

Amundsen: From Discovering to Securing Data

Mark Grover | Lyft | [@mark_grover](https://twitter.com/mark_grover)

Alyssa Ransbury | Square | [@alyssaran](https://twitter.com/alyssaran)

Slides: cutt.ly/a6n



Who we are



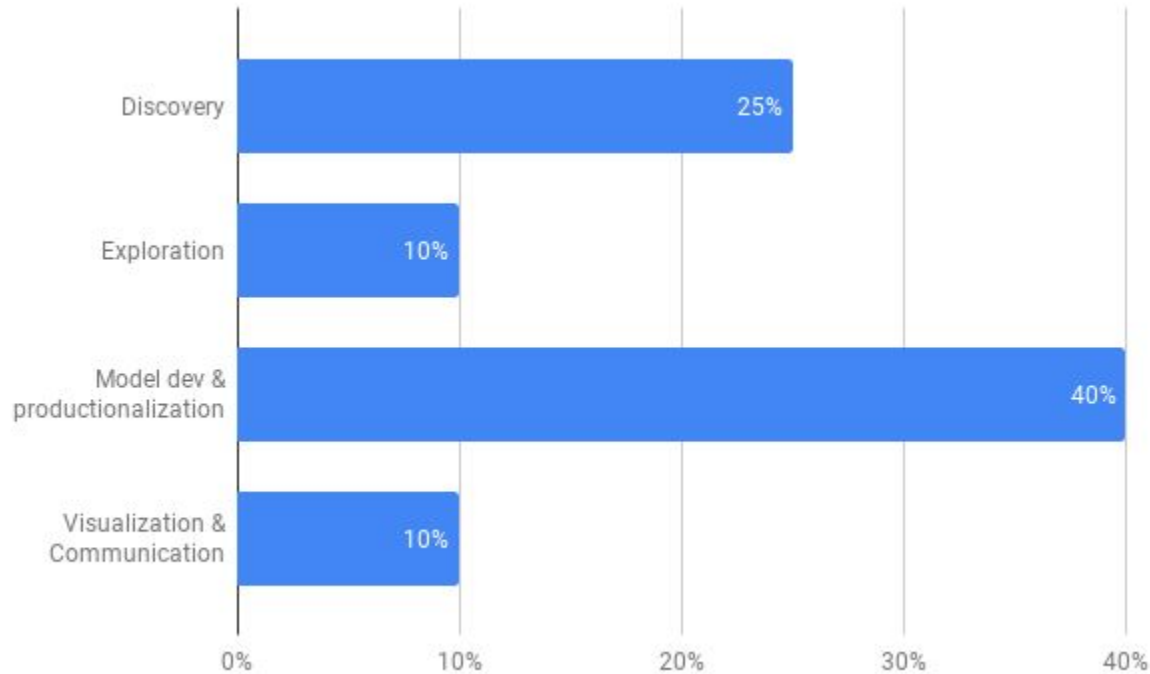
- Product Manager at Lyft
- Committer/PMC on various Apache projects
- Previously developer on Spark at Cloudera



- Security Engineer at Square
- Leads and supports on privacy engineering efforts across dozens of product teams

Problem

>35% consumer time spent on discovering & validating trustworthy data



Analyst/DS workflow and time spent on each step

Other side effects

1. Interrupt heavy culture

- What data should I use for X?
- Is that trustworthy?

2. Increased DB load

```
SELECT * ... LIMIT 100  
SELECT COUNT(*) over x days
```

Requirements & solutions

1. What kind of information? (aka *ABC of metadata*)

Application Context

Metadata needed by humans or applications to operate

- Where is the data?
- What are the semantics of the data?

Behavior

How is data created and used over time?

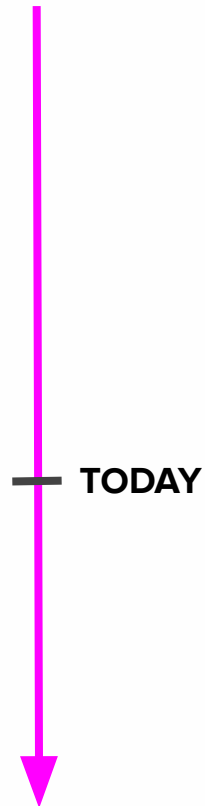
- Who's using the data?
- Who created the data?

} Data Lineage

Change

Change in data over time

- How is the data evolving over time?
- Evolution of code that generates the data



Terminology borrowed from [Ground](#) paper

2. About what data?

Short answer: Any data within your organization

Long answer:

Data stores



PostgreSQL



People



Employees

**Dashboard /
Reports**



looker

Processes



Notebooks



**Events /
Schemas**

Schema registry

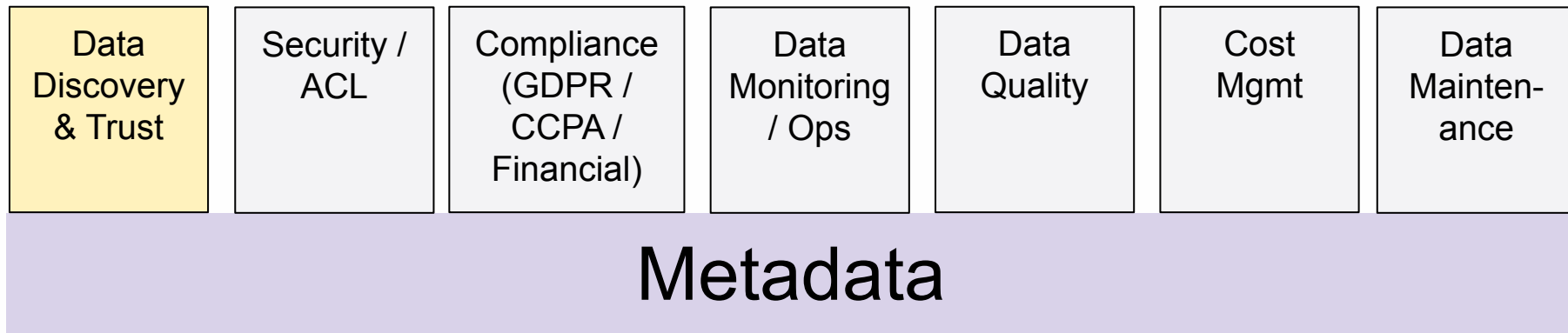
Segment

Streams


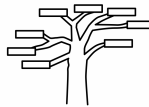




TODAY

Develop breadth of applications



Goal: Reduce time to find trusted data w/ versatile graph

Search based 	Lineage based 	Network based 	Programmatic 
<p>Where is the table/dashboard for X? What does it contain?</p>	<p>I am changing a data model, who are the owner and most common users?</p>	<p>I want to follow a power user in my team.</p>	<p>Access metadata programmatically</p>
<p>Does this analysis already exist?</p>	<p>This table's delivery was delayed today, I want to notify everyone downstream.</p>	<p>I want to bookmark tables of interest and get a feed of data delay, schema change, incidents.</p>	<p>Put (pull / push) metadata programmatically</p>

Other requirements

- Leverage as much data automatically as possible
- Preferably, open source and healthy community
- Easy to set up

Meet Amundsen

First person to discover the South Pole -
Norwegian explorer, Roald Amundsen

Search for data, or browse

AMUNDSEN

Announcements Browse



🔍 search for data resources...

Search within a category using the pattern with wildcard support 'category:*searchTerm*', e.g. 'schema:*core*'.
Current categories are 'column', 'database', 'schema', 'table', and 'tag'.

Browse Tags

tag1 1 tag2 1

My Bookmarks



test_schema.test_table2 ⭐

2nd test table

dynamo

Popular Tables ⓘ



test_schema.test_table1 ☆

1st test table

Hive

Search for datasets

AMUNDSEN

Announcements

Browse



🔍 **table**

🔍 **table** in Datasets

Datasets



test_schema.test_table2

2nd test table

dynamo



test_schema.test_table1




1st test table

Hive

[See all 2 Datasets results](#)


See details of the data set

AMUNDSEN Announcements Browse



<  **test_schema.test_table1** ☆
Datasets • Hive • gold high quality pii  Airflow  github Preview

Description
1st test table

Date Range
From: Apr 22, 2017
To: Sep 30, 2019

Frequent Users


Tags
tag1 tag2

Owners
 chris@example.org
 roald.amundsen@example.org

col1 col1 description	string
col2 col2 description	string
col3 col3 description	string
col4 col4 description	string
col5 col5 description	float

See detailed descriptions and profile of the column

col1

string

Description

This is an editable test description for the first column. This also supports **Markdown**.

***Column Statistics** Stats reflect data collected between May 22, 2015 and Jul 04, 2019.*

distinct values	8
-----------------	---

min	aardvark
-----	----------

num nulls	500320
-----------	--------

max	zebra
-----	-------

verified	230430
----------	--------

Search for existing dashboards (aka reports)

lyft AMUNDSEN

amundsen

Announcements Browse FAQ ? MG



Resource

- Datasets 60
- Dashboards 2
- People 0

Groups ⓘ

Name ⓘ


Tag ⓘ

 DPE amundsen_dashboard_table_lineage ☆	Mode	Last Successful Run Jun 12, 2020
 Global Ops Analytics - Scratchpad Clone of Amundsen Search Demystified ☆ Cloned copy of the report linked to in https://confluence.lyft.net/display/DATA/Amundsen+Search+Tutorial as-of 5/...	Mode	Last Successful Run May 25, 2020 >

Amundsen was last indexed on June 23rd 2020 at 5:30:49 pm


https://amundsen.lyft.net/dashboard/mode_dashboard%3A%2F%2Fgold.bc0496fe0072%2Fe52d535ae61b?index=1&source=search_results

Dashboard detail page

 AMUNDSEN Announcements Browse FAQ MG

[amundsen_dashboard_table_lineage](#) ☆
Dashboard in [DPE](#) Open Dashboard

Description
[Add Description in Mode](#)

Owners
 Tao Feng

Created
Jun 01, 2020 11pm PDT

Last Updated
Jun 12, 2020 5pm PDT






Recent View Count
1

Tags
[+ New](#)

Last Successful Run
Jun 12, 2020 5pm PDT

Last Run
Jun 12, 2020 5pm PDT
Succeeded

Tables (5) Queries (4)

-  [hivemetastore.partitions](#) ☆
Imported by sqoop on 2019/10/01 00:18:51 Hive
-  [events.event_hive_query_logged](#) ☆
This event fires when an hive query is created and another one when it is complet... Hive
-  [hivemetastore.dbs](#) ☆
Imported by sqoop on 2019/10/01 00:31:07 Hive
-  [hivemetastore.tbls](#) ☆ Hive
-  [default.event_security_audit](#) ☆
The event that is emitted when logging a security audit event Hive

Query 1

id	start_time	end_time	status	error_message	query_text
1	2020-06-01 21:03:47		failed	error: permission denied to view table	...
2	2020-06-01 21:03:49		failed	error: permission denied to view table	...
3	2020-06-01 21:03:50		failed	error: permission denied to view table	...
4	2020-06-01 21:03:51		failed	error: permission denied to view table	...
5	2020-06-01 21:03:52		failed	error: permission denied to view table	...
6	2020-06-01 19:23:09		failed	error: permission denied to view table	...
7	2020-06-01 19:23:10		failed	error: permission denied to view table	...

Amundsen was last indexed on June 23rd 2020 at 5:30:49 pm

Search for co-workers!


lyft AMUNDSEN

Q mark grover X

Announcements Browse FAQ ? MG

Resource

- Datasets 65
- Dashboards 2
- People 1

 **Mark Grover**
Product Manager • Data Tools & Productivity

User

Search for data owned and used by your peers!


lyft AMUNDSEN Announcements Browse FAQ ? MG

[MG](#) **Mark Grover**
Product Manager • Data Tools & Productivity • Manager: Matt Isanuk



[mgrover@lyft.com](#) [Employee Profile](#) [Github](#)

[Datasets \(56\)](#) [Dashboards \(0\)](#)


Owned (1)

 default.dummy ★	Hive
---	------

Bookmarked (2)

 default.dummy ★	Hive
 default.event_helloworld_hello_world ★ Helloworld - Helloworld Event for Eventingest Testing	Hive

Frequently Used (53)

 base.db_query_usage_metrics ☆	Hive
---	------

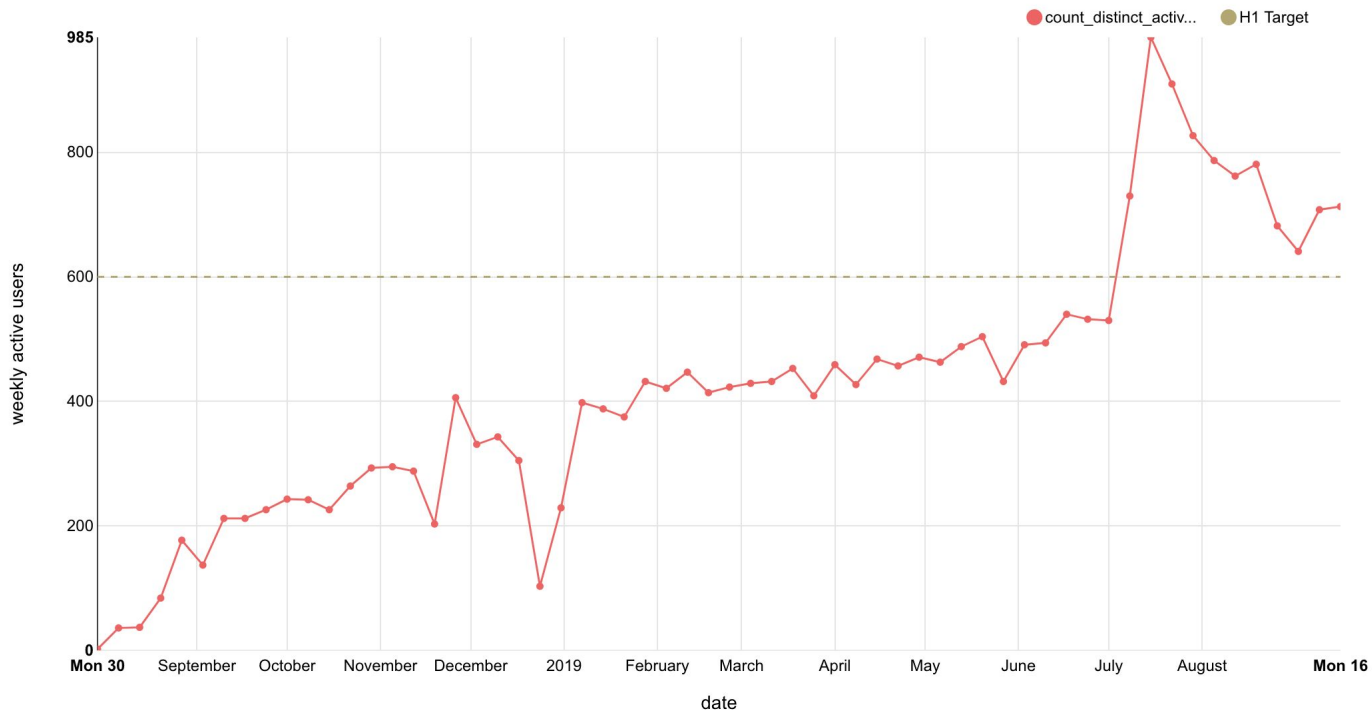
Amundsen was last indexed on June 23rd 2020 at 5:30:49 pm

Impact

A6n @ Lyft: 750+ WAUs, 150k+ tables, 4k+ employee pages

Weekly active users ☆ Altered s

60 rows 00:00:07.67



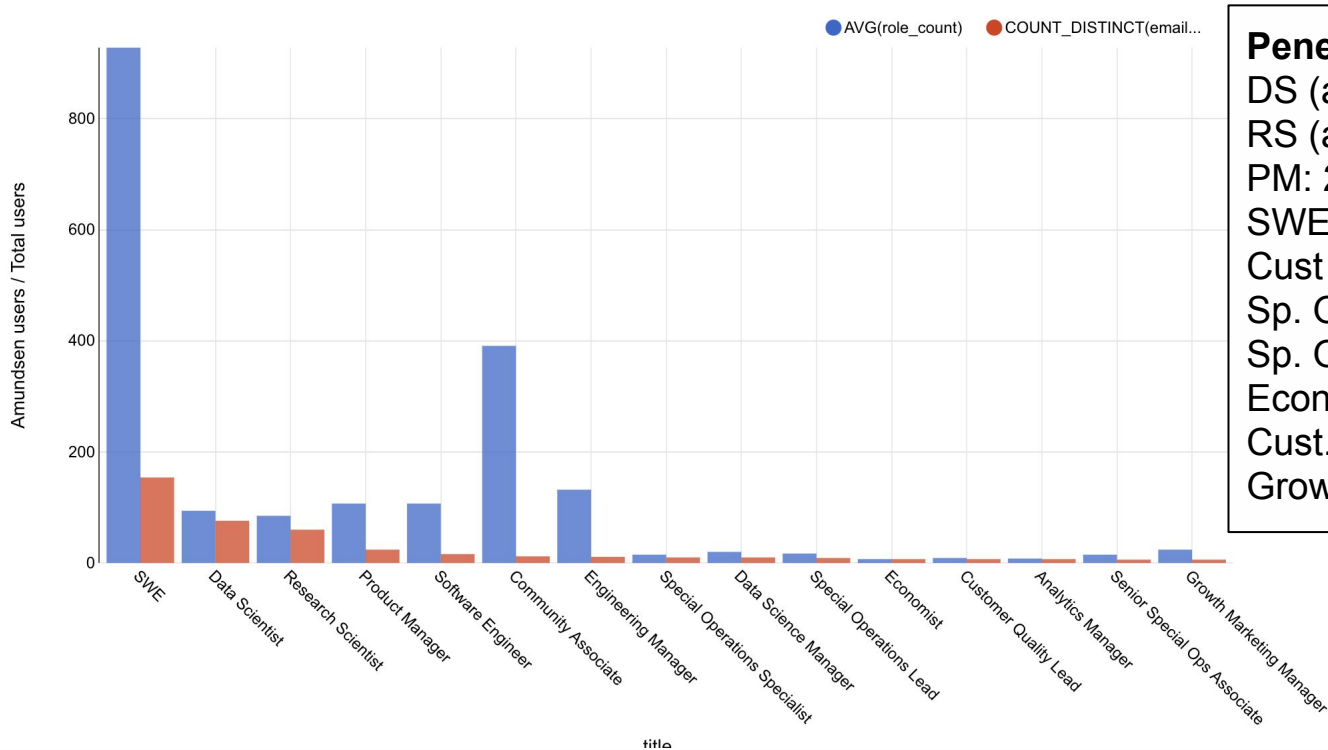
“This is God’s work” - Head of Analytics, Lyft

“I was on call and I’m confident 50% of the questions could have been answered by a simple search in Amundsen” - Data Scientist, Lyft

Roles of Amundsen users at Lyft

Weekly Amundsen user roles (Rolling 7 days) ☆ Altered

15 rows cached 00:00:00.15



Penetration rate:

DS (aka analyst): 81%

RS (aka DS): 71%

PM: 22%

SWE: 17%

Cust Serv: 7% (12/390)

Sp. Ops: 67% (10/15)

Sp. Op Leads: 53% (9/17)

Economist: 100% (7/7)

Cust. Quality: 78% (7/9)

Growth Mktg: 25% (6/24)

Roadmap



Roadmap (subject to change, not ordered)

- Tighter Lineage integration / visualization
- Better view integration
- ACL integration, allow only specific roles to edit descriptions
- Show search context for what matched
- Index more resources (notebooks, Kafka topics, etc.)

Amundsen Open Source

700

Community
members

150+

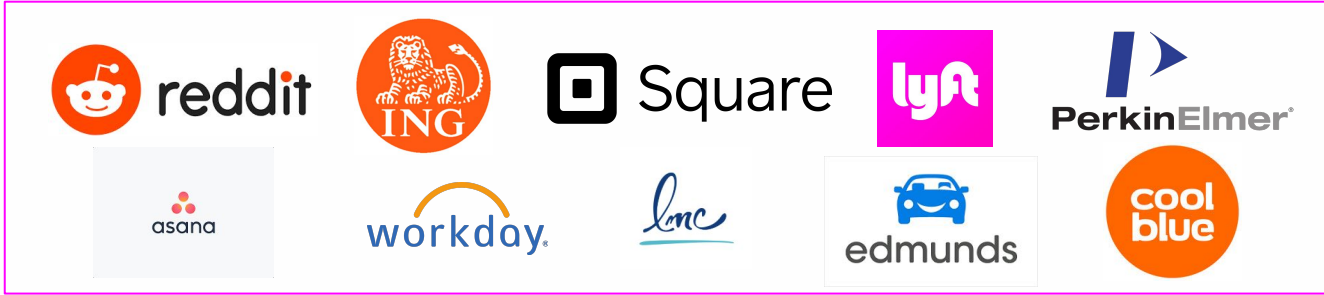
Companies in
the community

20+

Companies using
in production

Amundsen Open Source Community

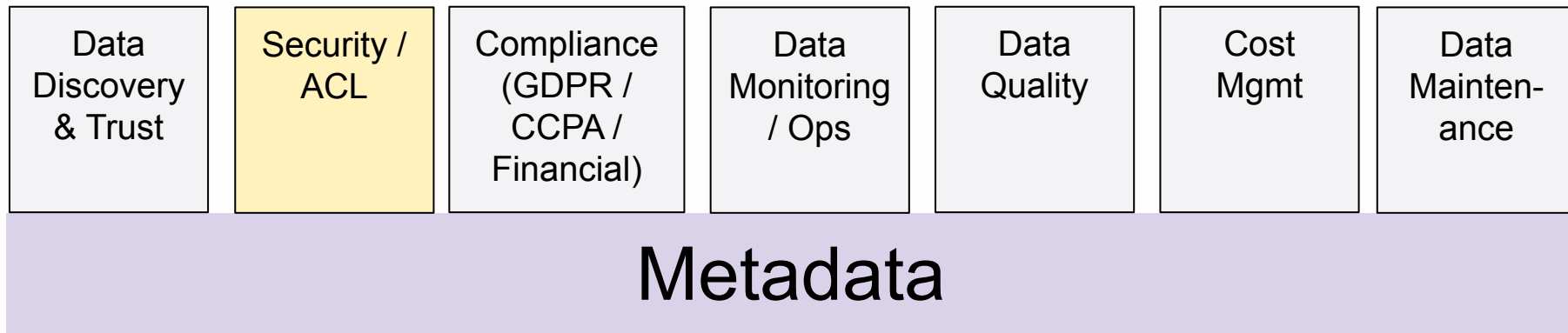
Prominent users



Active community

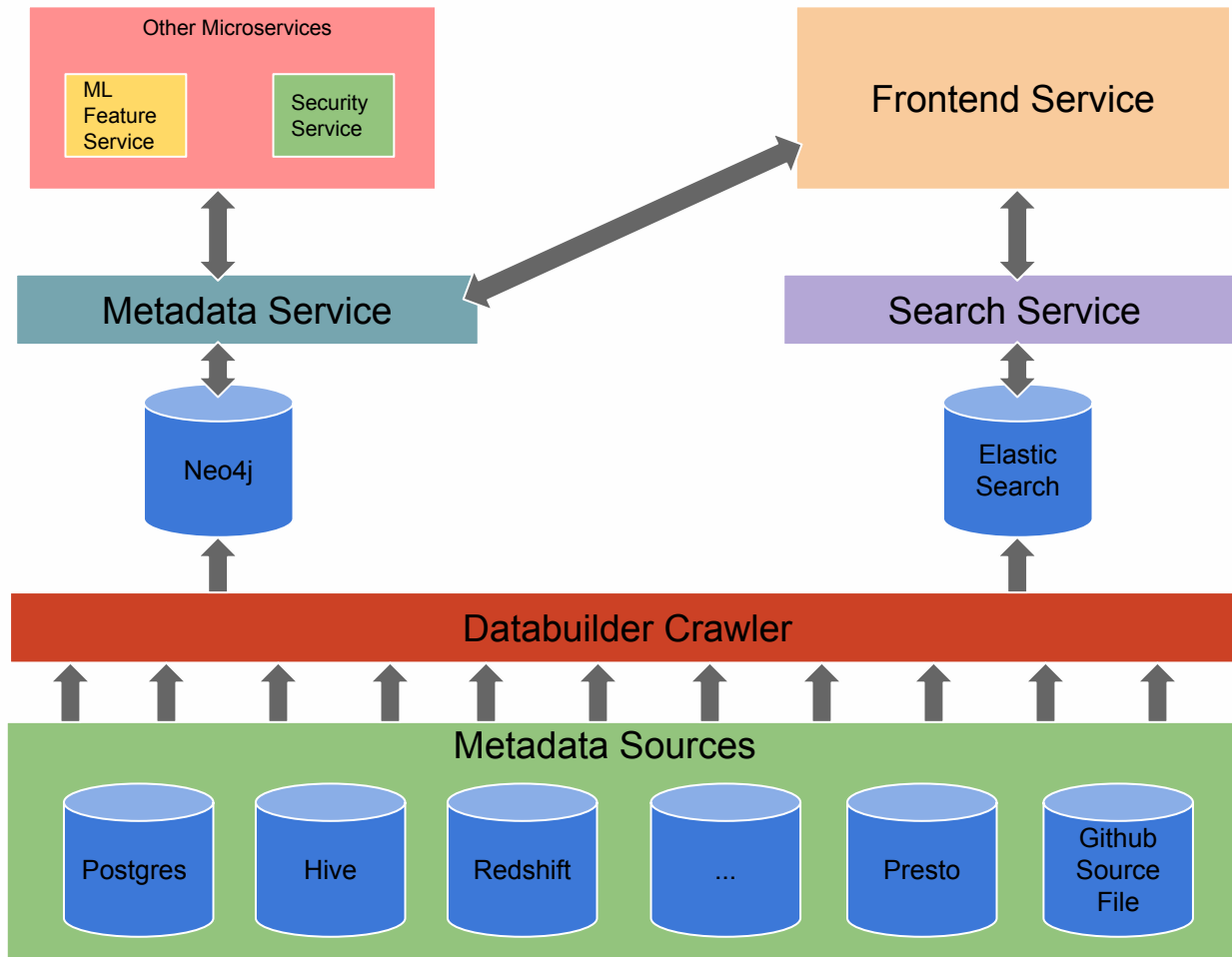


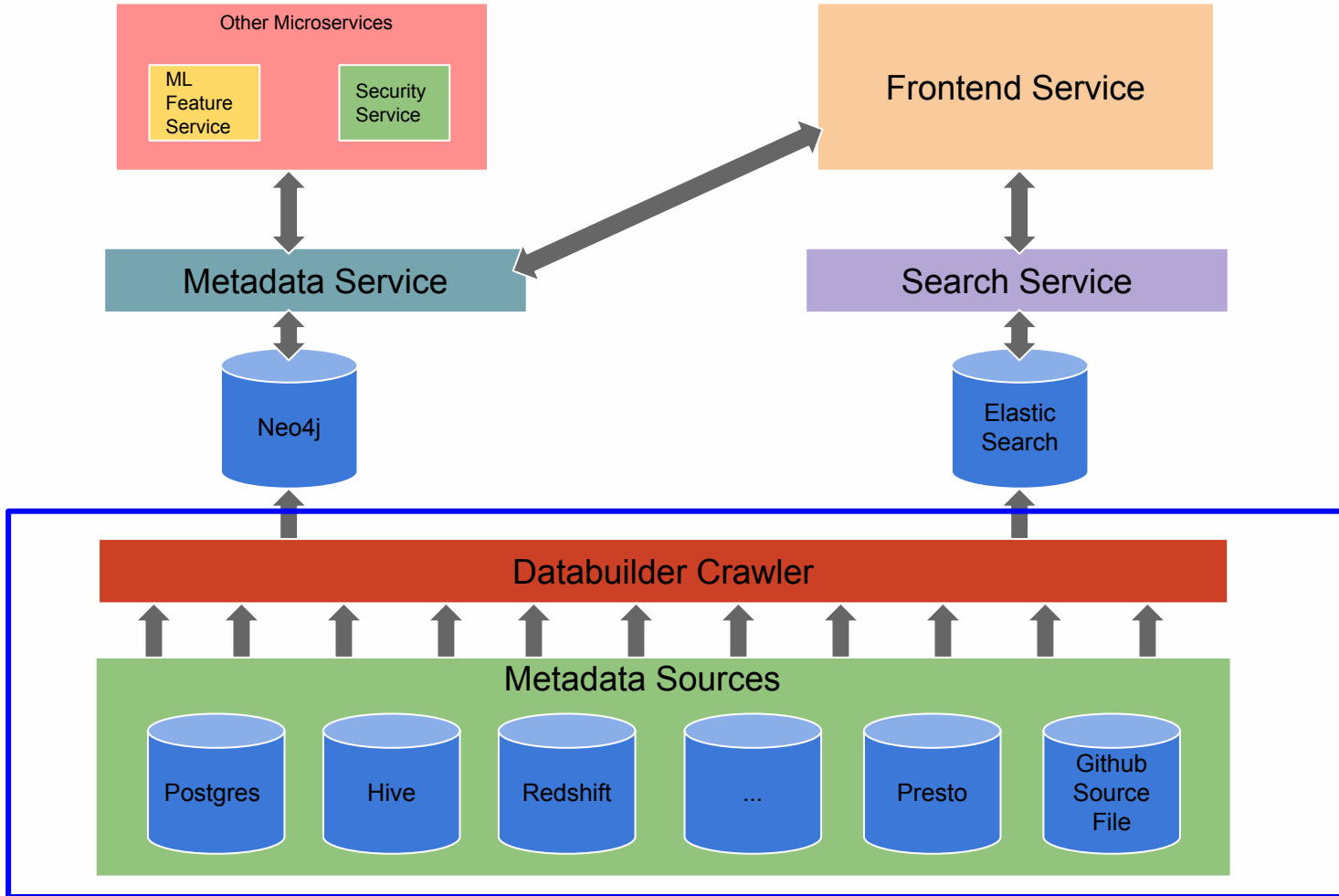
Develop breadth of applications

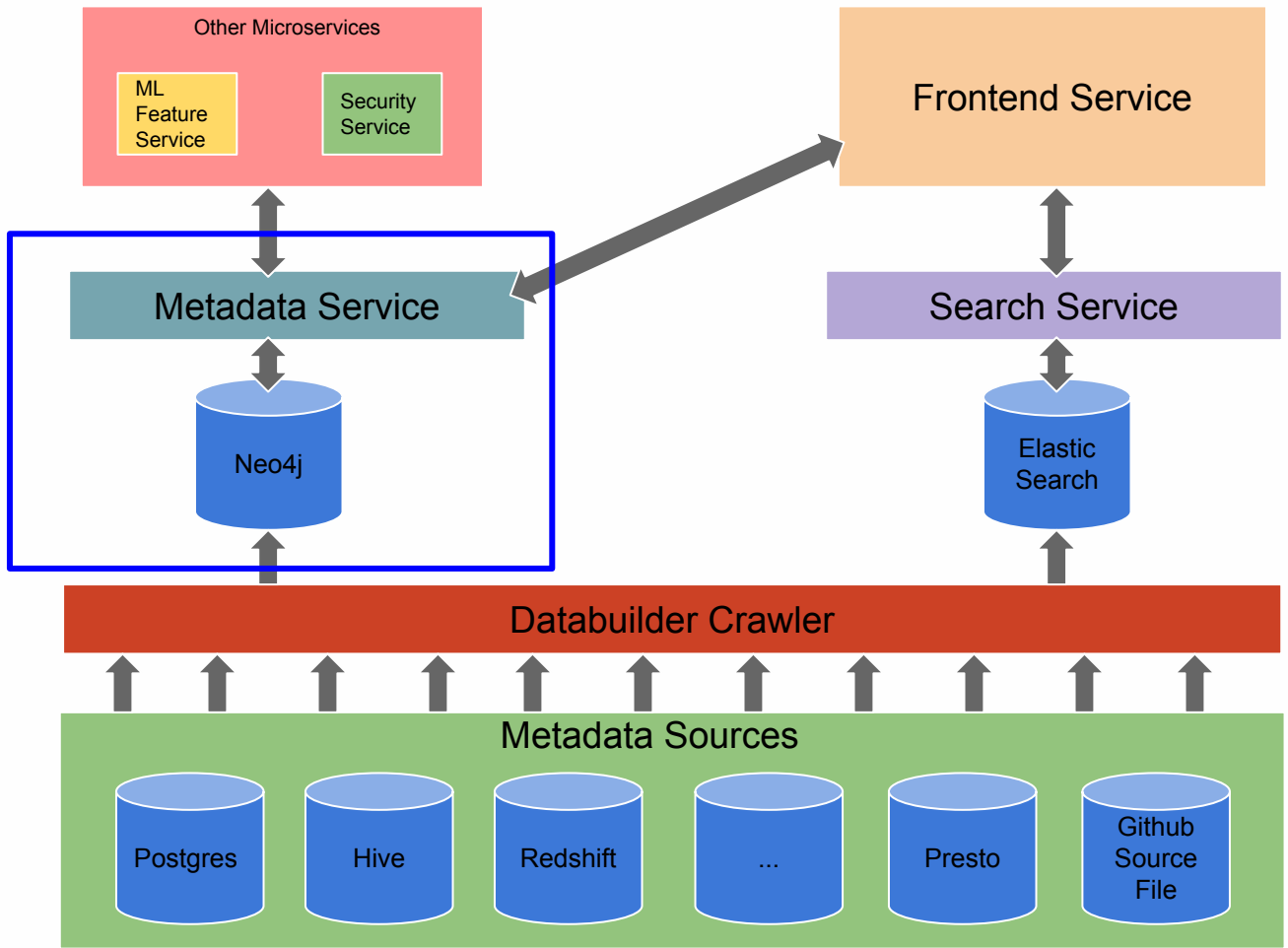


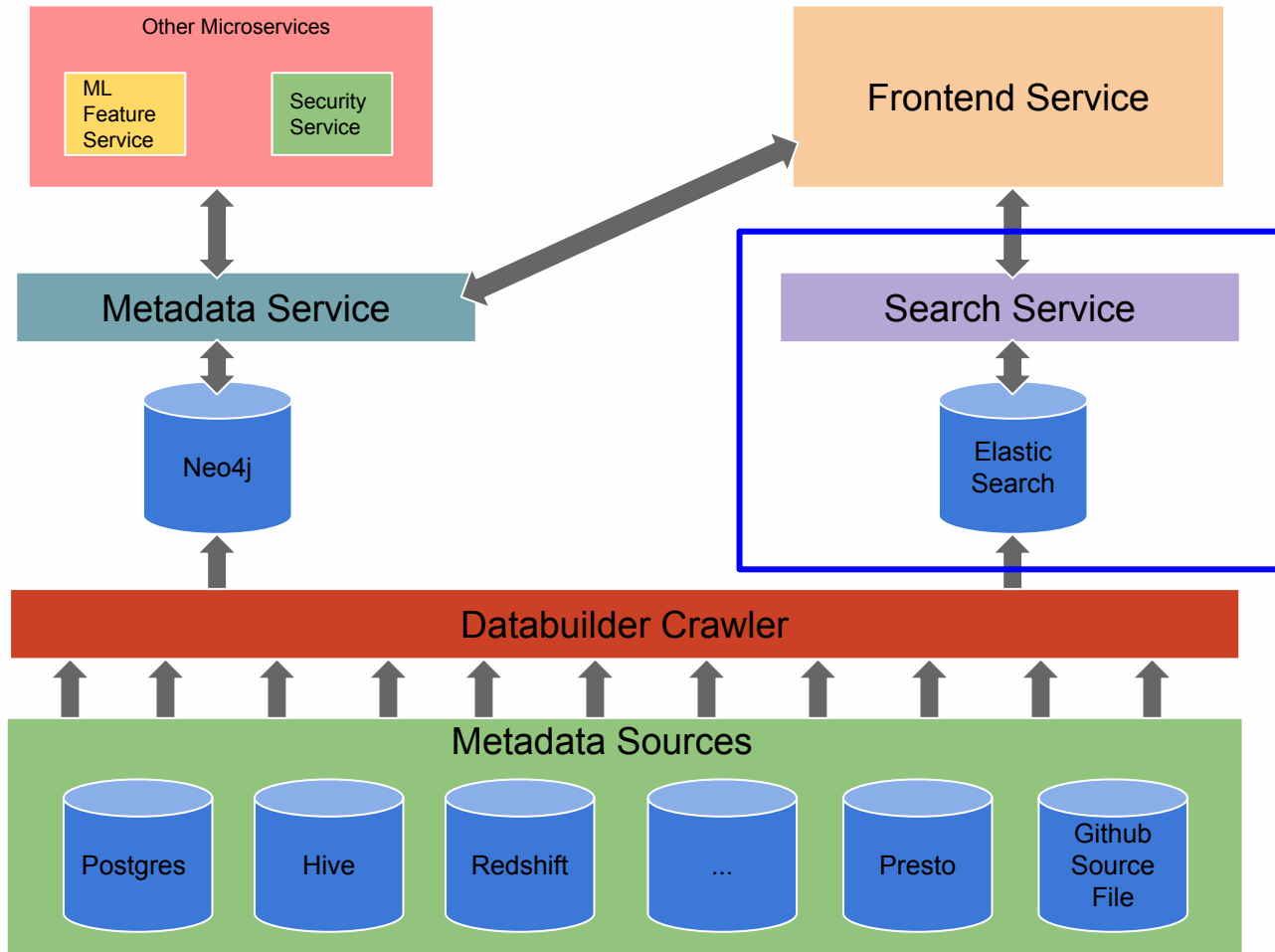
Square: Scaling Data Security / Privacy

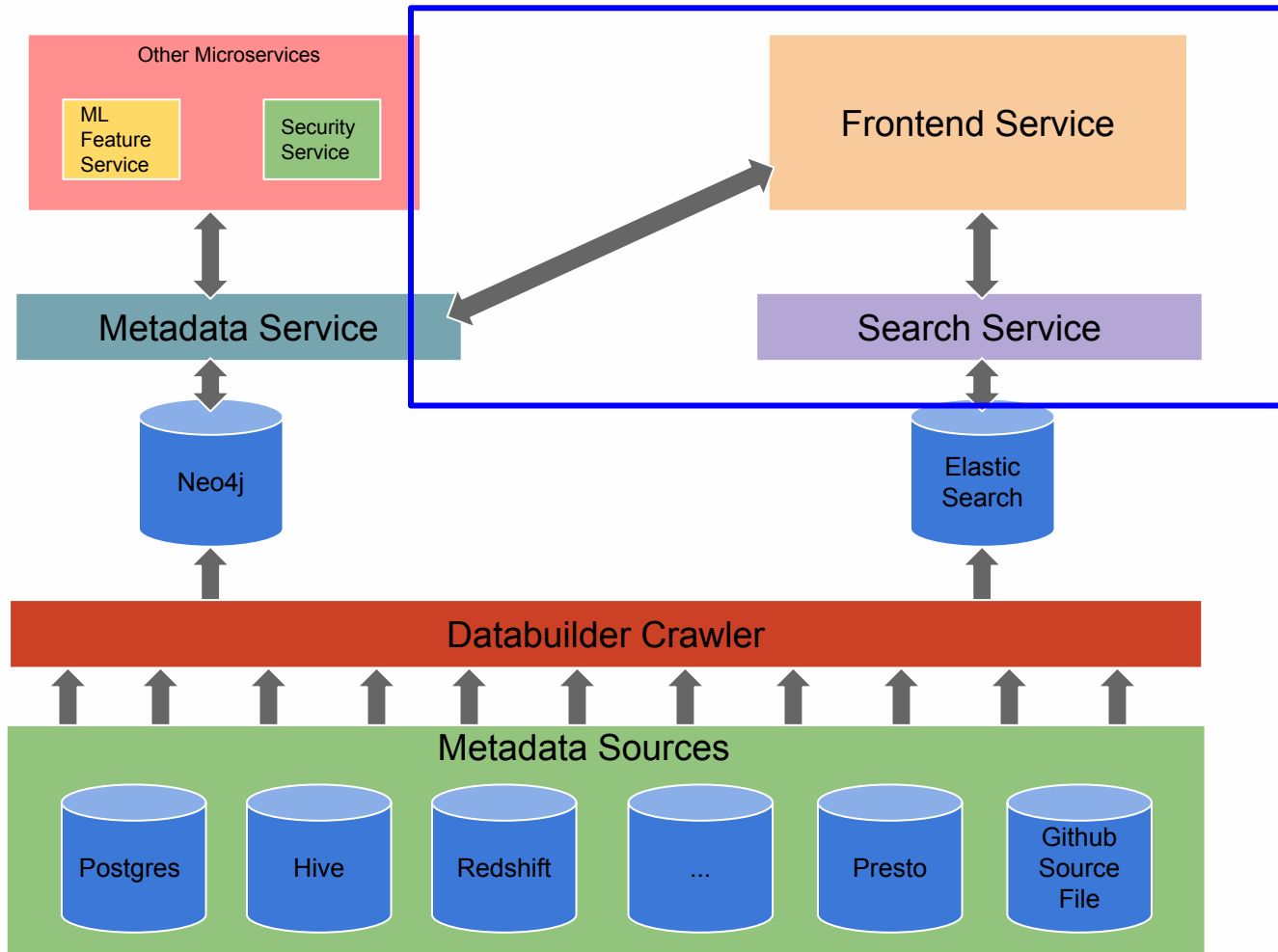




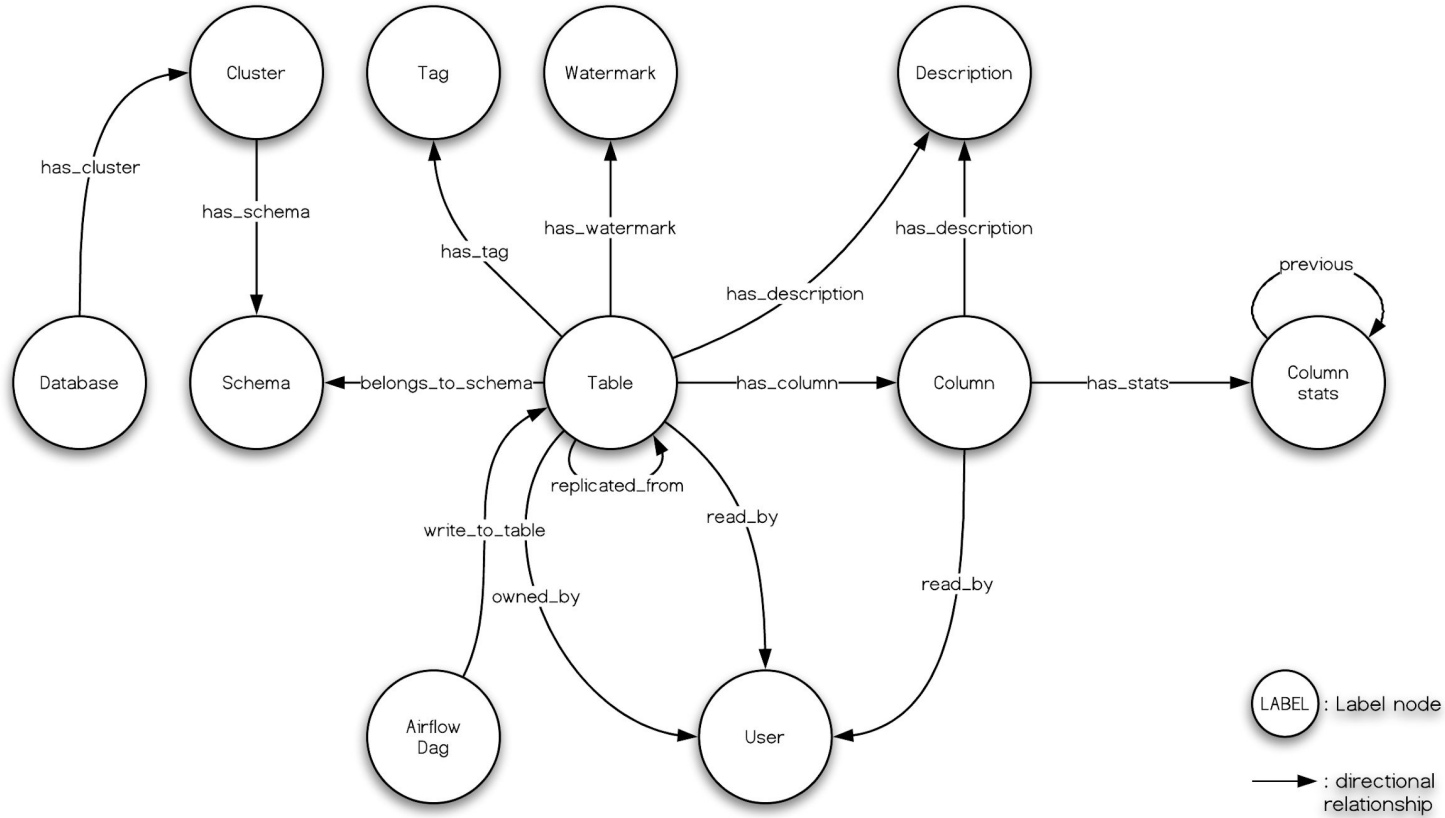




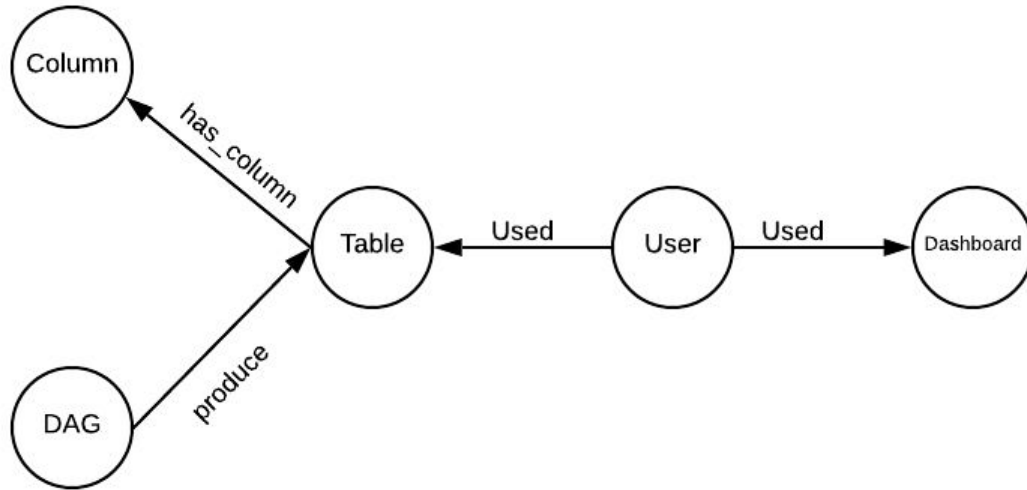




The graph

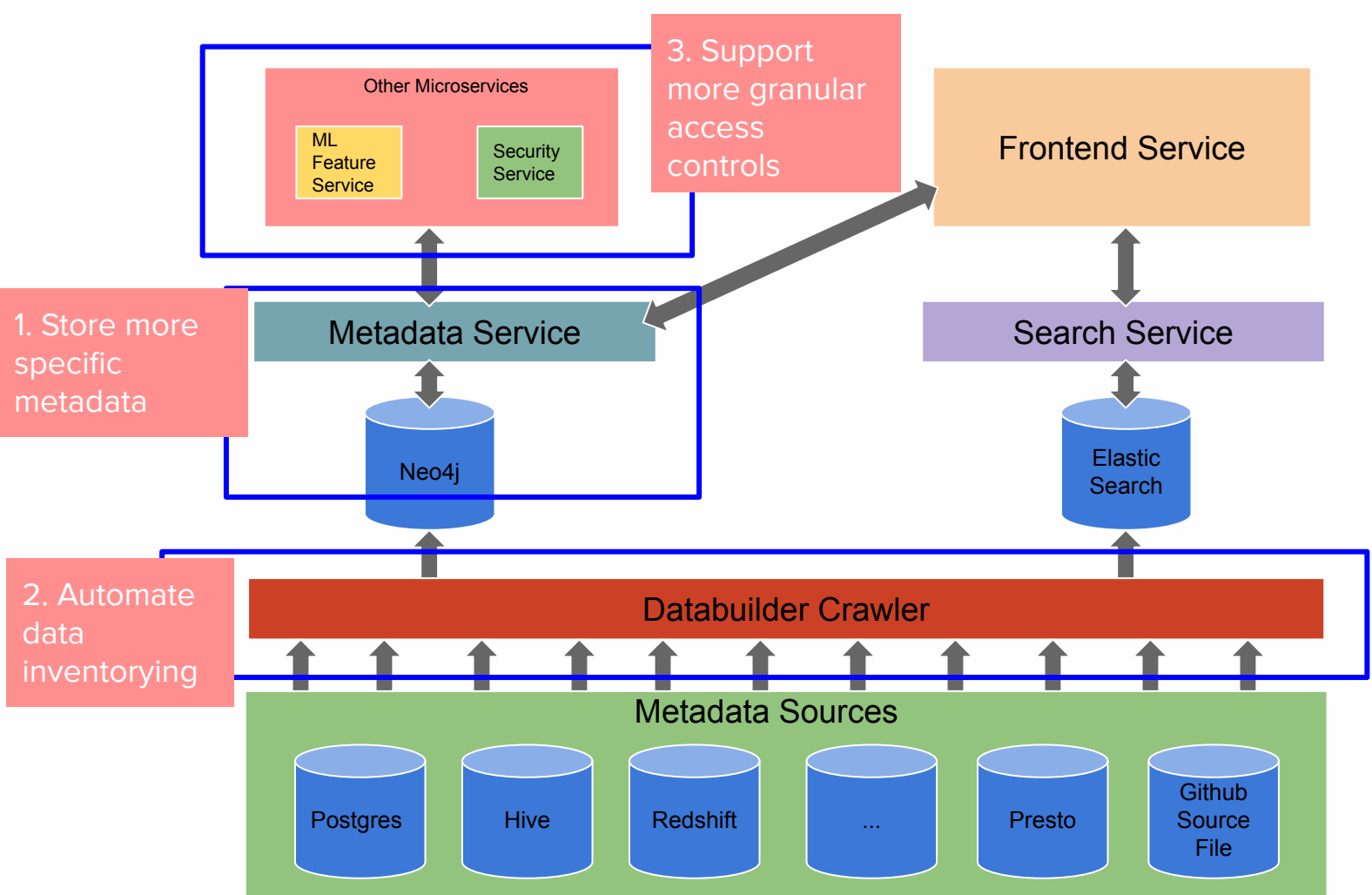


Benefits of a graph

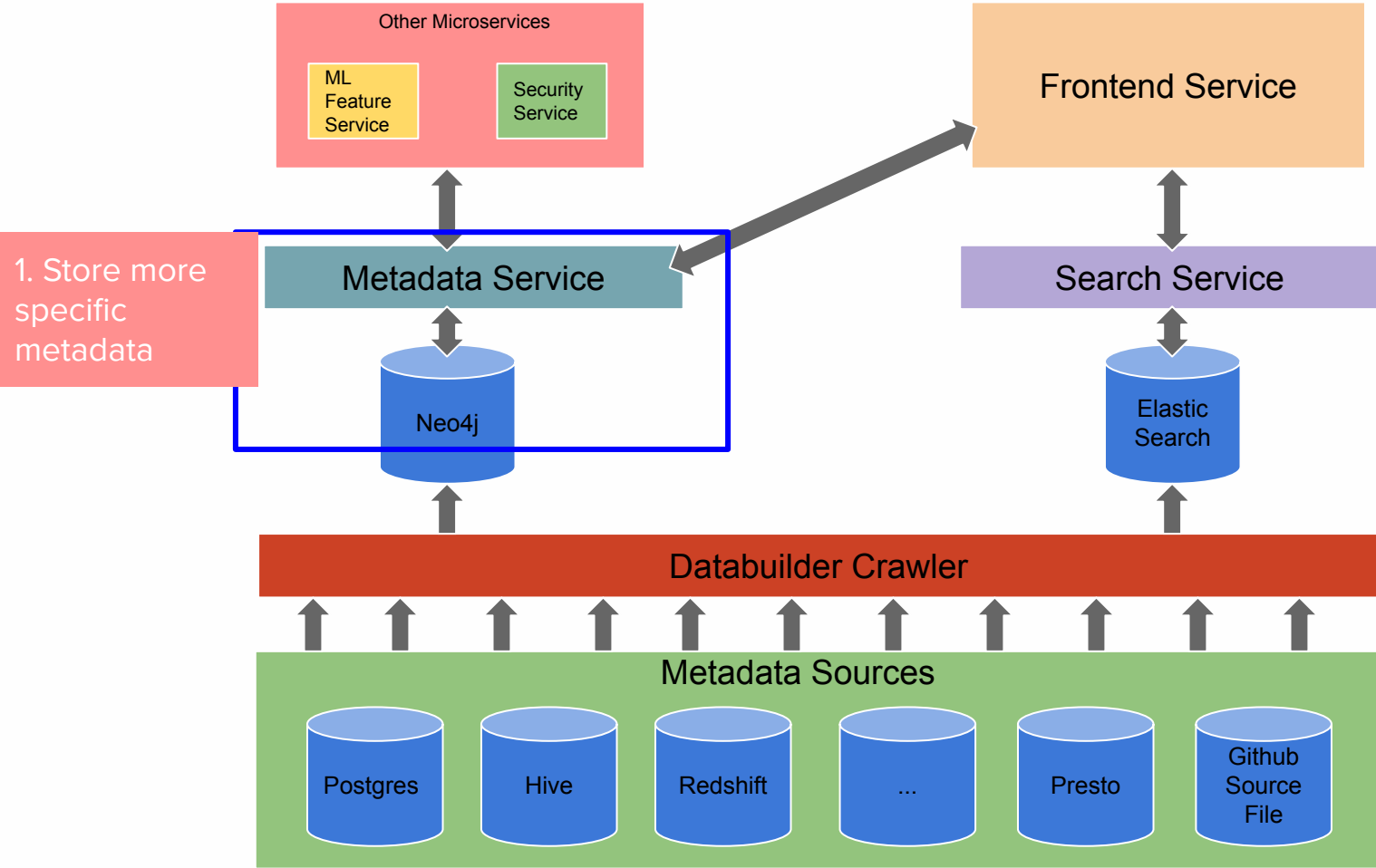


The Problem

- Relied mostly on manual work to understand our data
- We have A LOT of data, and growing
- NOT scalable
- Lots of goals related to data privacy on top of requirements from laws like GDPR / CCPA



1. Store more specific metadata



New Metadata Types



PII
Semantic
Type



Data
Subject
Type

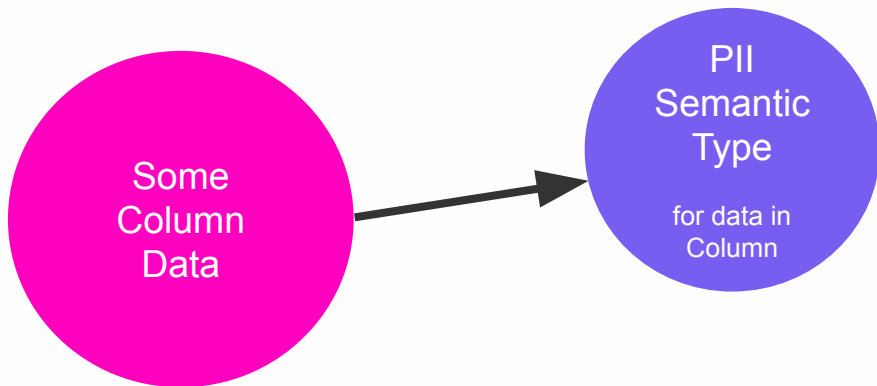


Data
Storage
Security

PII Semantic Type

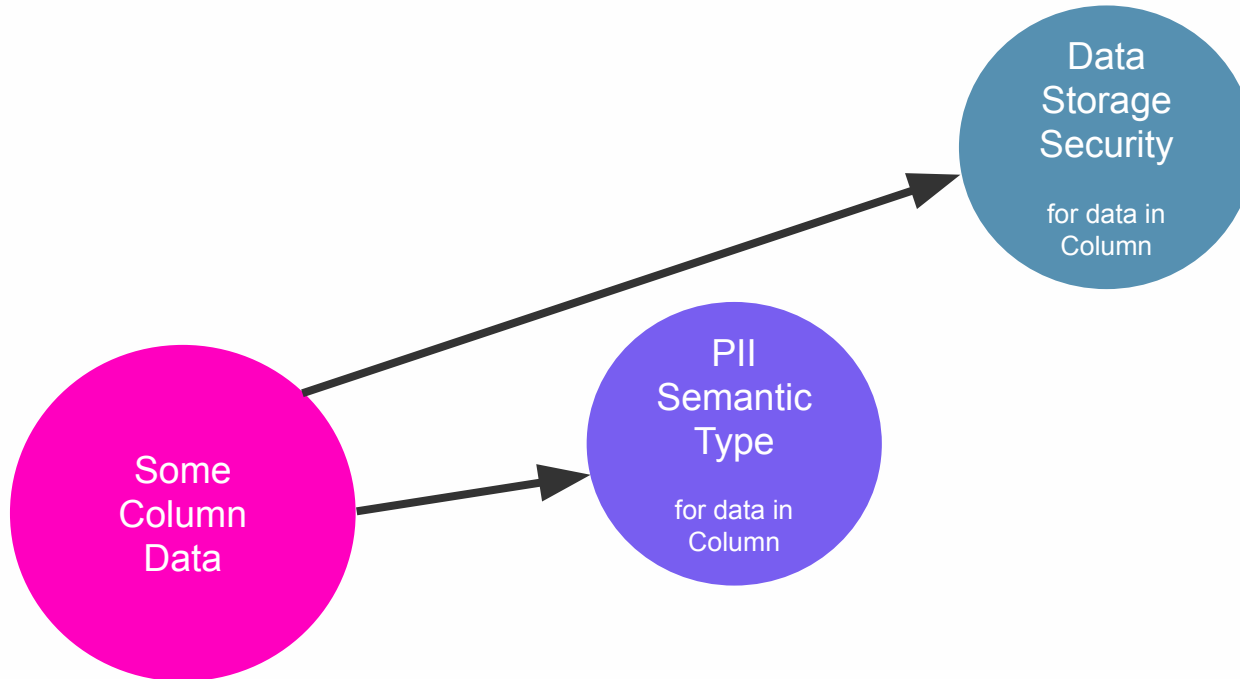
A set value describing the general contents of data in a column, with a focus on discovering data that fits into one of three buckets

- Sensitive by itself
- Could be sensitive when taken together with other data
- Links to sensitive data, like an internal identification token



Data Storage Security

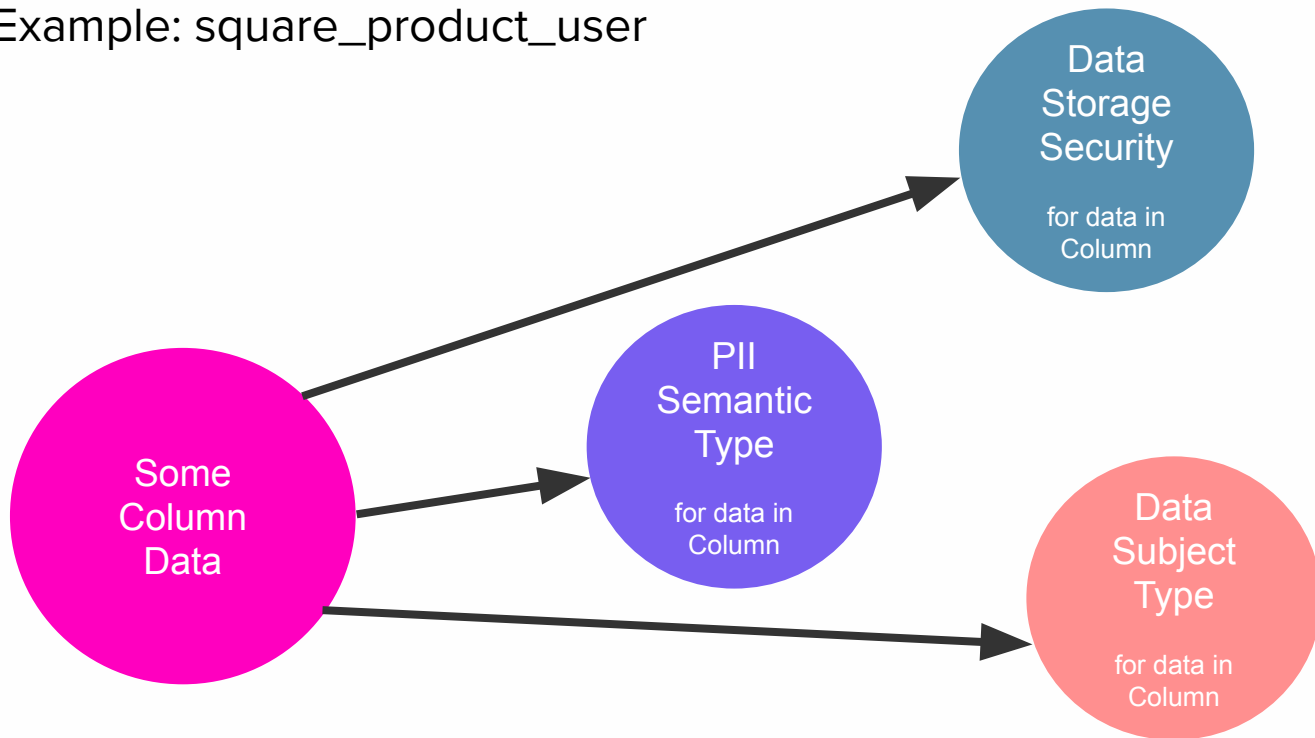
A set value describing the form of data in a column



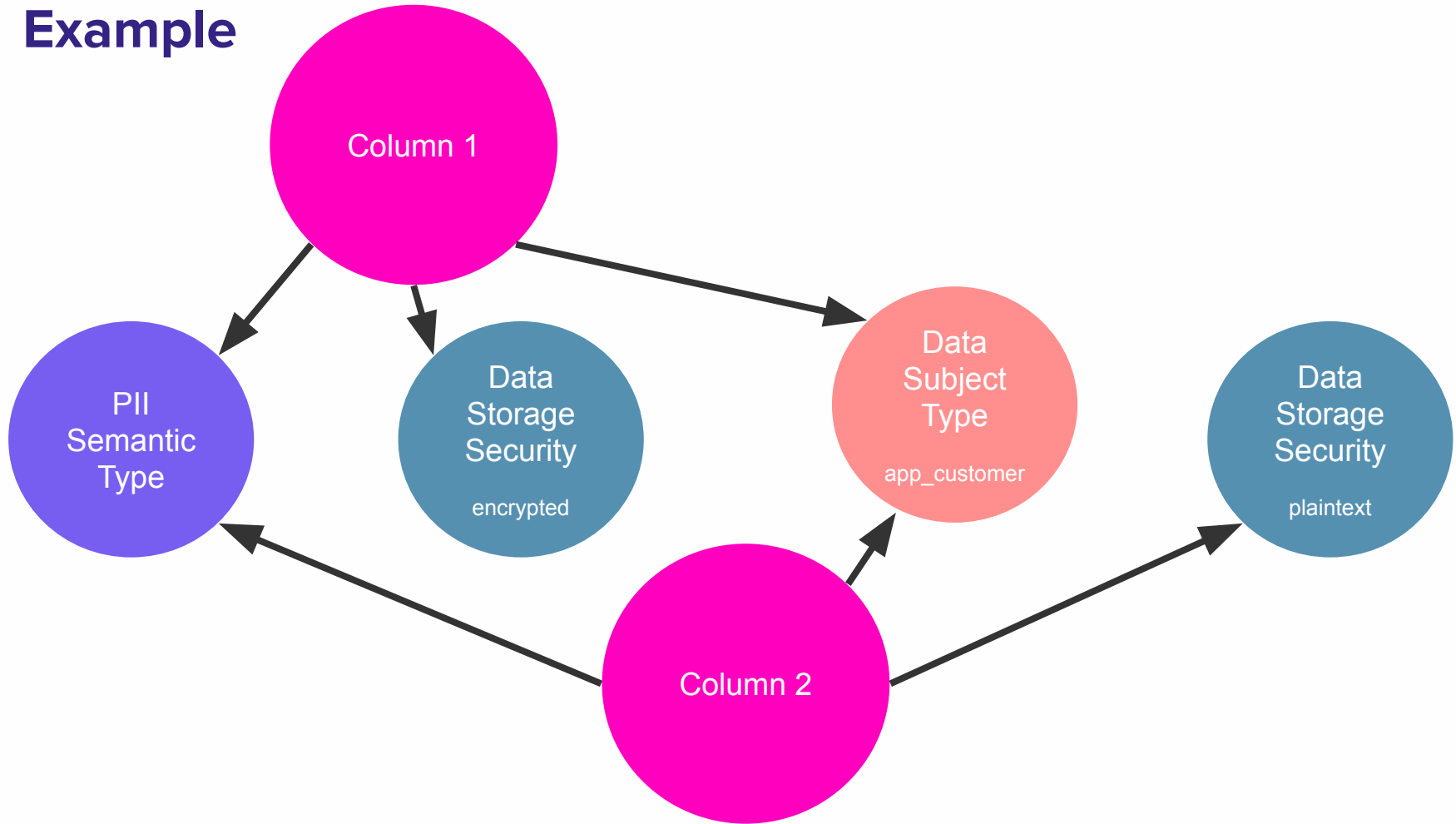
Data Subject Type

A set value describing users whose information exists in a column

- Example: square_product_user




Example



UI Changes

AMUNDSEN [About](#) [Browse](#) [?](#) [AR](#)

<  **foo.bar.wizards** ☆ [Preview](#)

Datasets • Snowflake • staging

Description
A table with basic details about wizards

Last Updated
Jul 08, 2020 12pm PDT

Tags
important

Frequent Users
No frequent users exist

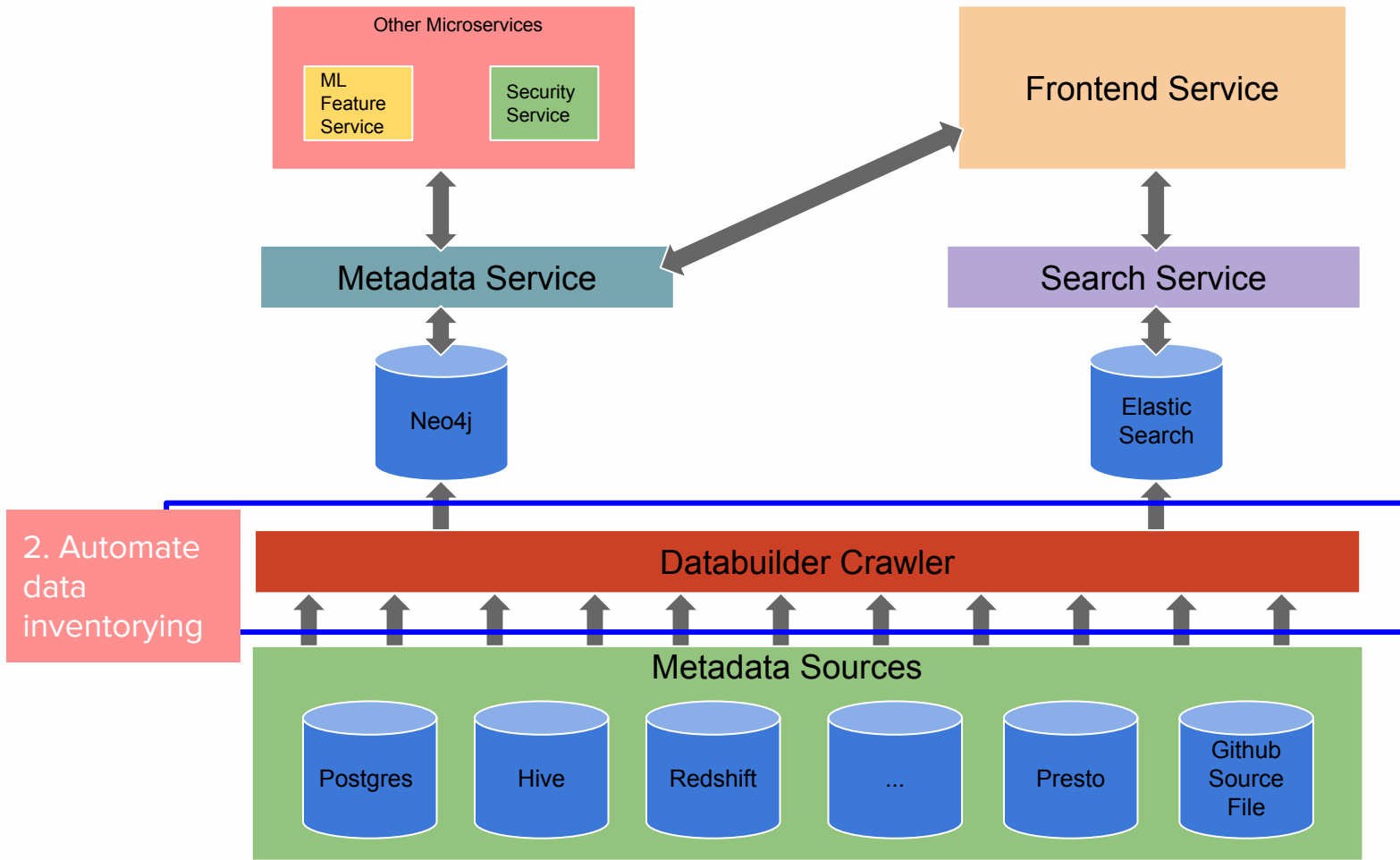
Owners
No owners exist

id The wizard id	varchar(10)
full_name The wizard name	varchar(10)
address The wizard physical address	varchar(24)
birthday Description The wizard's birthday	string

PII Semantic Type(s): BIRTHDATE Data Subject Type(s): SOME_PERSON

2. Automate data inventorying*

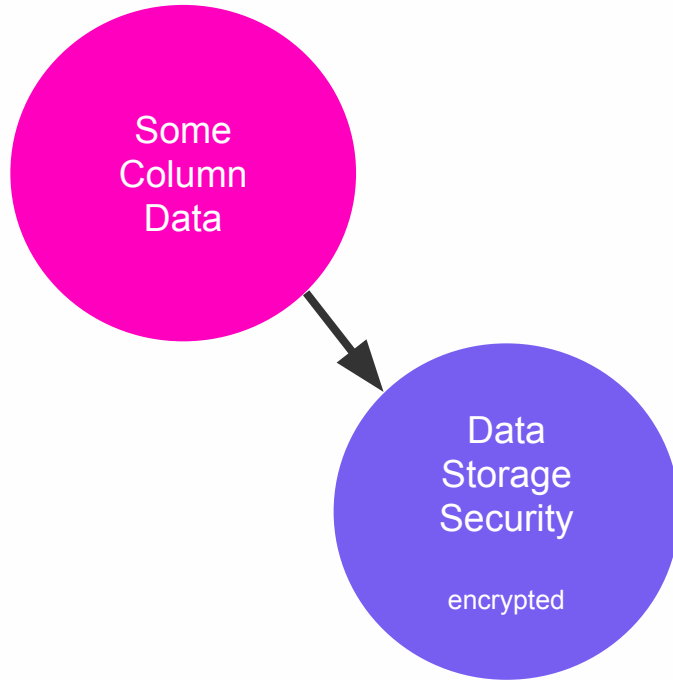
* where possible



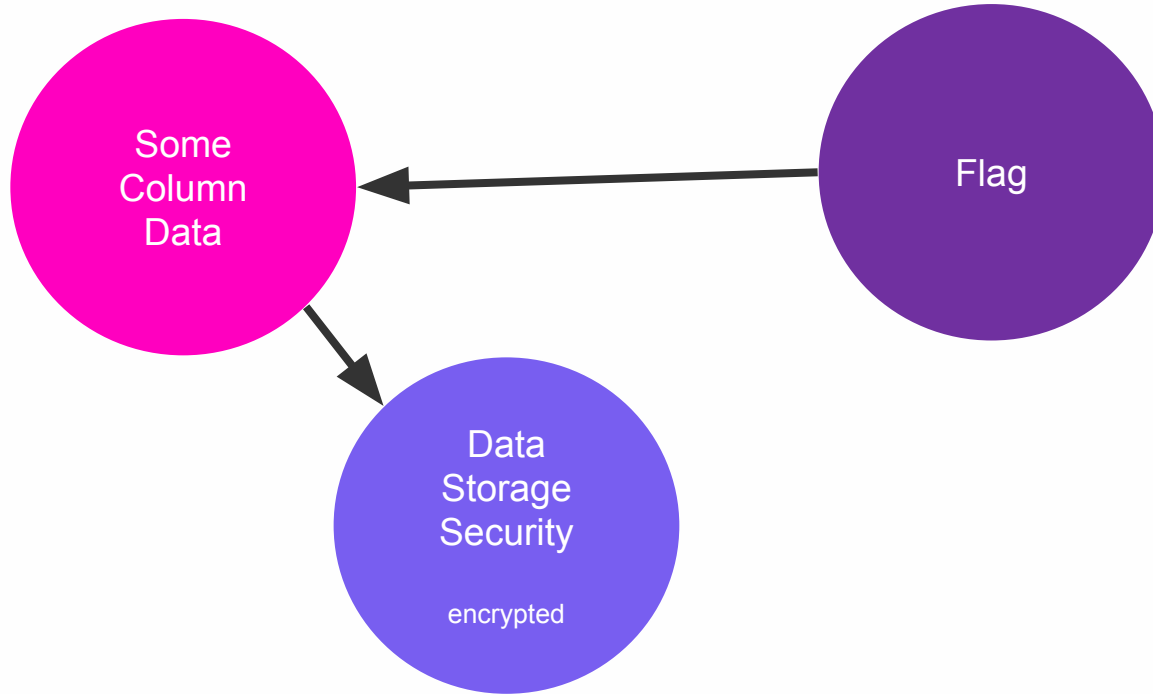
Why?

- We have a lot of data!
- Metadata should be kept current at scale
- New data sources get created all the time
- Mistakes happen

Flagging Possibly Sensitive Data Locations



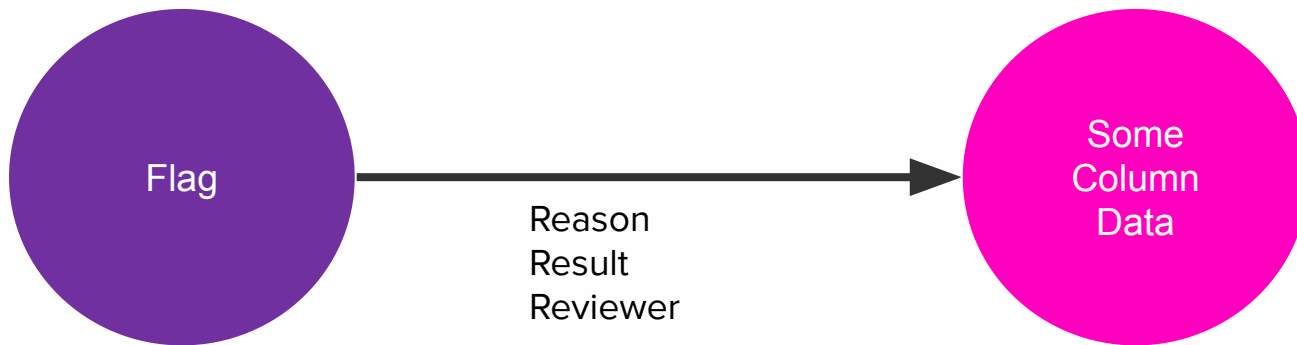
Flagging Possibly Sensitive Data Locations



Flags

Contain information about:

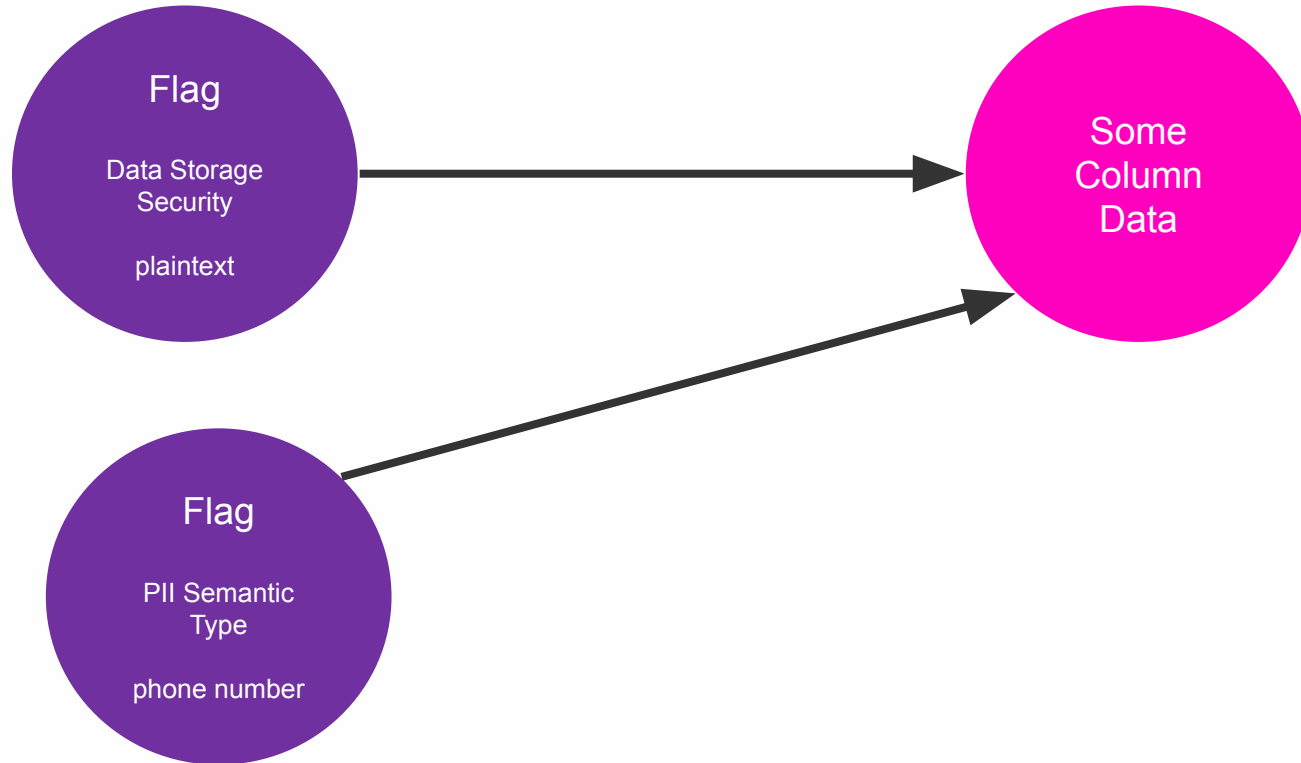
- The metadata type it relates to (i.e. PII Semantic Type, Data Storage Security)
- The possible value (i.e. encrypted for Data Storage Security)
- Human readable description



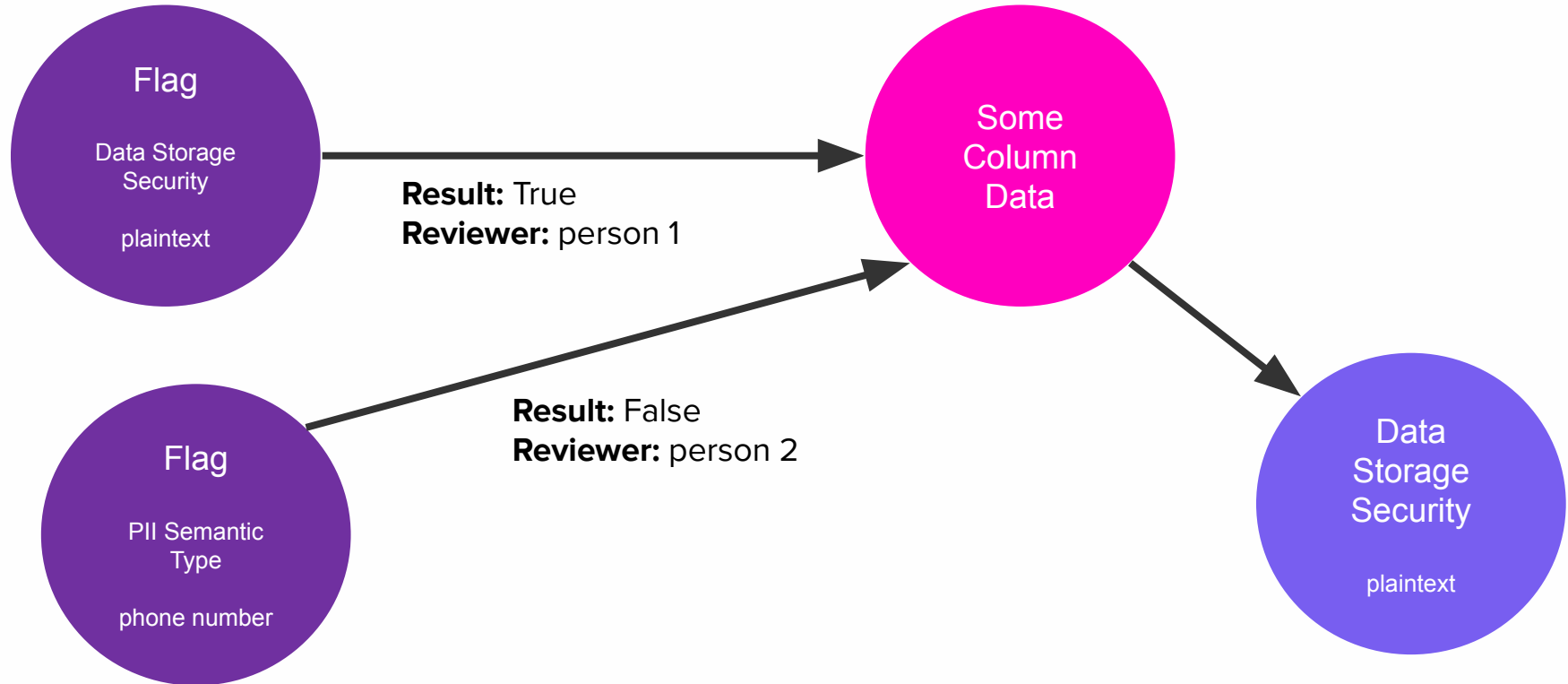
Checking for Data Sensitivity

- Column name
 - Keyword partial or full match
 - Literal PII match
- Column type
- Table name
- Sample values + Google DLP API

User Flow



User Flow






UI Changes

AMUNDSEN About Browse ? AR


Flagged Content

The tables and columns below have unresolved flags. Click "yes" or "no" to resolve these flags and tag the element with the suggested metadata. Read more about our data classification policy [here](#). See full list of metadata options [here](#).

 BigQuery://test-cluster.wizards/basic_pii/phone_number Flagged for: pii_semantic_type - PHONE_NUMBER Column may contain: Personal phone numbers	<input type="button" value="Yes"/> <input type="button" value="No"/>
 Snowflake://staging.foo.bar/wizards/birthday Flagged for: pii_semantic_type - BIRTHDATE Column may contain: individuals' date of birth, including birth year	<input type="button" value="Yes"/> <input type="button" value="No"/>
 Snowflake://staging.foo.bar/wizards/full_name Flagged for: pii_semantic_type - PERSON_NAME Column may contain: individuals' name	<input type="button" value="Yes"/> <input type="button" value="No"/>

UI Changes

AMUNDSEN [About](#) [Browse](#) [?](#) [AR](#)

<  **wizards.basic_pii** ☆ [Preview](#)

Datasets • BigQuery • test-cluster

Description

Includes sensitive information about wizards and their studies

Last Updated
Jul 14, 2020 12pm PDT

Tags
[+ New](#)

Frequent Users
No frequent users exist

Owners
No owners exist

id	int
name Name of the wizard	varchar(10)
age Age of the wizard	int
house Wizards house the wizard belongs to	varchar(10)
phone_number Phone number of the wizard	varchar(10)

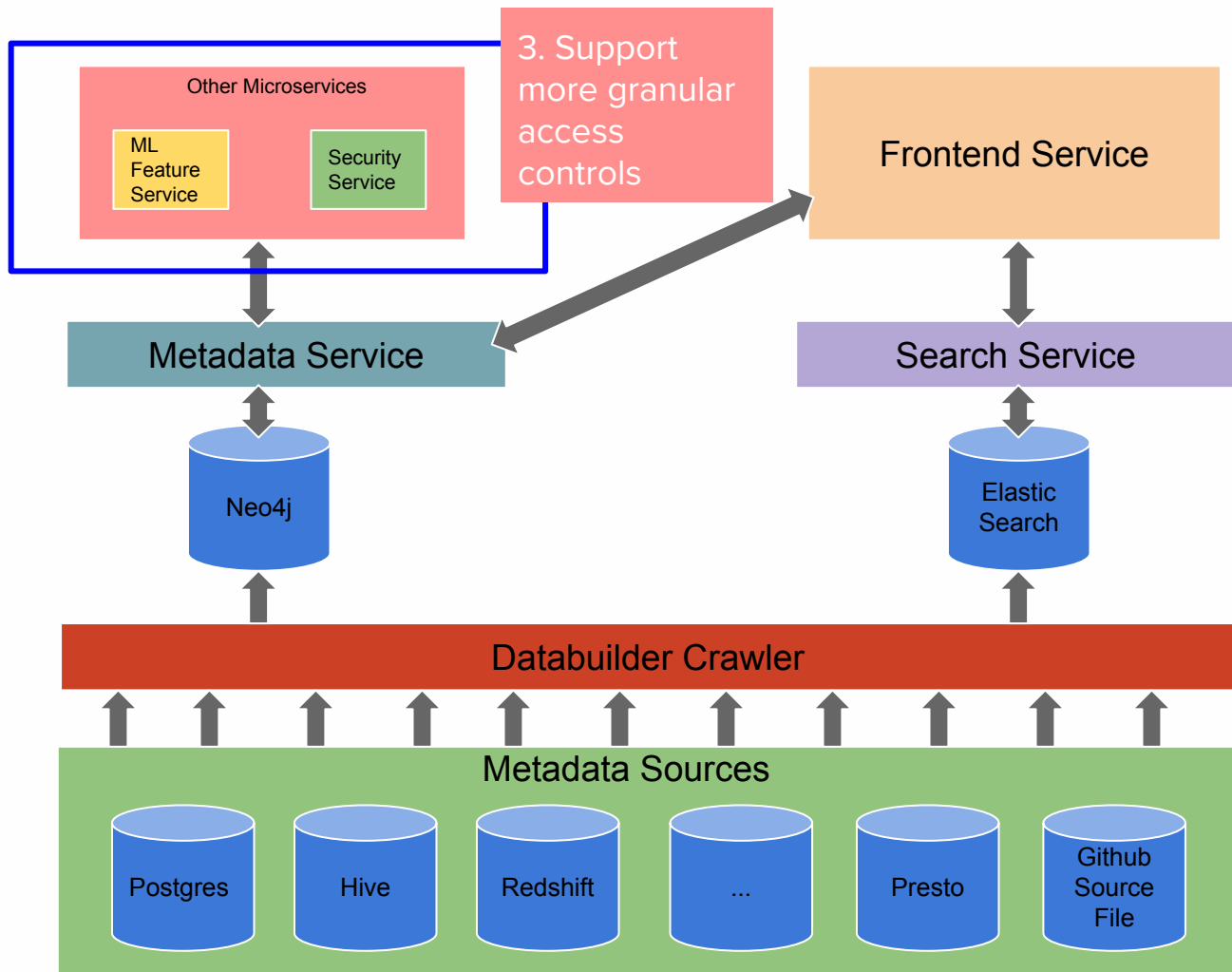
! WARNING ! This Column has been flagged for possibly containing sensitive information. It may include:

- Personal phone numbers

Do you think this flag makes sense? Please click "yes" or "no" next to the flag.

[Yes](#) [No](#)

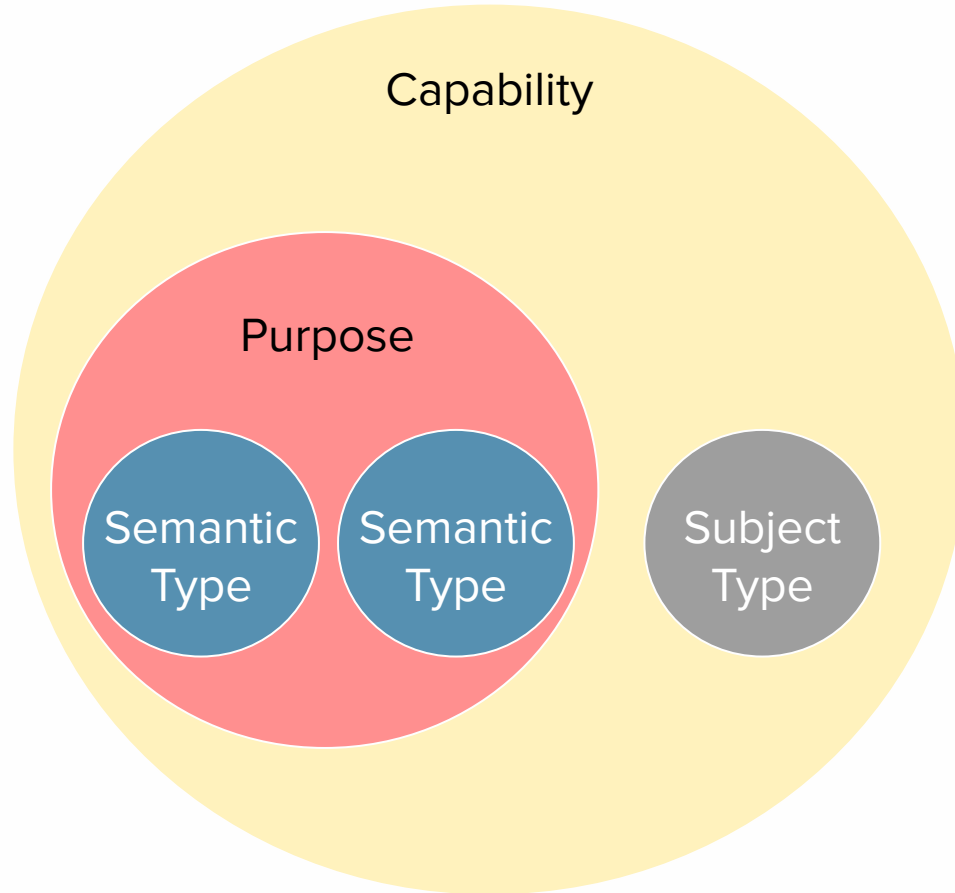
3. Support more granular access controls



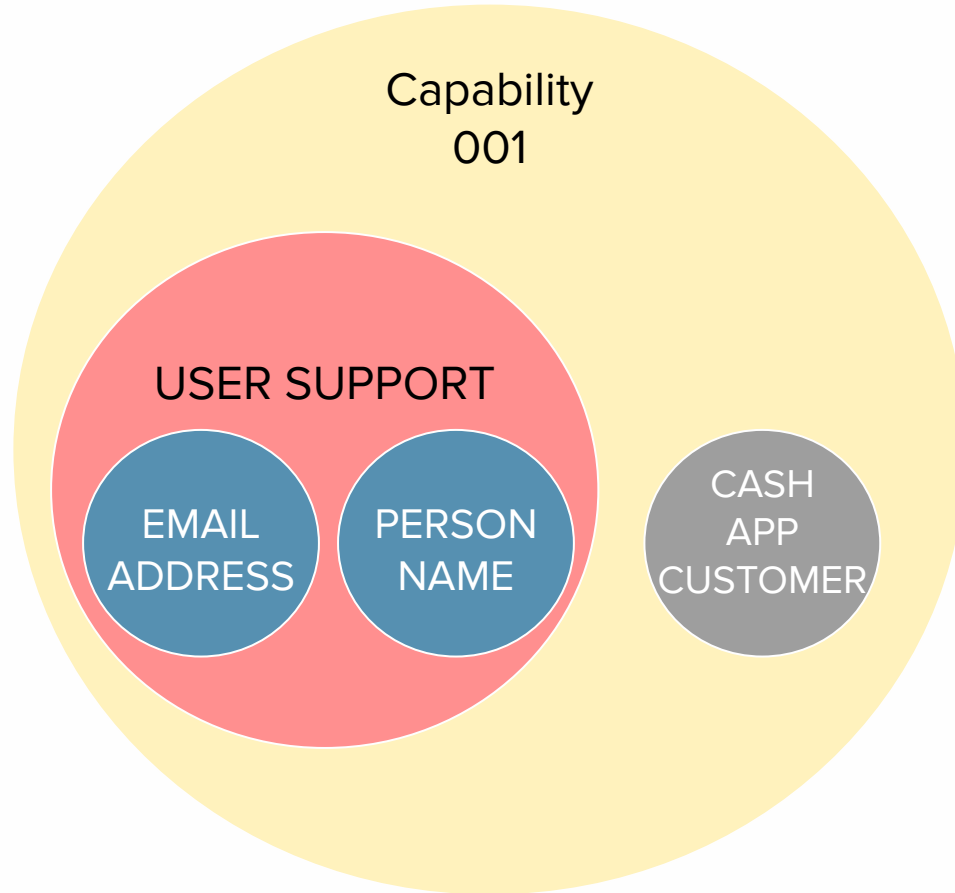
Capability-governed access control

- Purpose - An enumerated, legal-approved reason to access specific PII Semantic Types
- Capability - A combination of Purpose and Data Subject Type

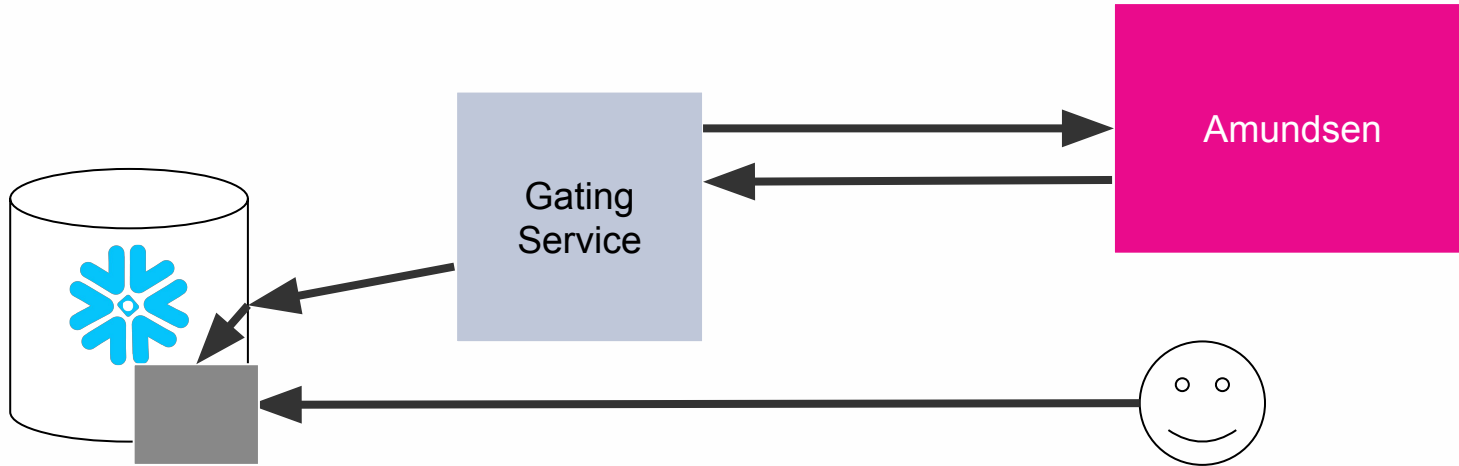
Capability-governed access control



Capability-governed access control



Capability-governed access control



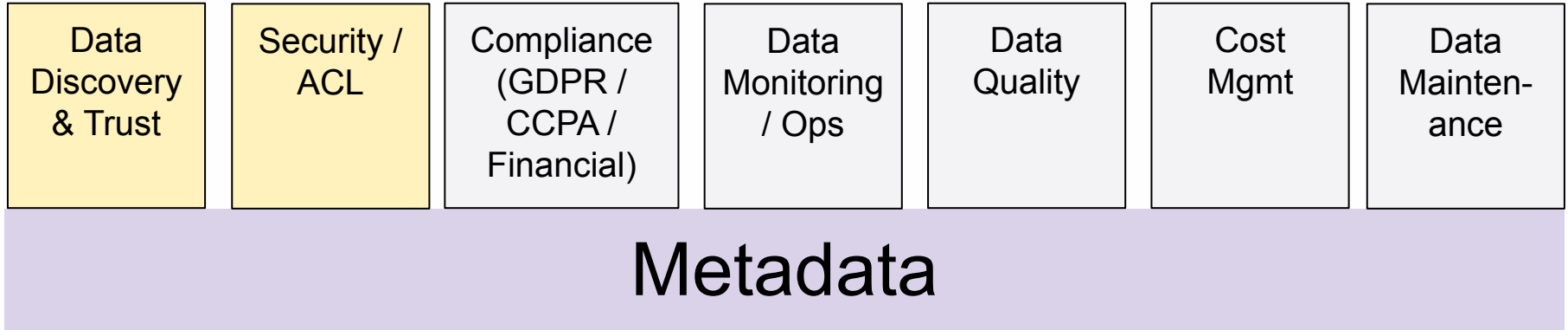
4. Future Work

Future and Work in Progress

- Gremlin Proxy
- UI components
 - show users exactly the data they have access to based on their capabilities
 - show the data that is tied to a specific capability
 - empower data owners to revoke broad access to some data
- Supporting additional data types
- And beyond! (we are hiring...)

Summary

Develop breadth of applications



Get started!

Check out Github:

github.com/lyft/amundsen

Join Slack (~700 users):

[Link to join on github](#)

Check out other blogs and videos:

[Amundsen for Privacy @ Square](#)

[Introducing Amundsen](#)

[Amundsen's architecture](#)

[YouTube channel](#)

Thanks!

Mark Grover | Lyft | @mark_grover

Alyssa Ransbury | Square | @alyssaran

Icons under Creative Commons License from <https://thenounproject.com/>