

The Data Practitioner's Guide to Data Discovery

Shirshanka Das & Maggie Hays

Data Council, Mar 2022, Austin

About Us



Shirshanka Das

Co-Founder & CEO

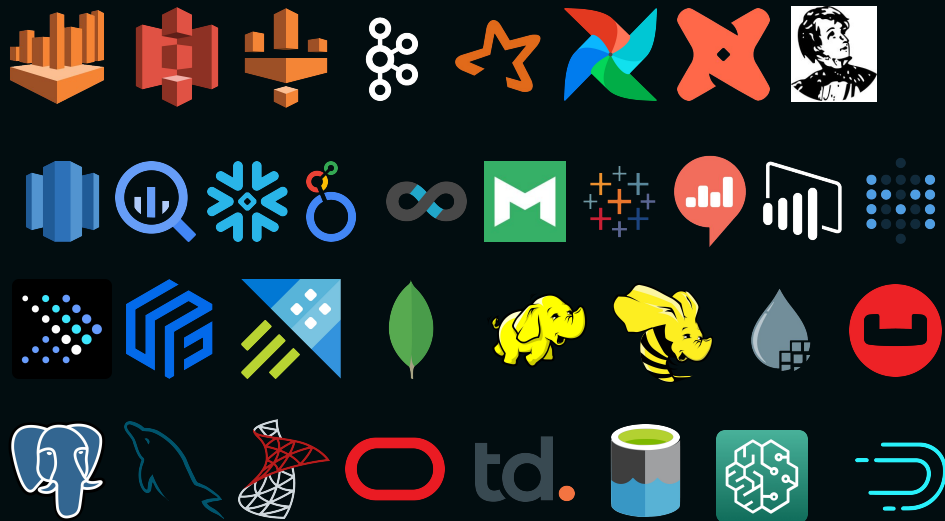


Maggie Hays

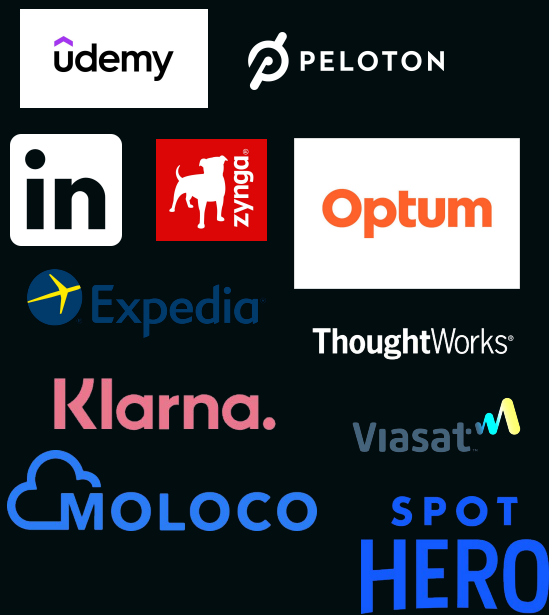
Founding Community Product
Manager

DataHub: #1 OSS Metadata Platform for the Modern Data Stack

DataHub Integrations



DataHub Adopters



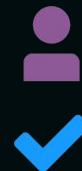


DataHub



2,675 Slack Members

10x YoY Growth
Across 56 Countries & 27 Local Time Zones



97 Contributors

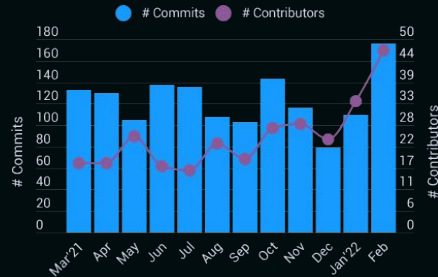
1,477 Commits

881,733 ++ 542,559 --

Top Member Roles



Top Member Industries



5.2k

GitHub Stars




665

YouTube Subscribers



371

Blog Subscribers



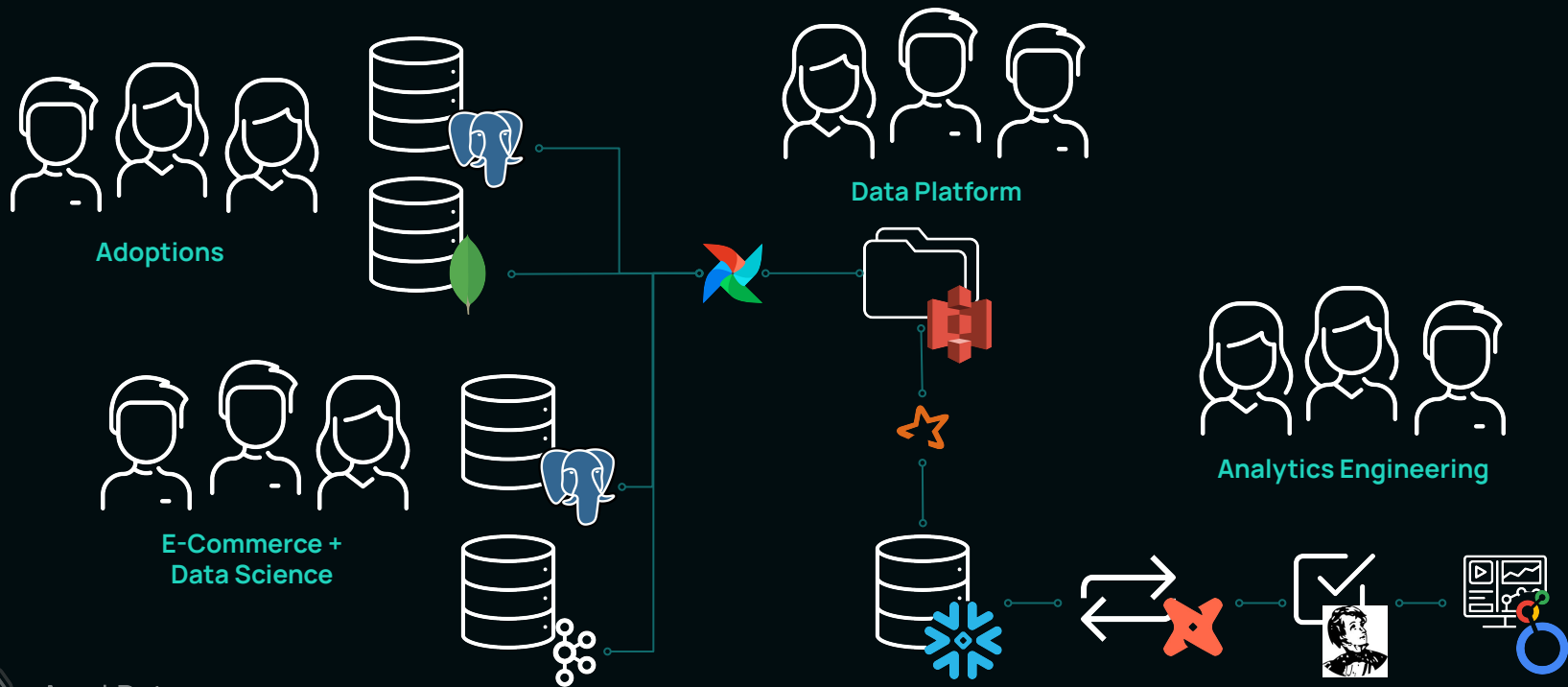
We Need to Re-Think Data Discovery



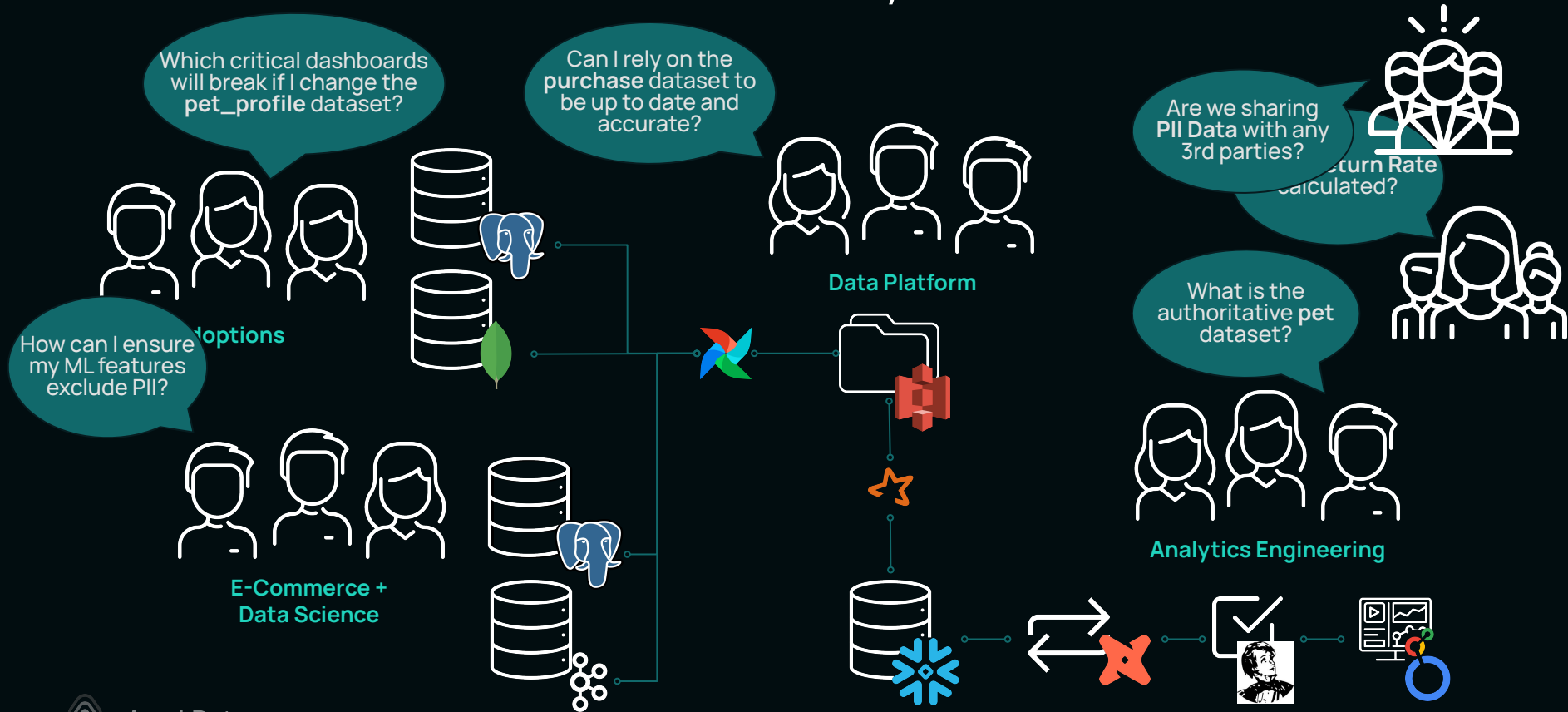
Example: Long Tail Companions

Where Every Pet is Exceptional

Long Tail Companions' Fragmented Data Stack



LTC Data Practitioners Routinely Ask:



A Typical Data Catalog Story

- **Typical Approach** → Connect to all systems, crawl, reconstruct truth, give consumers tools to add documentation in the application → Hope we get good discovery.
- **The Truth** → it doesn't work. Achieves basic discovery but doesn't really solve the questions we asked. All you get is a monolithic web-app where driving adoption is hard after the initial excitement wears off.

Top Issues with Most Data Catalogs

What is the authoritative **pet** dataset?



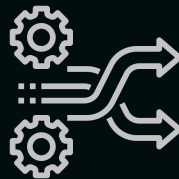
Physical Metadata is not intuitive to everyone

Can I rely on the **purchase** dataset to be up to date and accurate?



Crawl-Only Ingestion Leads to Stale Metadata

How can I ensure my ML features exclude PII?



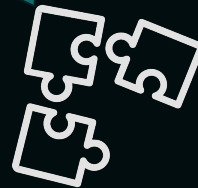
No approach to acting on changes in metadata

How is **Return Rate** calculated?



Manual enrichment of metadata leads to problems

Are we sharing **PII Data** with any 3rd parties?



Over-indexed on Data Warehouses

Which critical dashboards will break if I change the **pet_profile** dataset?

3 Must-Haves for Sustainable Data Discovery

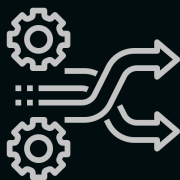
1. Metadata 360



Physical Metadata is not intuitive to everyone



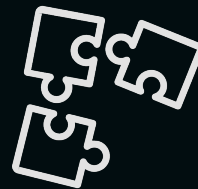
Crawl-Only Ingestion Leads to Stale Metadata



No approach to acting on changes in metadata



Manual enrichment of metadata leads to problems



Over-indexed on Data Warehouses

What Does this Look Like in Practice?

Metadata 360: Physical + Business Metadata

How is Return Rate calculated?



Manual enrichment of metadata leads to problems

A screenshot of the Acryl DataHub web interface. The browser address bar shows 'longtailcompanions.acryl.io/search?page=1&query=%2A'. The interface has a search bar at the top with the text 'Search Datasets, People, & more...'. Below the search bar, there are navigation tabs for 'Analytics', 'My Requests', 'Domains', 'Users & Groups', 'Ingestion', and 'Policies'. The main content area is divided into a left sidebar with filters and a main results pane. The filters include 'Type' (Datasets (77), Users (74), Glossary Terms (12), Groups (9), Charts (8), Containers (7), Domains (4), Tags (3), Dashboards (2)), 'Sub Type' (View (13), Source (9), Seed (8), Schema (6), Table (4), Explore (1), Database (1)), 'Platform' (Snowflake (53), dbt (21), Looker (20)), and 'Domain' (Pet Adoptions (12), E-Commerce (5)). The main results pane shows 'Showing 1 - 10 of 196 results' and lists several datasets: 'active_customer_itv' (Snowflake, Dataset, analytics), 'pet_profiles' (Snowflake, Dataset, adoption), 'pet_details' (Snowflake, Dataset, analytics), 'adoptions' (Dbt, Source), 'human_profiles' (Dbt, Source), 'humans' (Dbt, Source), and 'pet_profiles' (Dbt, Source). Each dataset entry includes a brief description and a 'Queried 1000+ times in the past month' indicator.

Metadata 360: Physical + Business Metadata

How is Return Rate calculated?



Manual enrichment of metadata leads to problems

```
return_rate.md
1  {% docs return_rate %}
2
3  The percentage of adopted animals that were returned to the shelter within 60 days of
4  adoption
5
6  ```sql
7  with adoption_date as (
8  select
9  profile_id
10 , as_of_date as adoption_date
11 from
12 analytics.pet_status_history
13 where
14 status = 'adopted'
15 )
16
17 , return_date as (
18 select
19 profile_id
20 , as_of_date as return_date
21 from
22 analytics.pet_status_history
23 where
24 status = 'returned to facility'
25 )
26
27 , days_until_return as (
28 select
29 adoption_date.profile_id
30 , adoption_date.adoption_date
31 , return_date.return_date
32 , date_diff(adoption_date.adoption_date, return_date.return_date) as
33 days_until_return
34 from
35 adoption_date
36 left join
37 return_date
38 using(profile_id)
39 )
40 select
41 count(distinct
42 case
43 when return_date is not null
44 and days_until_return < 60
45 then profile_id end)::numeric
46 / count(distinct(profile_id))*100.0 as return_rate
47 from
48 days_until_return
49 ``
50
51 {% enddocs %}
```

Glossary Terms > Adoption > ReturnRate

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'
)

, days_until_return as (
select
adoption_date.profile_id
, adoption_date.adoption_date
, return_date.return_date
, date_diff(adoption_date.adoption_date, return_date.return_date) as days_until_return
```

Metadata 360: Physical + Business Metadata

What is the authoritative pet dataset?



Physical Metadata is
not intuitive to
everyone



Acryl Data

Metadata 360: Physical + Business Metadata

What is the authoritative pet dataset?



Physical Metadata is not intuitive to everyone

Snowflake Table adoption

pet_profiles

Schema Documentation Properties Lineage Queries Stats Validation

Latest Historical

pet_profiles ✓

Schema Documentation Properties Lineage Queries Stats Validation

Assertions (2)

All assertions have passed
2 successful assertions, 0 failed assertions

Passed Dataset row count is greater than 20,000

Evaluations

Last evaluated on 3/22/2022 at 7:38:54 AM (America/Los_Angeles)

Mar 15 - Mar 22 2 passed 1 failed

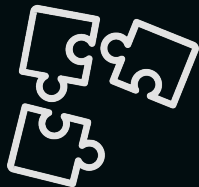
Row Count: 20500
3/20/2022, 7:38:54 AM

sex	unknown	unknown	unknown	unknown	0
spay_neutered	unknown	unknown	unknown	unknown	0
species	unknown	unknown	unknown	unknown	0

```
select
  pet_profiles.name
, popular_dog_names.dogname is not null as is_popular_dog_name
, sum(case when status = 'adopted' then 1 else 0 end) as cnt_adopted_dogs
, sum(case when status = 'ready for adoption' then 1 else 0 end) as cnt_adoptable_dogs
, count(*) as cnt_dogs
from
  adoption.pet_profiles
left join
```


Shift Left: Impact Analysis

Which critical dashboards will break if I change the `pet_profile` dataset?



Over-indexed on
Data Warehouses

Shift Left: Version Metadata as Code

The screenshot displays two Git commit history entries. The top entry is for the commit 'remove unneeded files' by shirshanka, with commit hash 33e3f9d and a timestamp of 2 days ago. Below this commit, a list of files is shown: a directory named 'common' (adding protobuf files), a directory named 'ecommerce' (adding protobuf files), and a directory named 'protobuf/meta' (remove unneeded files). The bottom entry is for the commit 'adding protobuf files' by shirshanka, with commit hash 28fa7cd and a timestamp of 2 days ago. Below this commit, a list of files is shown: 'ClickEvent.proto', 'ImpressionEvent.proto', and 'SearchEvent.proto', all of which are 'adding protobuf files'.

Commit Hash	Author	Commit Message	Timestamp
33e3f9d	shirshanka	remove unneeded files	2 days ago
28fa7cd	shirshanka	adding protobuf files	2 days ago

Shift Left: Schema Annotations

```
1 syntax = "proto3";
2 package common;
3 import "protobuf/meta/meta.proto";
4
5
6 /**
7  | Context attached to all events
8  */
9 message EventContext {
10
11  | // the device specific identifier
12  | string device_id = 1;
13  | // the device type associated with this event
14  | string device_type = 2;
15  | // the user id associated with this event
16  | string user_id = 3;
17  | // the UTC timestamp associated with this event, in mi
18  | int32 timestamp=4;
19  | // the ip address of the client
20  | string ip_address=5
21  | [(meta.fld.classification) = "Classification.Sensitive"]
22  |
23  | }
24
25
26
```

```
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  | }
```



Shift Left: Schema Annotations

```
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 }
```

Datasets > prod > kafka > ecommerce > searchevent

Kafka Dataset

Queries Stats Validation

Description Tags Terms

the device specific identifier

the device type associated with this event

the user id associated with this event

the UTC timestamp associated with this event, in milliseconds since epoch

the ip address of the client

search_id (Number)

search_term (String)

> result_array[] (Array)

Details Lineage 0 upstream, 0 downstream

About

The event emitted whenever a Search is executed

+ Add Link

Tags

team.Ecommerce X + Add Tag

Glossary Terms

Sensitive X + Add Term

Owners

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

+ Add Owner

Domain

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

Set Domain

Sensitive

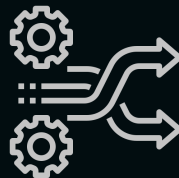
Streaming Metadata: Monitor and React to Changes

Can I rely on the purchase dataset to be up to date and accurate?



Crawl-Only
Ingestion Leads to
Stale Metadata

How can I ensure my ML features exclude PII?



No approach to
acting on changes
in metadata

But First, How is DataHub Architected?



Generation 1

Crawl-based ingestion

Opinionated inflexible metadata model



Generation 2

Single Microservice with DB

Opinionated inflexible metadata model or completely generic metadata model



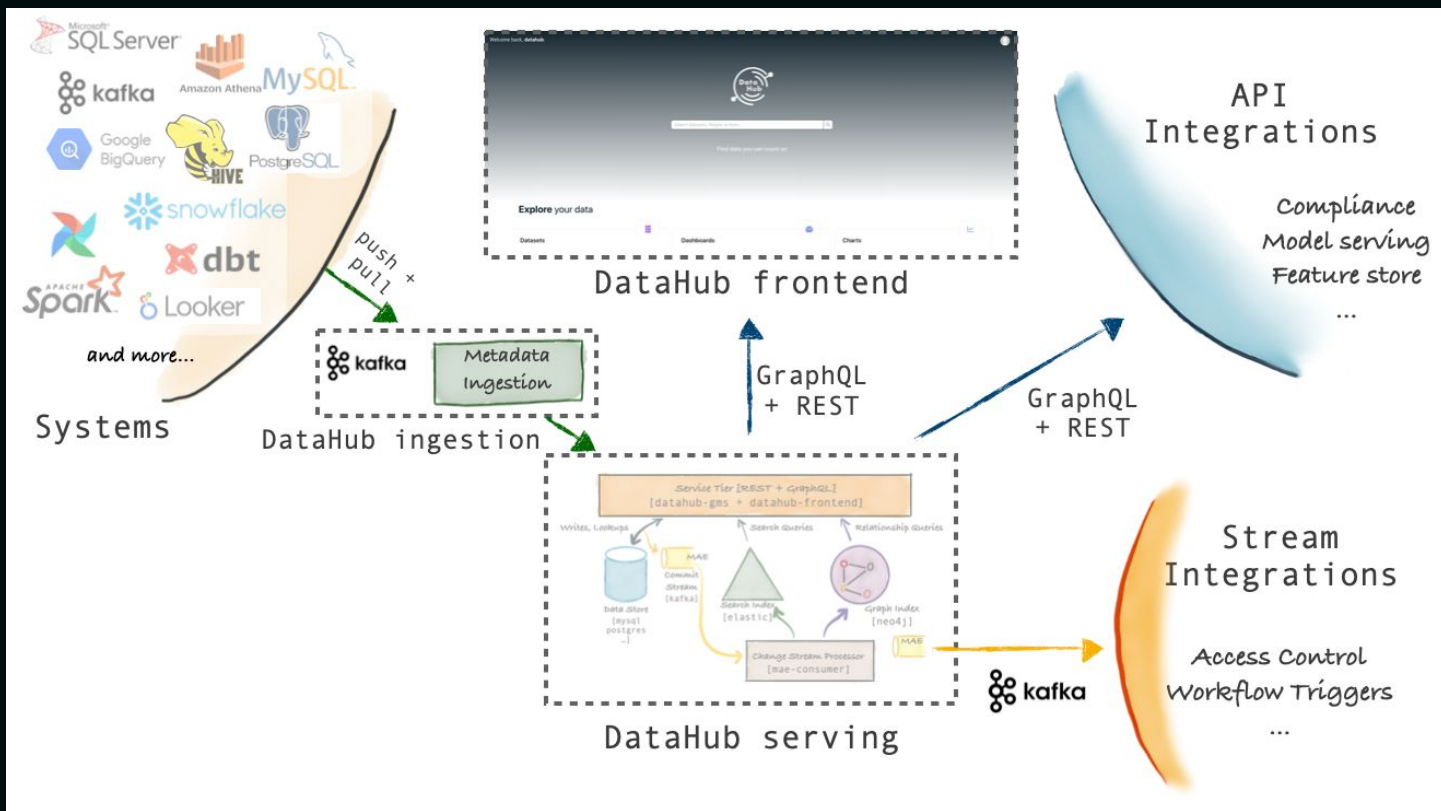
Generation 3

Stream-oriented loosely coupled architecture

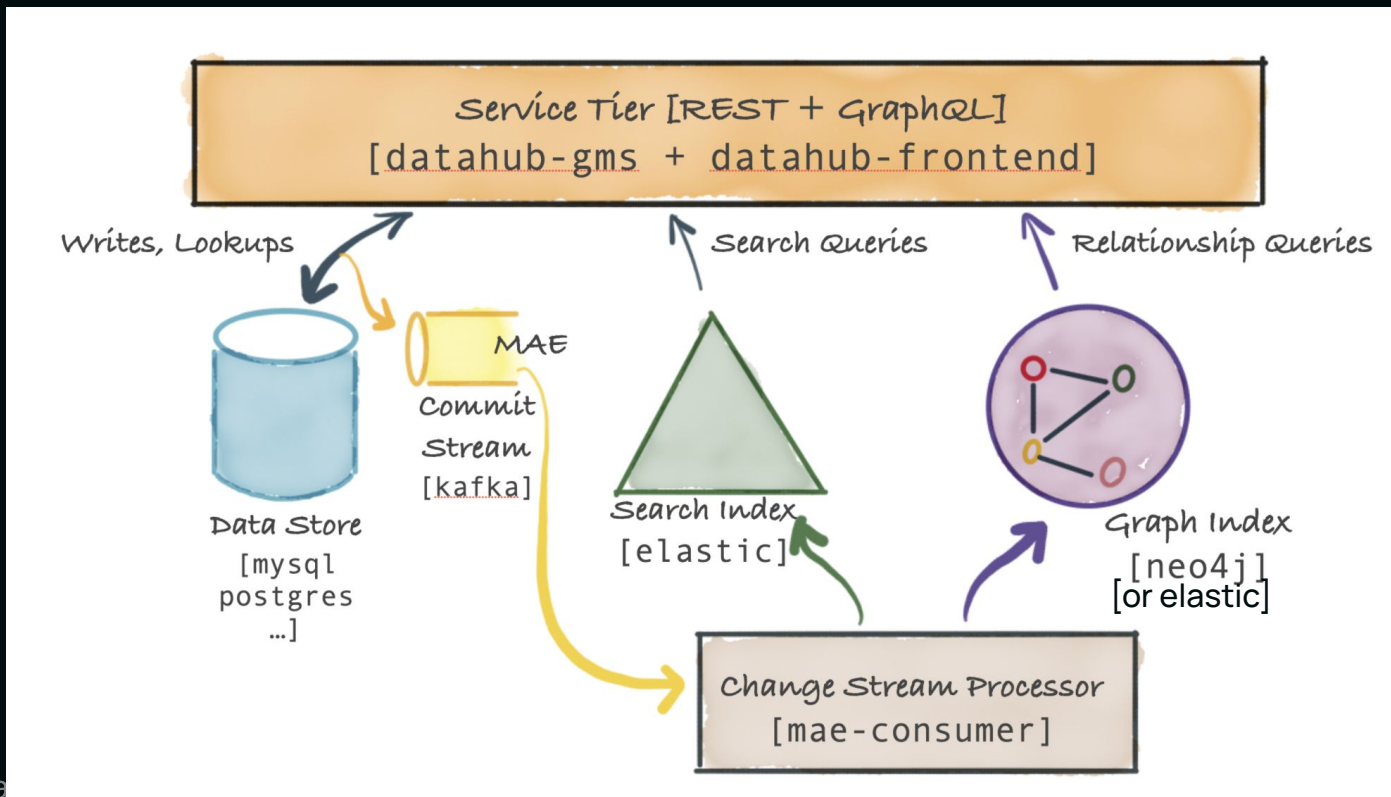
Push-based with easy integration for crawlers

Extensible metadata model with useful core model

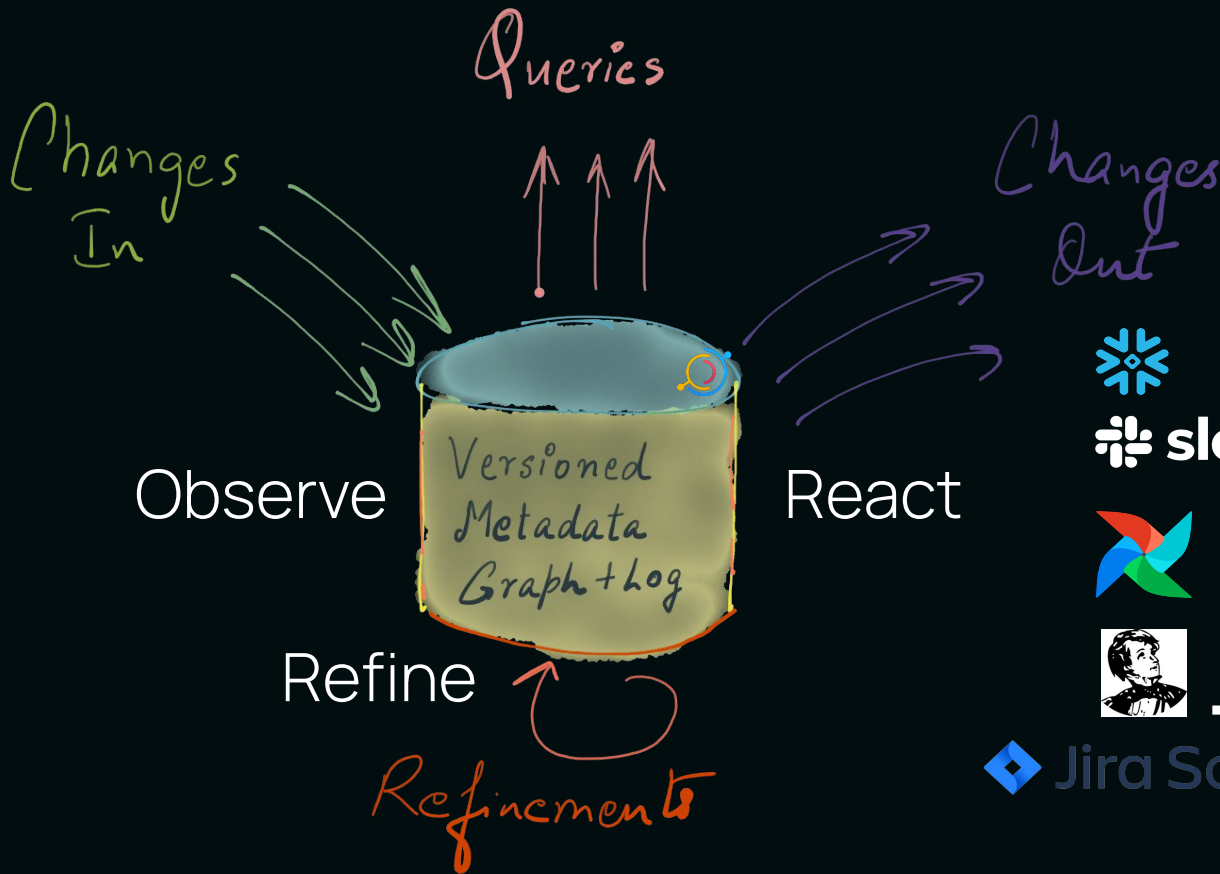
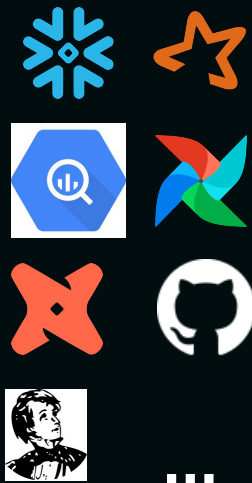
DataHub Architecture



DataHub Storage Architecture



Streaming Metadata enables the Control Plane



Discovery is Passive – Streaming is Active

Tag propagation

- Automatically tag/mask datasets

Impact Analysis + Timeline API

- All of the tools you have at your disposal to look back in time
- Eg. push a breaking change, see a version change via timeline API; fetch all downstream entities & alert owners

Streaming Metadata: Time-Travel for Schema History

Can I rely on the purchase dataset to be up to date and accurate?



Crawl-Only
Ingestion Leads to
Stale Metadata

```
~/workspace/long-tail-companions/ecommerce/kafka/protobuf-schemas > main +4 |12 ?12 datahub |
```

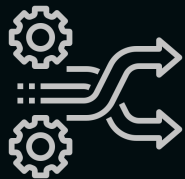
Streaming Metadata: Time-Travel for Schema History

```
> datahub timeline --urn "urn:li:dataset:(urn:li:dataPlatform:kafka,ecommerce.SearchEvent,PROD)" -c technical_schema -c tag
http://localhost:8080/openapi/timeline/v1/urn%3Ali%3Adataset%3AX28urn%3Ali%3AdataPlatform%3Akafka%2Cecommerce.SearchEvent%2CPROD%29?categories=TECHNICAL_SCHEMA,TAG
2022-03-20 22:54:37 - 0.0.0-computed
  ADD TAG dataset:kafka:ecommerce.SearchEvent (urn:li:tag:team.Ecommerce): Tag 'team.Ecommerce' added to entity 'urn:li:dataset:(urn:li:dataPlatform:kafka,ecommerce.SearchEvent,PROD)'.
  ADD TECHNICAL_SCHEMA dataset:kafka:ecommerce.SearchEvent (field:result_array): A forwards & backwards compatible change due to the newly added field 'result_array'.
  ADD TECHNICAL_SCHEMA dataset:kafka:ecommerce.SearchEvent (field:result_array.item_id): A forwards & backwards compatible change due to the newly added field 'result_array.item_id'.
  ADD TECHNICAL_SCHEMA dataset:kafka:ecommerce.SearchEvent (field:context): A forwards & backwards compatible change due to the newly added field 'context'.
  ADD TECHNICAL_SCHEMA dataset:kafka:ecommerce.SearchEvent (field:context.timestamp): A forwards & backwards compatible change due to the newly added field 'context.timestamp'.
  ADD TECHNICAL_SCHEMA dataset:kafka:ecommerce.SearchEvent (field:context.device_id): A forwards & backwards compatible change due to the newly added field 'context.device_id'.
  ADD TECHNICAL_SCHEMA dataset:kafka:ecommerce.SearchEvent (field:context.device_type): A forwards & backwards compatible change due to the newly added field 'context.device_type'.
  ADD TECHNICAL_SCHEMA dataset:kafka:ecommerce.SearchEvent (field:context.user_id): A forwards & backwards compatible change due to the newly added field 'context.user_id'.
  ADD TECHNICAL_SCHEMA dataset:kafka:ecommerce.SearchEvent (field:search_id): A forwards & backwards compatible change due to the newly added field 'search_id'.
  ADD TECHNICAL_SCHEMA dataset:kafka:ecommerce.SearchEvent (field:search_term): A forwards & backwards compatible change due to the newly added field 'search_term'.
2022-03-20 22:56:30 - 0.1.0-computed
  ADD TECHNICAL_SCHEMA dataset:kafka:ecommerce.SearchEvent (field:context.ip_address): A forwards & backwards compatible change due to the newly added field 'context.ip_address'.
2022-03-20 22:58:19 - 1.0.0-computed
  REMOVE TECHNICAL_SCHEMA dataset:kafka:ecommerce.SearchEvent (field:search_term): A backwards incompatible change due to removal of field: 'search_term'.
2022-03-20 23:02:59 - 1.1.0-computed
  ADD TAG dataset:kafka:ecommerce.SearchEvent (urn:li:tag:pii): Tag 'pii' added to entity 'urn:li:dataset:(urn:li:dataPlatform:kafka,ecommerce.SearchEvent,PROD)'.
~/workspace/long-tail-companions/ecommerce/kafka/protobuf-schemas > main +4 !12 ?12 data_council Py
>
```



Streaming Metadata: Tag Propagation

How can I ensure
my ML features
exclude PII?



No approach to
acting on changes
in metadata

3 Must-Haves for Sustainable Data Discovery

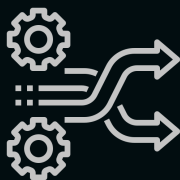
1. Metadata 360



Physical Metadata is not intuitive to everyone



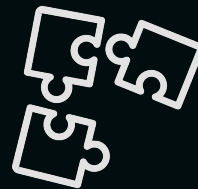
Crawl-Only Ingestion Leads to Stale Metadata



No approach to acting on changes in metadata



Manual enrichment of metadata leads to problems



Over-indexed on Data Warehouses

Join the Movement

slack.datahubproject.io

github.com/datahub-project/datahub

@datahubproject

```
> pip install acryl-datahub
```

```
> datahub docker quickstart
```

Acryl Data is Hiring!

CAREERS

Join Our Team

Join us in bringing clarity to data by enabling delightful search and discovery, data observability, and federated governance across data ecosystems.

Culture

At Acryl Data, collaboration is key, curiosity inspires action, and ambition and empathy is our (not so) secret sauce.

Values

We are a community-first, impact-driven team committed to representing the lived experiences, unique perspectives, and communities around us.



Acryl Data