LTC Data Practitioners Routinely Ask:

How can I ensure my ML features exclude PII?

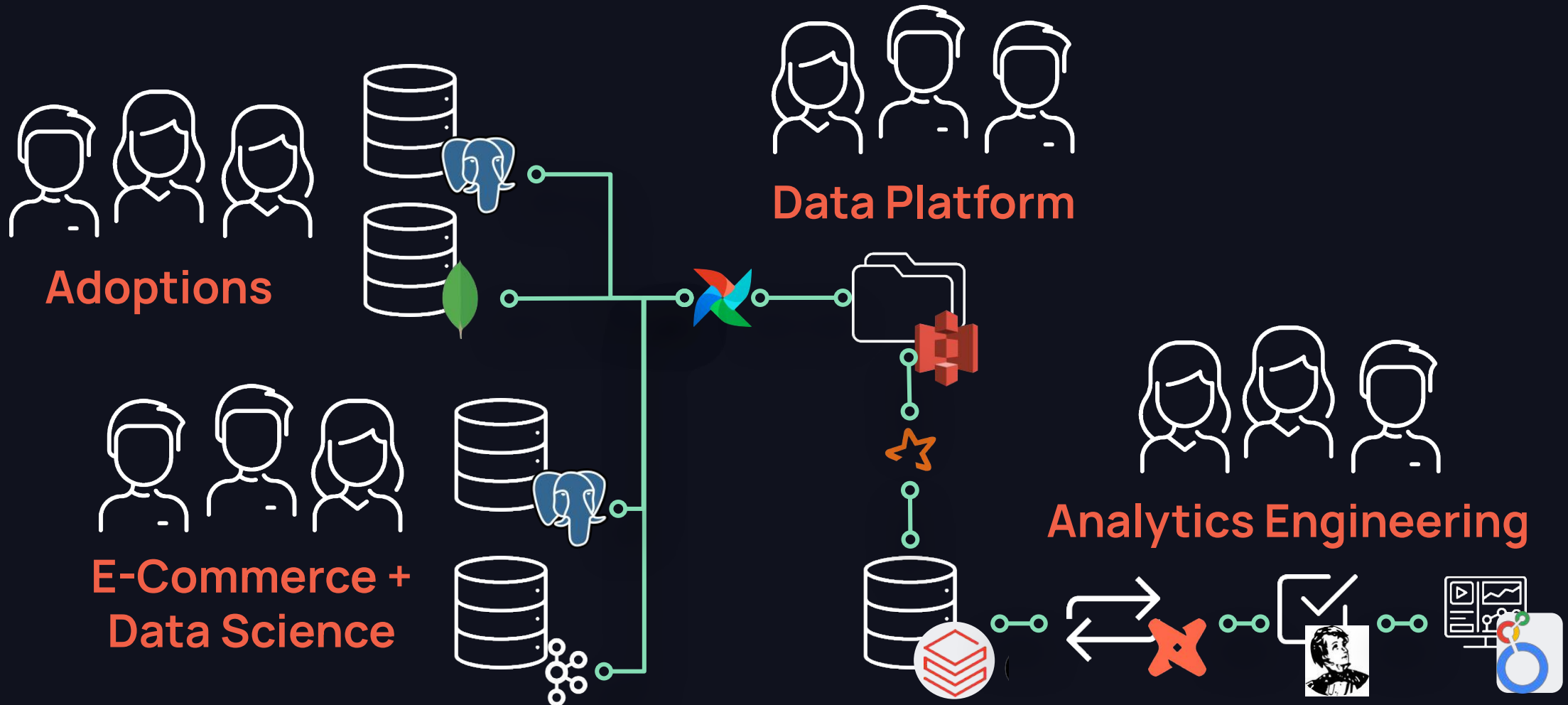
E-Commerce + Data Science

Data Platform

Analytics Engineering



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

The screenshot shows the DataHub web interface. On the left is a navigation sidebar with a search icon and a list of integration categories: ClickHouse, CSV, Data lake files, dbt, Delta Lake, Druid, Elastic Search, Feast, File Based Lineage, File, Glue, SAP HANA, and Hive (which is highlighted in blue). The main content area displays a configuration snippet for a DataBricks Hive source. The configuration is as follows:

```
# -----  
# 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/xxxxxyzzzaasa/1234-567890-hello123'  
  
sink:  
  # sink configs
```

Metadata 360

Integrations: Delta Lake (new!)

DataHub

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.

```
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

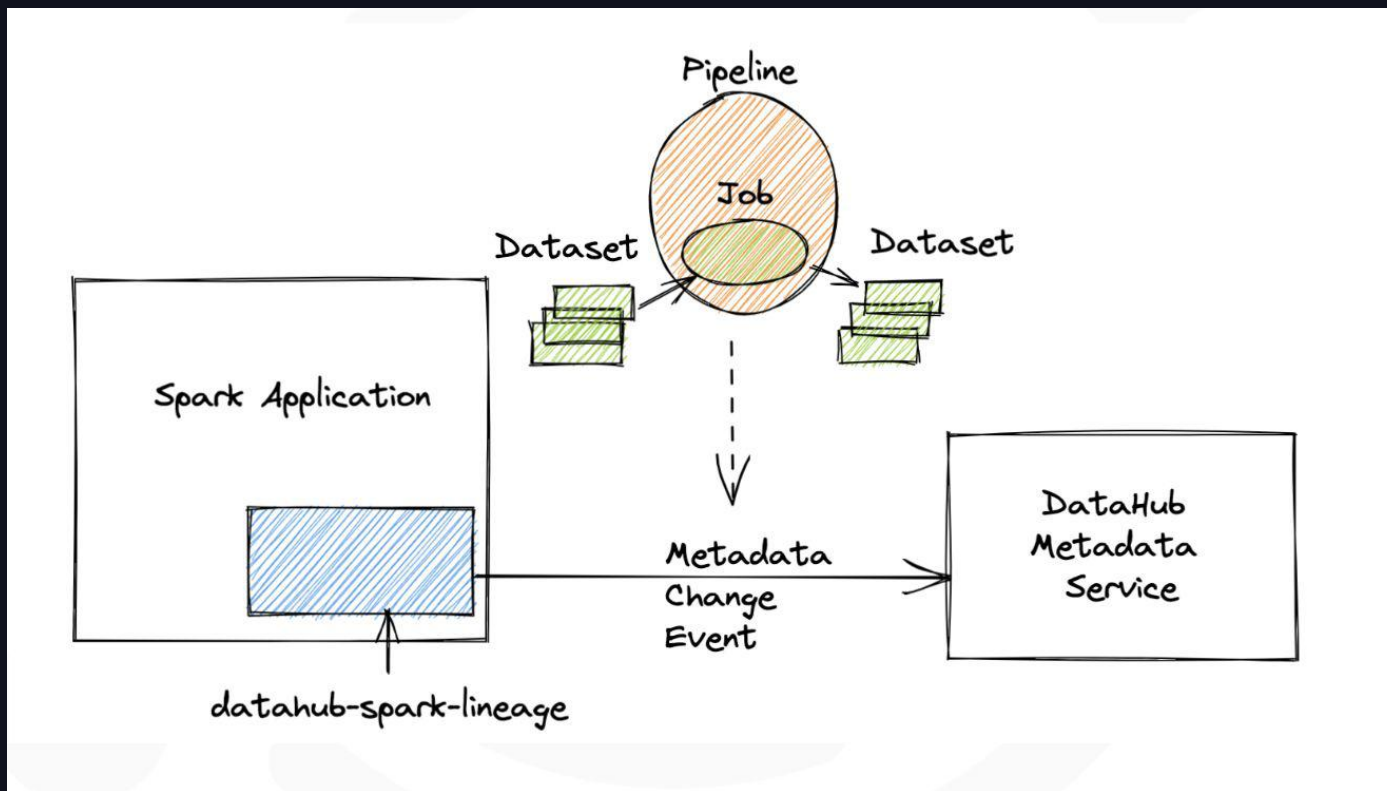
Integrations: Delta Lake (new!)

The screenshot shows the Metadatalabs interface for a Delta Lake dataset. At the top, there is a search bar and a breadcrumb trail: Datasets > prod > delta-lake > long_tail_companions > adoption > pet_profiles. Below the breadcrumb, the dataset name 'long_tail_companions/adoption/pet_profiles' is displayed, along with a 'Dataset' label and 'Delta Lake' provider. A navigation bar includes 'Schema' (selected), 'Documentation', 'Properties', 'Lineage', 'Queries', 'Stats', and 'Validation'. A status bar indicates the schema was reported at 29/06/2022, 11:21 GMT+5:30, with a 'Normal' status, a 'Blame' button, and a version '0.0.0 - 14 seconds ago'. The main content is a table with columns for Field, Description, Tags, and Terms.

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 during intake		
coat_type (String)	Coat type of pet, as determined during intake		
name (String)	Name assigned to pet during intake		


Metadata 360

Integrations: Push-based Integration with Spark



Search for groups, artifacts, categories

Home » io.acryl » datahub-spark-lineage

 **Datahub Spark Lineage**
Library to push data lineage from spark to datahub

License	Apache 2.0
Tags	spark io
Ranking	#391846 in MvnRepository (See Top Artifacts)

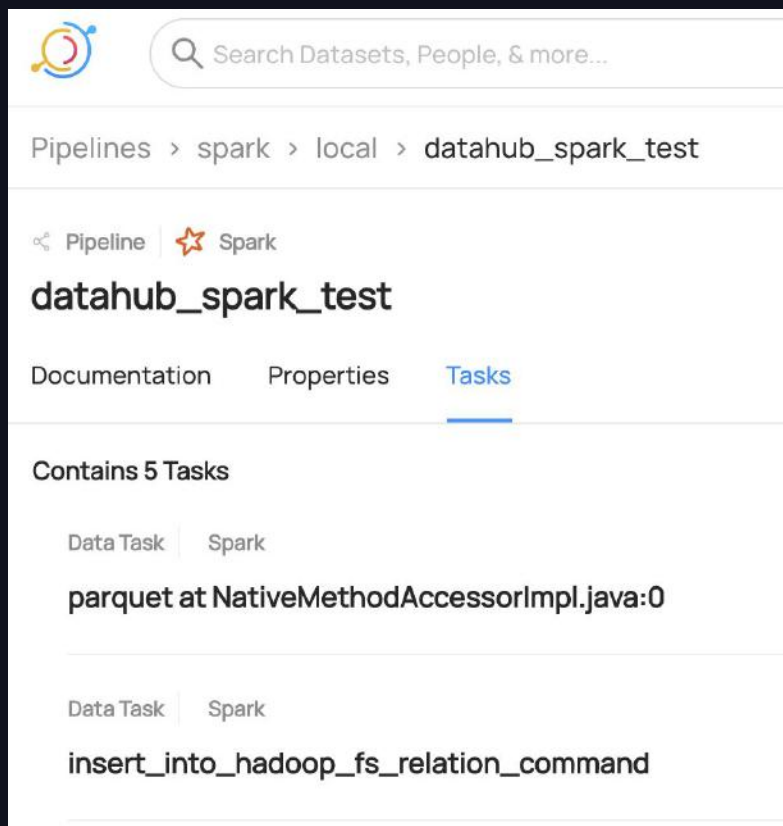
Central (15)

Version
0.8.38
0.8.36

```
.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



Search Datasets, People, & more...

Pipelines > spark > local > datahub_spark_test

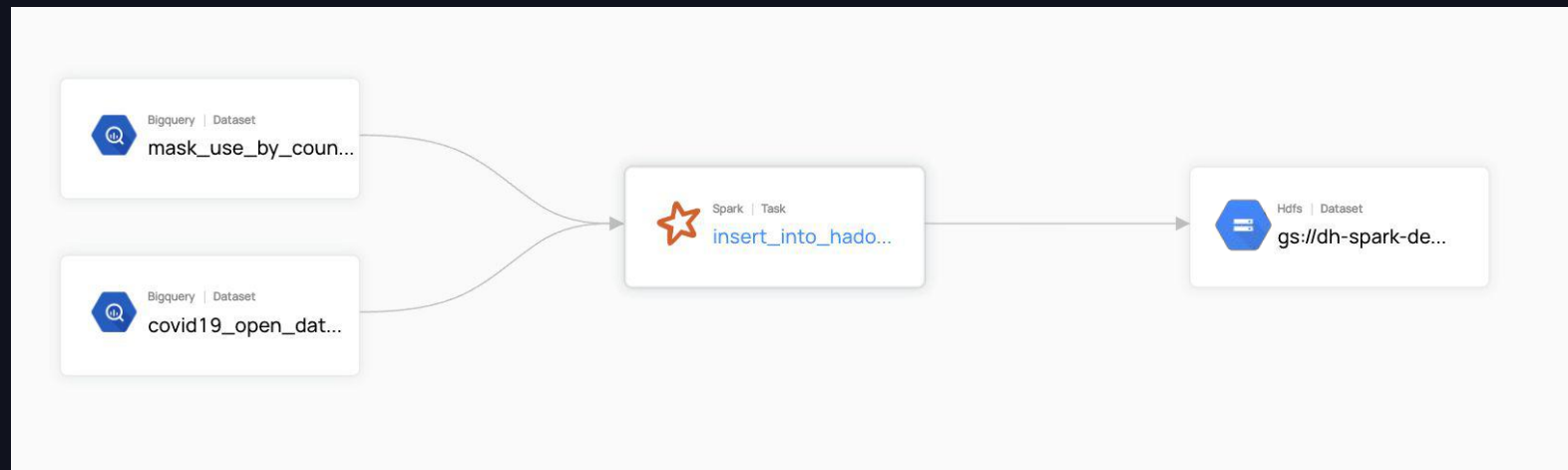
Pipeline | Spark

datahub_spark_test

Documentation | Properties | **Tasks**

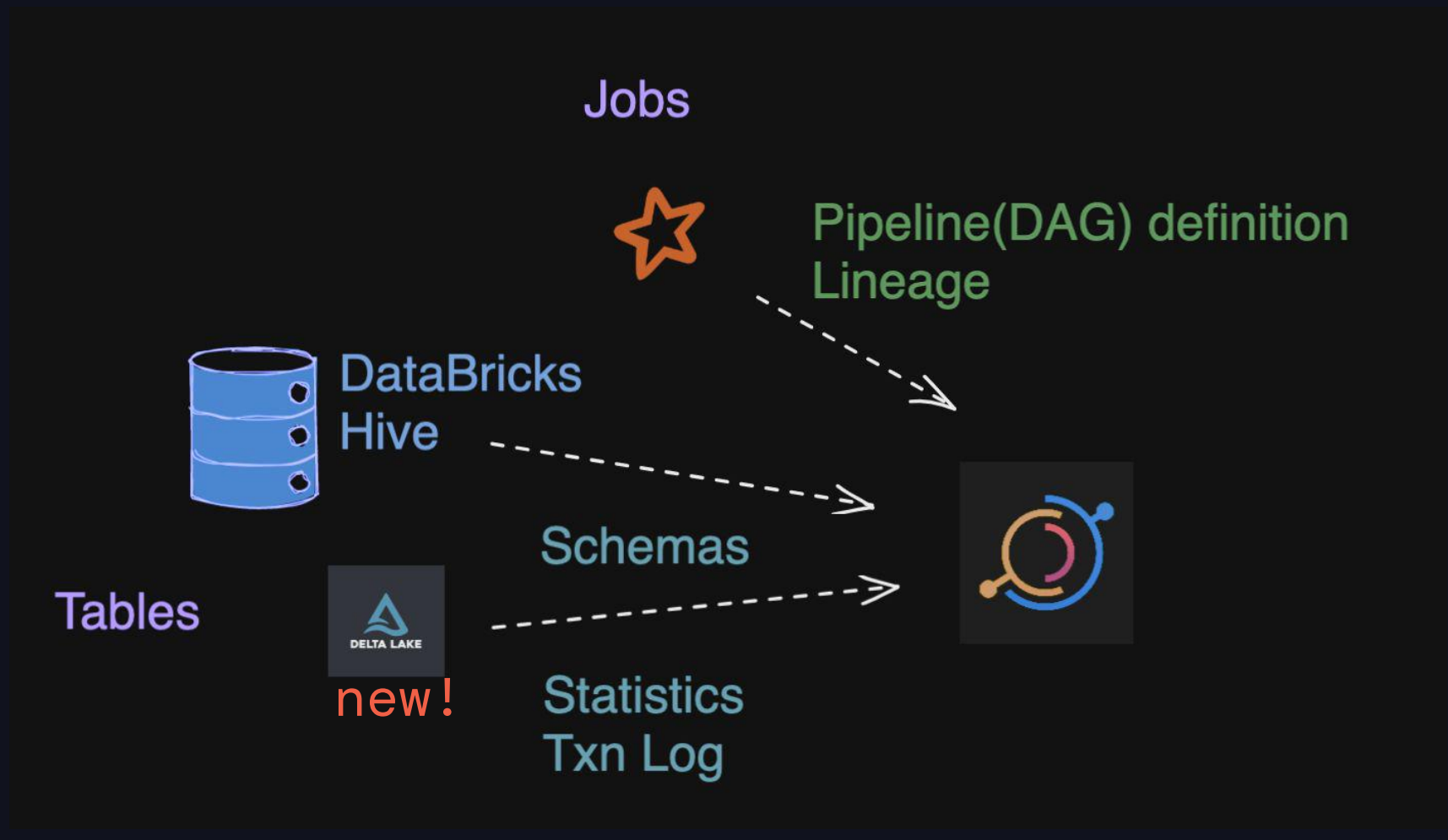
Contains 5 Tasks

- Data Task | Spark
parquet at NativeMethodAccessorImpl.java:0
- Data Task | Spark
insert_into_hadoop_fs_relation_command



Metadata 360

Integrations Recap



Metadata 360

Searching for pet_profile

The screenshot shows the Acryl DataHub interface with a search query for 'pet_profile'. The search results are displayed in a list format, showing various data assets related to the search term. The left sidebar contains filters for Type, Tag, Glossary Term, Domain, Owned By, Sub Type, and Platform. The main content area shows the following results:

- 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.petstathistory • 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):** pet_profiles

Metadata 360

Dive into technical metadata

The screenshot displays a metadata management interface for a table named 'pet_profiles'. The interface includes a search bar at the top, navigation breadcrumbs, and a sidebar with tabs for Schema, Documentation, Properties, Lineage, Queries, Stats, Validation, and Incidents. The 'Schema' tab is active, showing a table with columns for Field, Description, Tags, and Terms. A yellow highlight is placed over the 'View Technical Schema' button. The table lists fields such as profile_id, species, breed, sex, and color, each with a description and associated tags like Sensitive, Confidential, Breed, and Email. The right sidebar provides details about the table, including its location, a description, and statistics such as 43,517 rows and 13 columns.

Search Datasets, People, & more...

Analytics MyRequests 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 X
species (String)	Species of pet, either feline or canine		Confidential X
breed (String)	Breed of pet as determined during intake	(edited)	Breed X
sex (String)	Sex of pet		Confidential X
color (String)	Color of pet, as determined during intake		Email X

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 X + Add Tags

Glossary Terms
Confidential X Sensitive X + Add Terms

Metadata 360

Dive into technical metadata

The screenshot displays the Databricks Metadata 360 interface for a table named 'pet_profiles'. The breadcrumb navigation shows the path: Datasets > prod > databricks > long_tail_companions > adoption > pet_profiles. The table is located in the 'long_tail_companions' > 'adoption' folder. The interface includes a search bar at the top, navigation tabs for Schema, Documentation, Properties, Lineage, Queries, Stats, Validation, and Incidents. The 'pet_profiles' table is reported at 3/20/2022, 09:23 PM PDT with a 'Normal' status and 'Blame' assigned. The table schema is shown in a table with columns for Field, Description, Tags, and Terms. The 'About' panel on the right provides usage statistics: 43,517 rows, 13 columns, 2710 monthly queries, and was last updated on 5/23/2022, 07:12 PM CDT. The 'Tags' section shows 'Tier1' and an 'Add Tags' button. The 'Glossary Terms' section is also visible.

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 X
species String	Species of pet, either feline or canine		Confidential X
breed String	Breed of pet as determined during intake (edited)		Breed X
sex String	Sex of pet		Confidential X
color String	Color of pet, as determined during intake		Email X

About

Understand Usage

Stats [More stats >](#)

Rows	Columns
43,517	13
Monthly Queries	Top Users
2710	B M +2
Last Updated	
5/23/2022, 07:12 PM CDT	

Tags

Tier1 X + Add Tags

Glossary Terms



Metadata 360

Understand Business Context

pet_profiles ✓

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	Breed of pet as determined during intake <i>(edited)</i>	Breed	
sex String	Sex of pet		
color String	Color of pet, as determined during intake		
coat_type String	Coat type of pet, as determined during intake		
name String	Name assigned to pet during intake		

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

Apply Logical/Business Metadata

Tags

Tier1 X + Add Tag

Glossary Terms

Sensitive X Confidential X + Add Term



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 <i>(edited)</i>	Breed ×
sex	String	Sex of pet	
color	String	Color of pet, as determined during intake	
coat_type	String	Coat type of pet, as determined during 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

The screenshot shows the Metadata 360 interface for the 'pet_profiles' table. The breadcrumb path is 'Datasets > prod > snowflake > long_tail_companions > adoption > pet_profiles'. The table is reported at 3/20/2022, 11:23 PM CDT with a 'Normal' status. The 'View Operational Detail' callout box is positioned over the 'Queries' tab in the navigation menu.

Field	Description	Tags	Terms
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 <small>(edited)</small>	Breed	

Stats

Rows	Columns
43,517	13
Monthly Queries	Top Users
2710	B M +26
Last Updated	
5/23/2022, 07:12 PM CDT	

Metadata 360

Stats

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

Table Stats

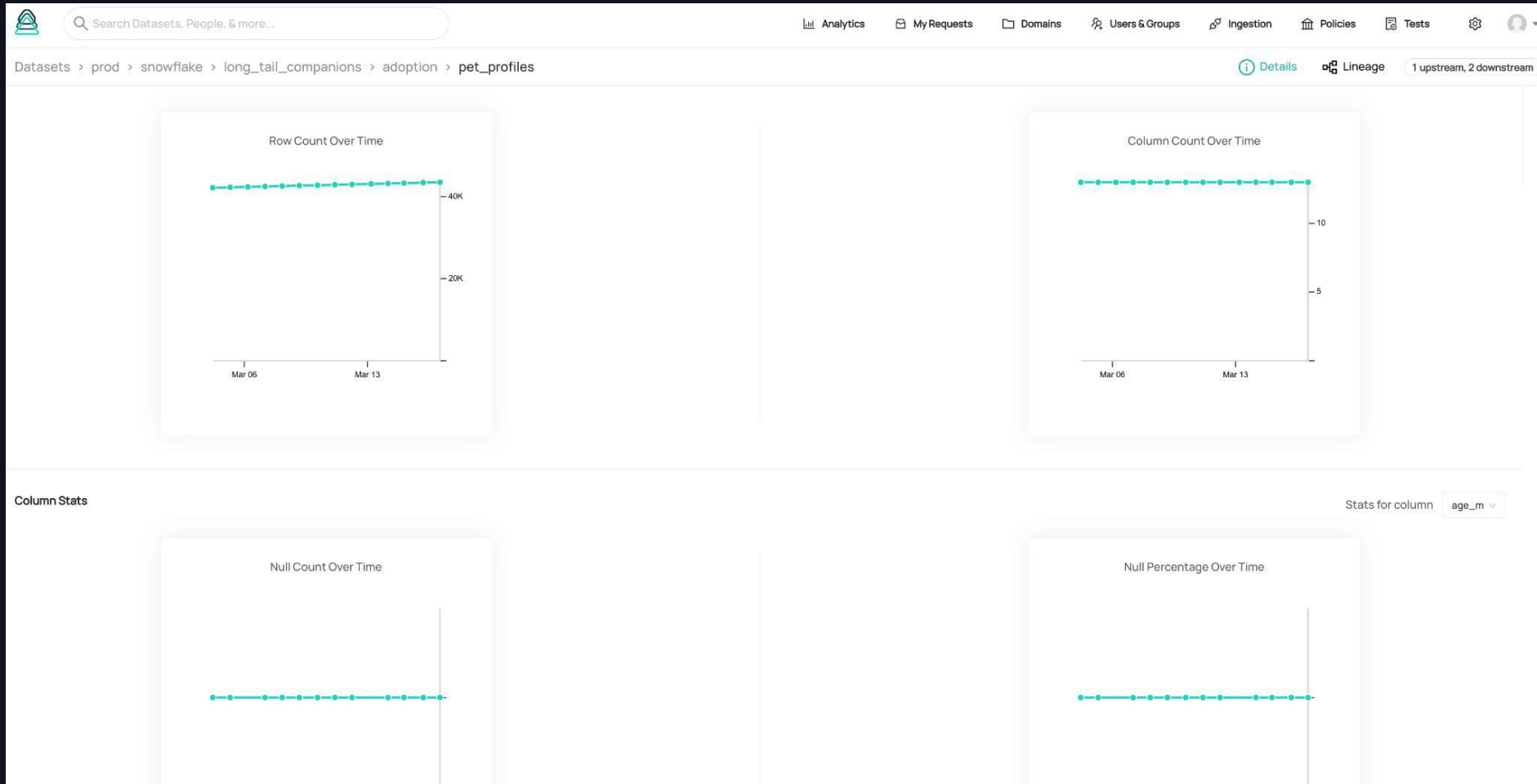
Rows: **43.5K** Columns: **13**

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

The screenshot displays the Acryl DataHub interface for a table named `pet_profiles`. The breadcrumb navigation is `Datasets > prod > snowflake > long_tail_companions > adoption > pet_profiles`. The `pet_profiles` table is selected, and the `Queries` tab is active. The main content area shows a SQL query:

```
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
```

Below the query, a second query is partially visible:

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

The right-hand sidebar contains several sections:

- About:** "this table contains profile details of all pets" and an `+ Add Link` button.
- Stats:** A table showing `Rows: 43,517` and `Columns: 13`. Below it, `Monthly Queries: 2710` and `Last Updated: 5/23/2022, 07:12 PM CDT`. A `More stats >` link is also present.
- Tags:** A `Tier1` tag and an `+ Add Tag` button.
- Glossary Terms:** `Sensitive` and `Confidential` tags, with an `+ Add Term` button.
- Owners:** A list of users with their initials in a circle: `Adoption`, `Ricea MacAnespie`, `Prentiss Matussov`, and `Avigdor Bramhall`.

Metadata 360

Data Quality assertions

The screenshot displays the Acryl DataHub interface for a dataset named 'pet_profiles'. The breadcrumb navigation is 'Datasets > prod > snowflake > long_tail_companions > adoption > pet_profiles'. The 'Validation' tab is active, showing a summary: 'All assertions have passed' with 8 successful and 0 failed assertions. Below this, six individual assertions are listed, all marked as 'Passed':

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

The right-hand sidebar provides additional information about the table:

- About:** this table contains profile details of all pe...
- Stats:** 43,517 Rows, 13 Columns. Includes a 'More stats >' link.
- Monthly Queries:** 2710
- Top Users:** B, M, and a user with 26 queries.
- Last Updated:** 5/23/2022, 07:12 PM CDT
- Tags:** Tier1 (with a close icon) and an 'Add Tag' button.
- Glossary Terms:** (Section header)



Metadata 360

Business Context using Business Glossary

The screenshot displays the Metadatalytics Business Glossary interface. At the top, there is a search bar labeled "Search Datasets, People, & more...". Below it, the breadcrumb navigation shows "Glossary > PersonalInformation > Email".

On the left side, there is a "Search Glossary" input field and a tree view of categories. The "PersonalInformation" category is expanded, and "Email" is highlighted in light green. Other categories include zO, R&D updated, AIDE, Classification, TestGrp2, Space, OKRs, Sales, ClientsAndAccounts, Semantix, Adoptions, Marketing, MyRootGrp1, TASA, Age, Age name, and Age name.

The main content area shows the "Email" glossary term. It has a "Glossary Term" icon and a link. Below the title are tabs for "Documentation", "Related Entities", "Related Terms", and "Properties". The "Related Terms" tab is active, showing a table with two columns: "Contains" and "Inherits".

Contains	Inherits
	<p>Glossary Term</p> <p>Confidential</p> <p></p>

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
1  version: 2
2
3  models:
4    - name: pet_details
5      description: Table with all pet-related details
6      meta:
7        owner: "shannon@longtail.com"
8        model_maturity: prod
9        contains_pii: false
10       business_critical: true
11       domain: "Pet Adoptions"
12
```

Shift Left

The screenshot displays the Acryl DataHub interface for a dataset named 'pet_details'. The breadcrumb navigation shows the path: Datasets > prod > dbt > long_tail_companions > analytics > pet_details. The main content area shows the dataset name 'pet_details' with tabs for Schema, View Definition, Documentation, Properties, Lineage, and Queries. The 'Documentation' tab is active, displaying a description: 'Table with all pet-related details'. A code editor window is overlaid on the bottom left, showing the schema definition in 'schema.yml' format:

```
1 version: 2
2
3 models:
4   - name: pet_details
5     description: Table with all pet-related details
6     meta:
7       owner: "shannon@longtail.com"
8       model_maturity: prod
9       contains_pii: false
10      business_critical: true
11      domain: "Pet Adoptions"
```

The right sidebar contains metadata sections: 'Tags' with 'business_critical' and 'prod_model', 'Glossary Terms' with 'DaysInStatus', 'Owners' with 'Shannon Lovett', and 'Domain' with 'Pet Adoptions'. The top navigation bar includes links for Analytics, My Requests, Domains, Users & Groups, Ingestion, Policies, and Tests.




Shift Left

Declare & collect metadata at the source



```
1  syntax = "proto3";
2  package ecommerce;
3  import "protobuf/meta/meta.proto";
4  import "common/context.proto";
5
6
7  message SearchResult {
8  |   int32 item_id=1;
9  }
10
11  /**
12  |   The event emitted whenever a Search is executed
13  | */
14  message SearchEvent {
15  |   option(meta.msg.classification) = "Classification.Sensitive";
16  |   option(meta.msg.team) = "Ecommerce";
17
18  |   common.EventContext context = 4;
19  |   // the search identifier
20  |   int32 search_id = 1;
21  |
22  |   repeated SearchResult result_array = 3;
23  }
```

Details **Lineage** 1 upstream, 1 downstream

About 

The event emitted whenever a Search is executed



+ Add Link

Tags



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

+ Add Tag

Glossary Terms

 Sensitive  + Add Term

Owners

 Ecommerce  + Add Owner



Shift Left

Declare metadata at source => high quality metadata

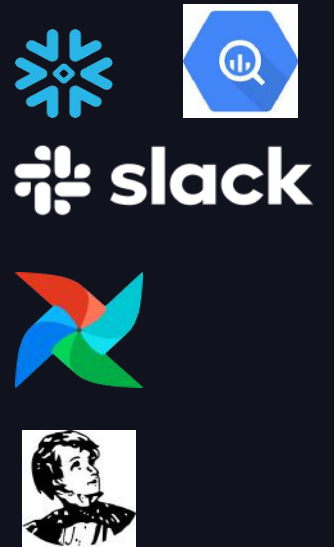
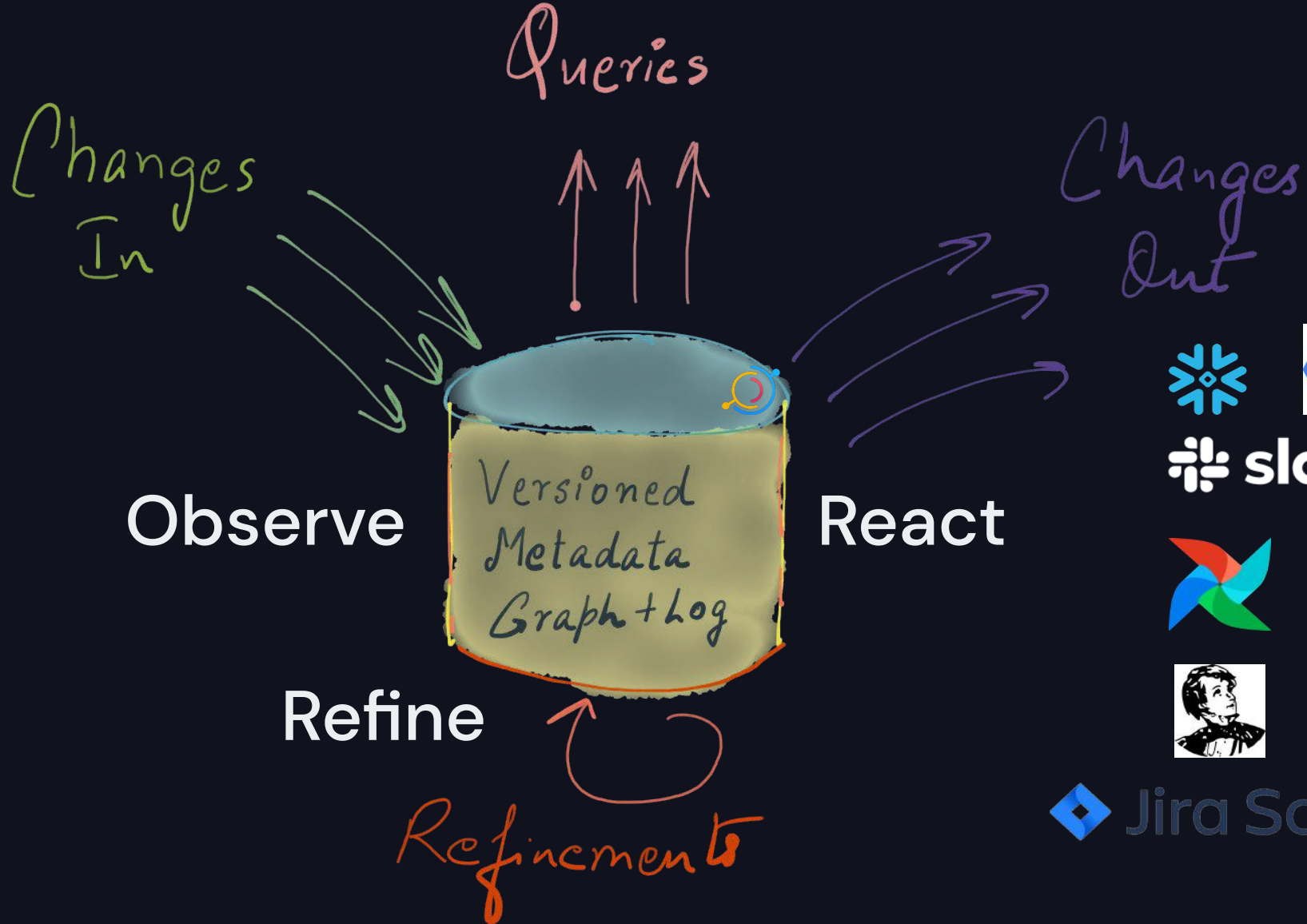
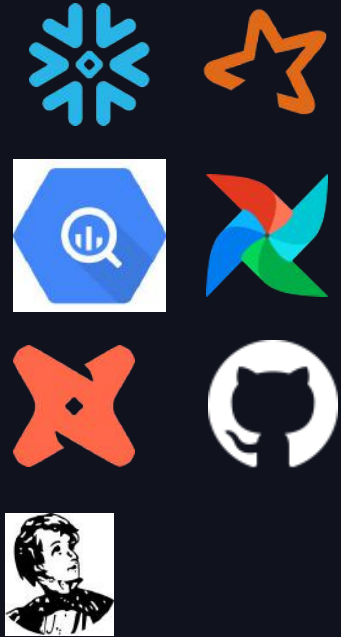
What is the authoritative pet_profile dataset?



Active Metadata

Inject metadata into the operational plane

⚡ DataHub's Architecture





Active Metadata

DataHub Actions let you react to changes in real-time

```
# 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
```

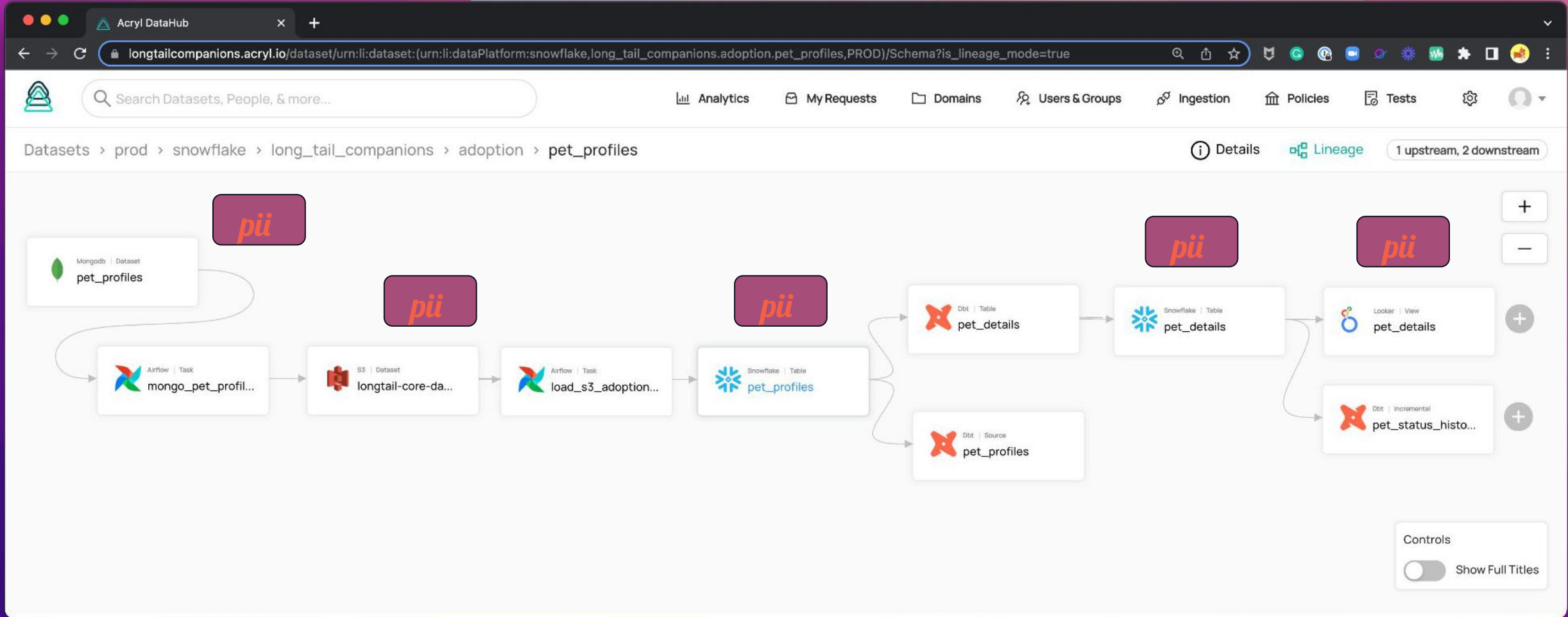
⚡ Active Metadata

DataHub Actions let you react to changes in real-time

```
31
32 # A basic example of a DataHub action that prints all
33 # events received to the console.
34 class HelloWorldAction(Action):
35     @classmethod
36     def create(cls, config_dict: dict, ctx: PipelineContext) -> "Action":
37         action_config = HelloWorldConfig.parse_obj(config_dict or {})
38         return cls(action_config, ctx)
39
40     def __init__(self, config: HelloWorldConfig, ctx: PipelineContext):
41         self.config = config
42
43     def act(self, event: EventEnvelope) -> None:
44         print("Hello world! Received event:")
45         message = json.dumps(json.loads(event.as_json()), indent=4)
46         if self.config.to_upper:
47             print(message.upper())
48         else:
49             print(message)
50
51     def close(self) -> None:
52         pass
```

Active Metadata

Tag Propagation using real-time actions



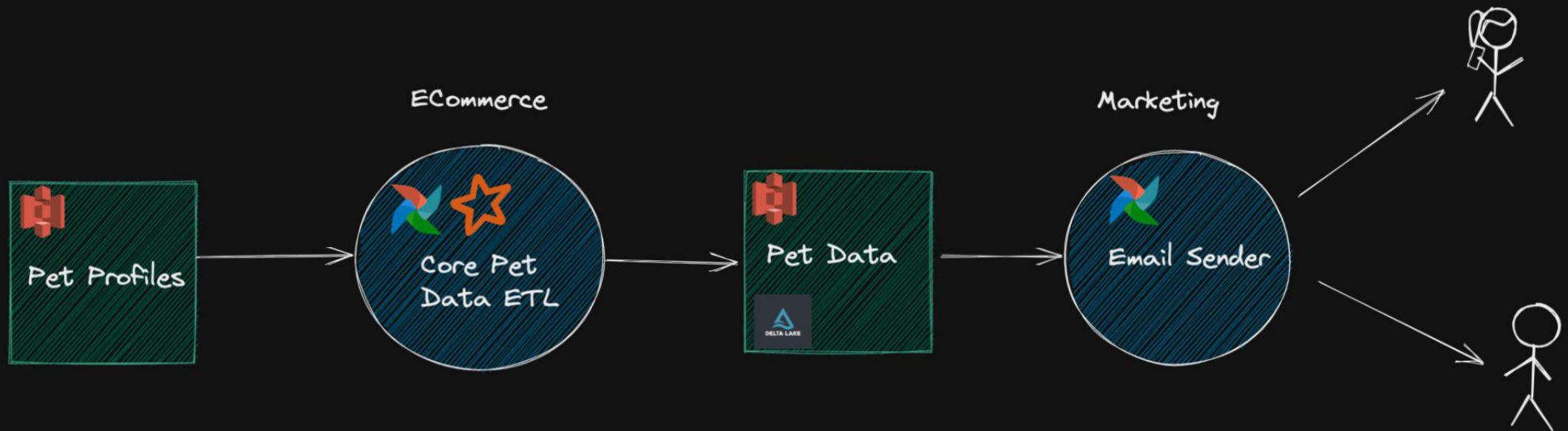


Active Metadata

Prevent bad data
from causing more damage

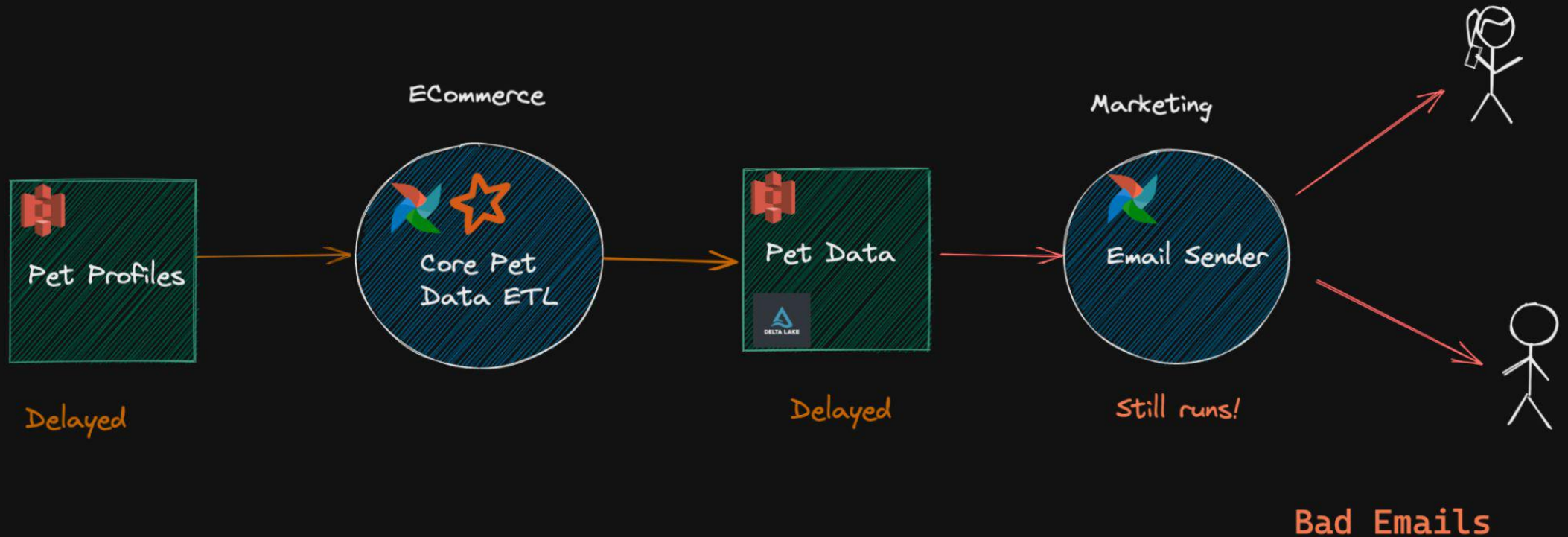
⚡ A Typical Scenario

Data Standardization followed by Marketing Emails



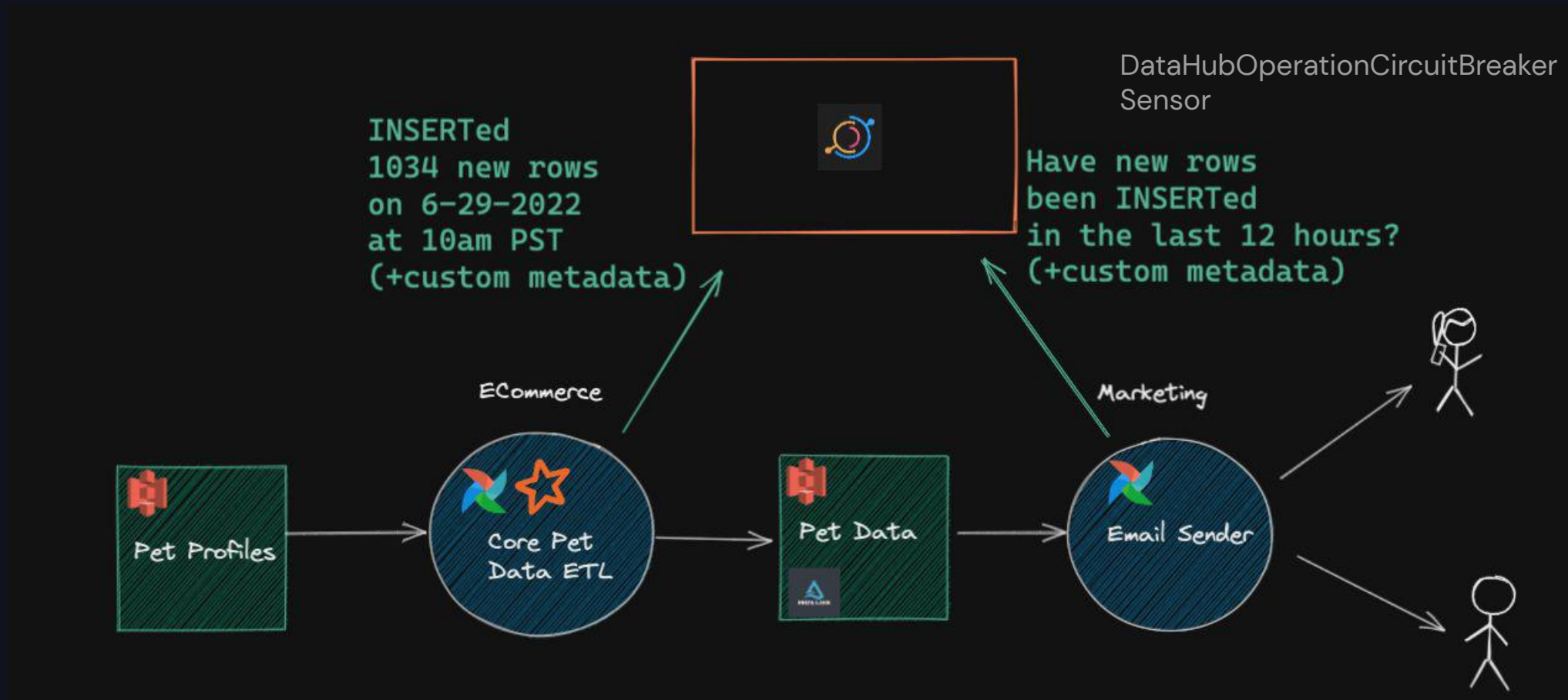
⚡ A Problem: Delayed Data

One day...



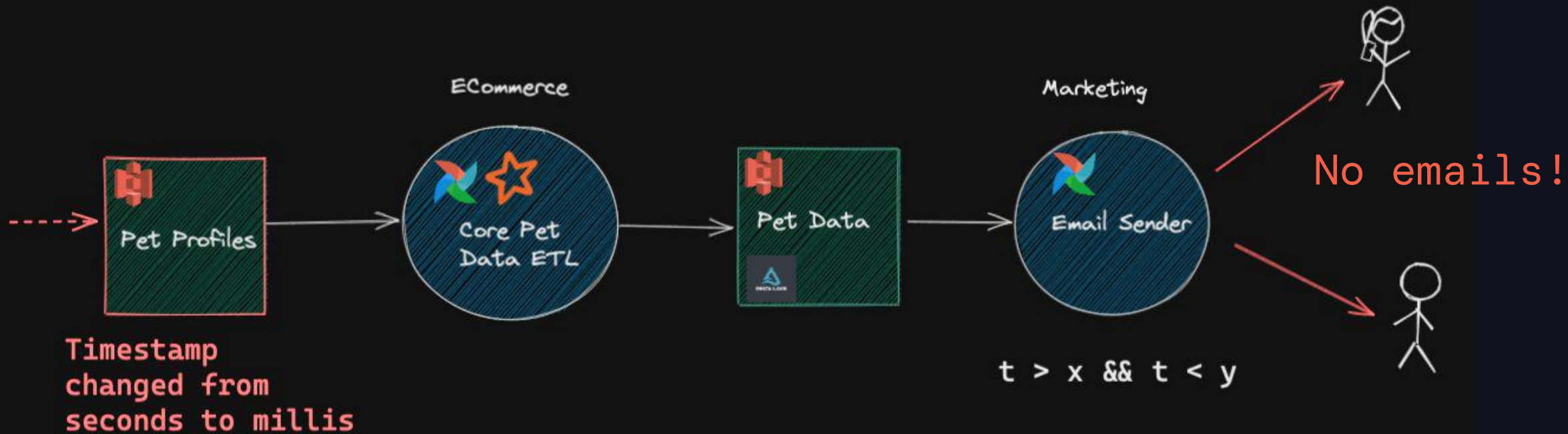
⚡ A Problem: Delayed Data

Step 2: Verify



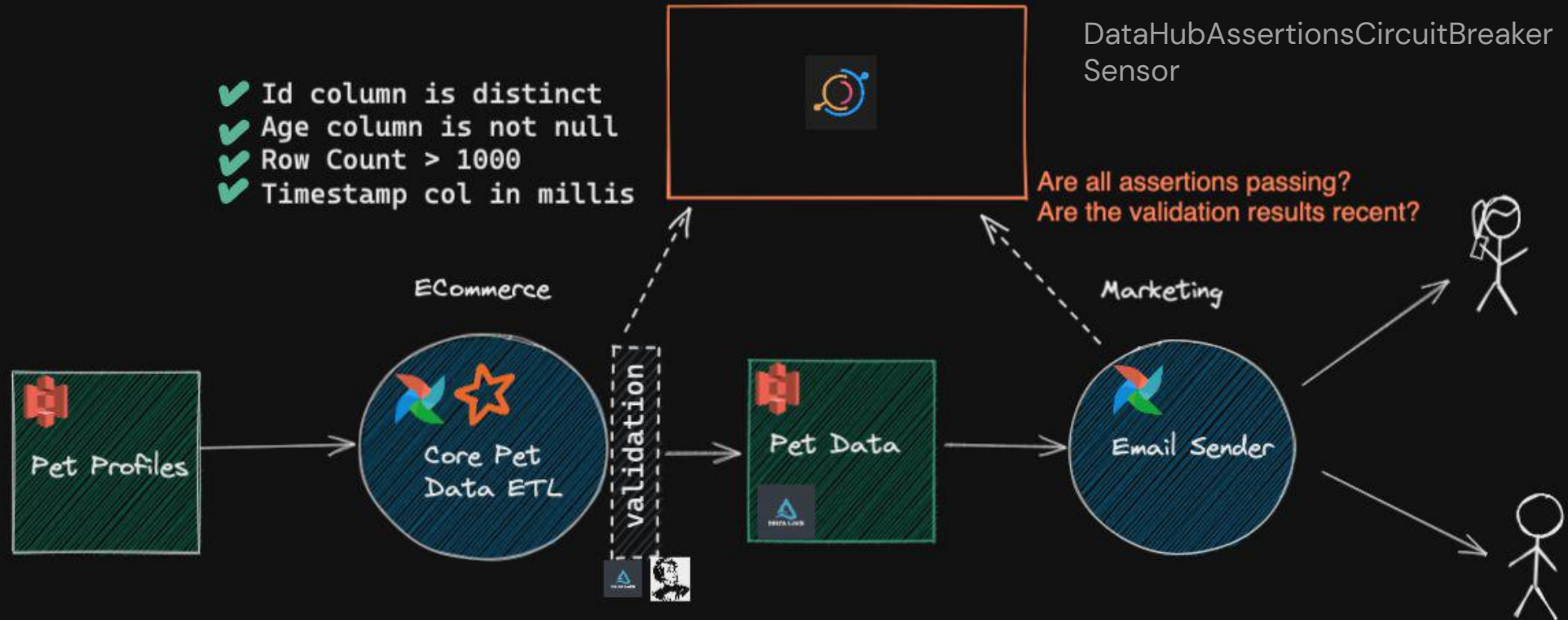
⚡ Another Problem: Broken Data

A few months later...



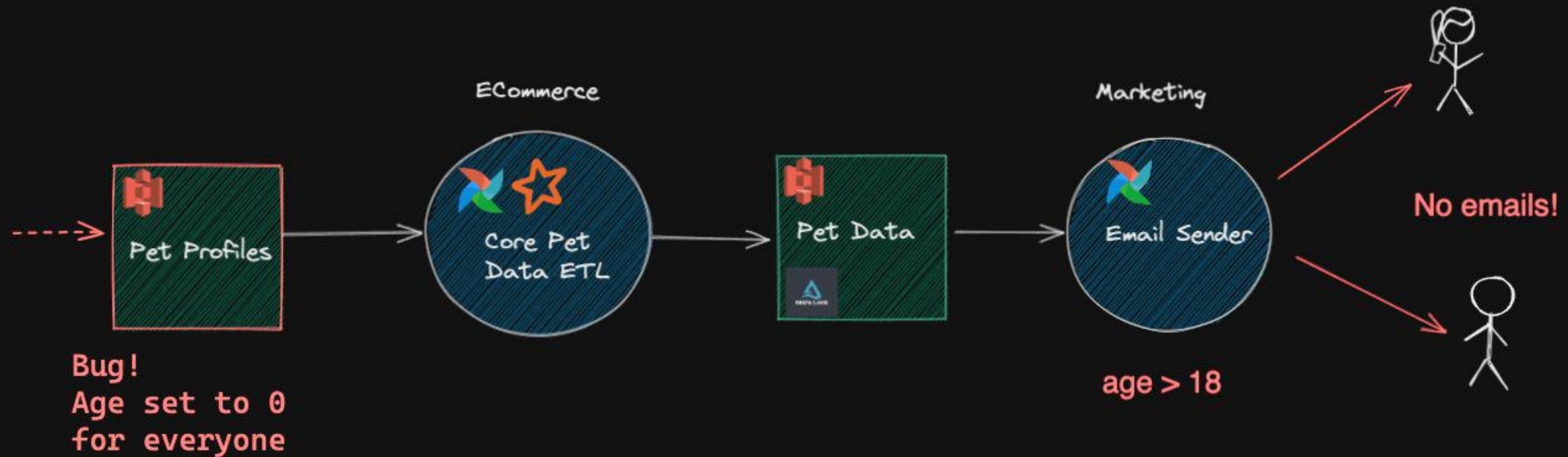
⚡ DataHub Assertions

Step 2: Verify



⚡ Another Problem: Broken Data

A few weeks later..



Tests can't catch everything

DataHub Incidents

Step 1: Raise Incident

+ Raise Incident

ions > adoption > pet_profiles

tion

Queri

Operati

with age

Raise Incident

Type

Operational

* Title

Data Backfill - Age set to 0 for all newly added profiles.

* Description

Starting on May 20, 2022 new profiles were c

Cancel Add

Table Snowflake > long_tail_companions > adoption

pet_profiles ✓ 📄 ⚠️

Schema Documentation Properties Lineage Queries Stats Validation Incidents

+ Raise Incident All

⚠️ **There are 2 active incidents**
2 active incidents, 2 resolved incidents

Data Backfill - Age set to 0 for all newly added profiles. Operational

Starting on May 20, 2022 new profiles were mistakenly created with age = 0 on the profiles. This is currently being backfilled. ✓ Resolve ⚠️

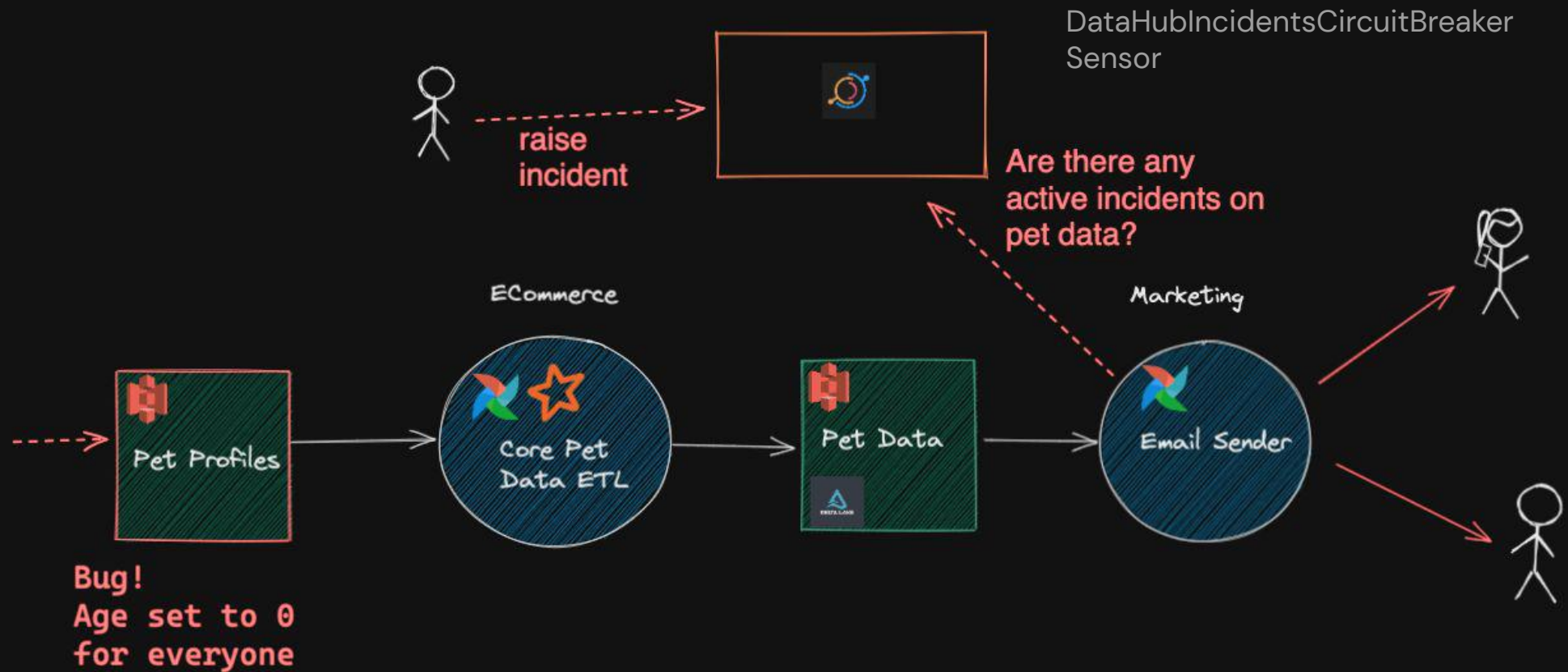
22 May 2022 (America/Los_Angeles)

Incident raised because of DAG failure Operational

Run manual__2022-05-20T12:09:32.735583+00:00 failed for dag: marketing-send_emails because task_failure Resolved on 20 May 2022 by Admin ✓

⚡ DataHub Incidents

Step 2: Verify



The Reliability Hierarchy of Needs

Toolkit

DataHub Incidents



DataHub Assertions



DataHub Operations



?

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

DataHub Tests



Step 1: Define Tests

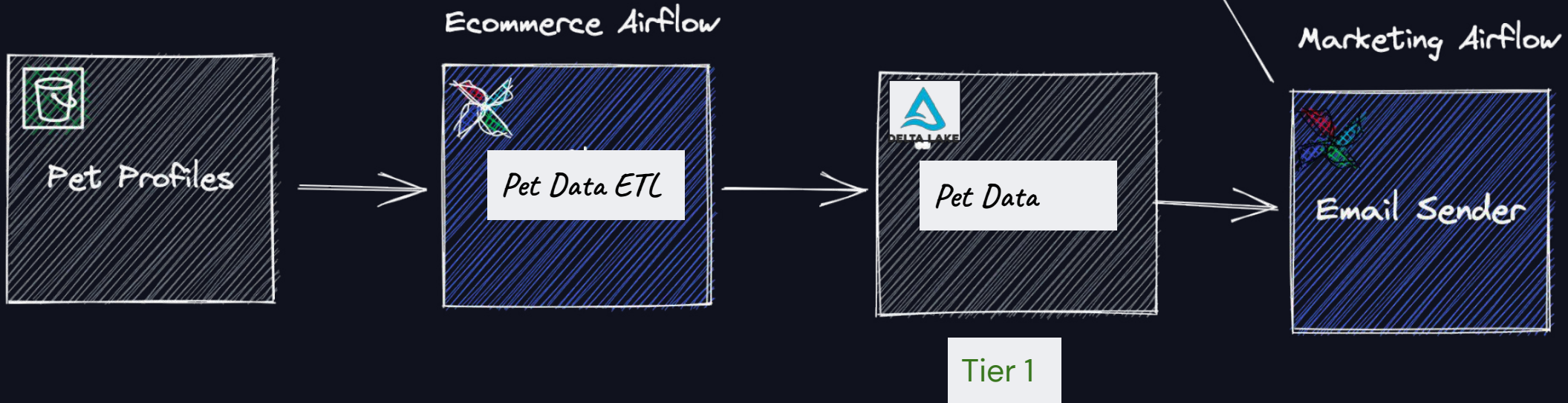
"All Tier1 Datasets should have passing assertions"



DataHubTestsCircuitBreaker
Sensor

Step 2: Verify

"Are my inputs passing Tests?"



DataHub Tests

Central policy definition, distributed enforcement

Name
Give your test a name.

Category
The category of your test.

Description
An optional description to help keep track of your test.

Define your Test
For more information about how to configure a Test, check out the [DataHub Tests Guide](#).

```
1 on:
2   dataset:
3     -
4       field: tags
5       condition: EQUALS
6       values:
7         - 'urn:li:tag:Tier1'
8   rules:
9     -
10      field: assertions
11      condition: EXISTS
12    -
13      field: assertions.runEvents.result.type
14      condition: NOT_EQUALS
15      value: FAILURE
16
```

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*

```
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

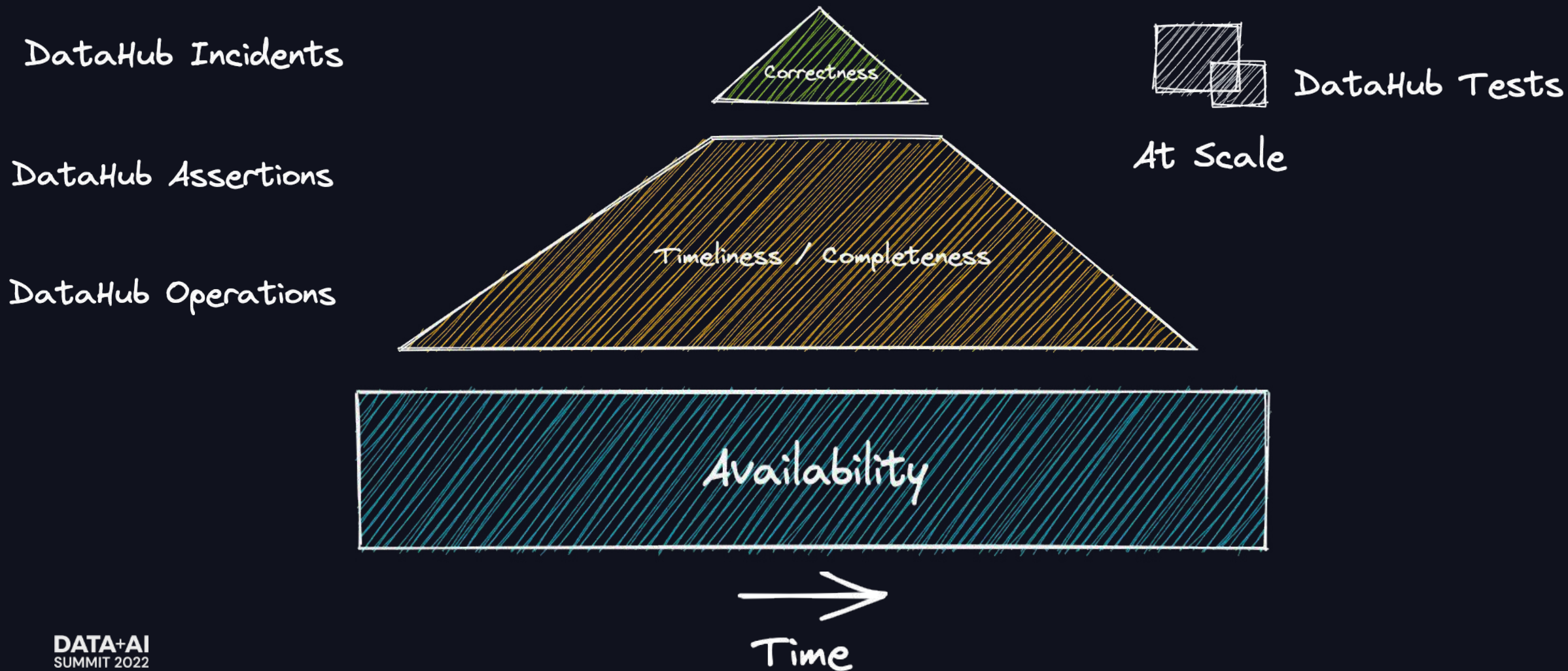
Create a Metadata Test
Circuit Breaker

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



Active Metadata

Inject metadata into the operational plane

How can I ensure my ML features exclude PII?



Can I rely on critical data products to be up to date and accurate?



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

