

Technical Advisory Council Meeting

July 30, 2020

 THE **LINUX** FOUNDATION

 **LF AI**

Antitrust Policy Notice

- › Linux Foundation meetings involve participation by industry competitors, and it is the intention of the Linux Foundation to conduct all of its activities in accordance with applicable antitrust and competition laws. It is therefore extremely important that attendees adhere to meeting agendas, and be aware of, and not participate in, any activities that are prohibited under applicable US state, federal or foreign antitrust and competition laws.
- › Examples of types of actions that are prohibited at Linux Foundation meetings and in connection with Linux Foundation activities are described in the Linux Foundation Antitrust Policy available at <http://www.linuxfoundation.org/antitrust-policy>. If you have questions about these matters, please contact your company counsel, or if you are a member of the Linux Foundation, feel free to contact Andrew Updegrave of the firm of Gesmer Undergone LLP, which provides legal counsel to the Linux Foundation.

Recording of Calls

Reminder:

TAC calls are recorded and available for viewing on the [TAC Wiki](#)

Reminder: LF AI Useful Links

Web site: lfai.foundation
Wiki: wiki.lfai.foundation
GitHub: github.com/lfai
Landscape: landscape.lfai.foundation or l.lfai.foundation
Mail Lists: <https://lists.lfai.foundation>

LF AI Logos: <https://github.com/lfai/artwork/tree/master/lfai>

LF AI Presentation Template:

https://drive.google.com/file/d/1eiDNJvXCqSZHT4Zk_-czASlz2GTBRZk2/view?usp=sharing

Events Page on LF AI Website: <https://lfai.foundation/events/>

Events Calendar on LF AI Wiki (subscribe available):

<https://wiki.lfai.foundation/pages/viewpage.action?pageId=12091544>

Event Wiki Pages: <https://wiki.lfai.foundation/display/DL/LF+AI+Foundation+Events>

Agenda

- › Roll Call
- › Approval of Minutes
- › Amundsen Incubation Project Proposal + TAC Vote
- › Upcoming TAC Meetings
- › Open Discussion

TAC Voting Members

| Member | Contact | Email |
|-------------------|----------------------|--|
| AT&T | Reuben Klein | rk1518@att.com |
| Baidu | Daxiang Dong | dongdaxiang@baidu.com |
| Ericsson | Rani Yadav-Ranjan | rani.yadav-ranjan@ericsson.com |
| Huawei | Huang Zhipeng | huangzhipeng@huawei.com |
| Nokia | Pantelis Monogioudis | pantelis.monogioudis@nokia.com |
| Tech Mahindra | Nikunj Nirmal | nn006444@techmahindra.com |
| Tencent | Bruce Tao | brucetao@tencent.com |
| Zilliz | Jun Gu | jun.gu@zilliz.com |
| ZTE | Wei Meng | meng.wei2@zte.com.cn |
| Acumos AI Project | Nat Subramanian | natarajan.subramanian@techmahindra.com |
| Angel Project | Bruce Tao | brucetao@tencent.com |
| ONNX Project | Jim Spohrer* | spohrer@us.ibm.com |

* TAC Chairperson

Approval of Minutes

Draft minutes from the June 18th & July 16th meeting of the TAC were previously distributed to the TAC members

Proposed Resolution:

- › That the minutes of the June 18th & July 16th meeting of the Technical Advisory Council of the LF AI Foundation are hereby approved

Project Contribution Proposal: Amundsen

Project Contribution Proposal: Amundsen

Amundsen is a metadata driven application for improving the productivity of data analysts, data scientists and engineers when interacting with data. It does that today by indexing data resources (tables, dashboards, streams, etc.) and powering a page-rank style search based on usage patterns (e.g. highly queried tables show up earlier than less queried tables). Think of it as Google search for data. The project is named after Norwegian explorer [Roald Amundsen](#), the first person to discover the South Pole.

- › **GitHub:** <https://github.com/lyft/amundsen>
- › **Projects Level:** Incubation
- › **Presenter(s):** Mark Grover and Tao Feng
- › **Proposal:** <https://github.com/lfai/proposing-projects/blob/master/proposals/tamundsen.adoc>

Amundsen card on the LF AI landscape



Amundsen

Lyft

Data · Operations

Amundsen is a metadata driven application for improving the productivity of data analysts, data scientists and engineers when interacting with data.

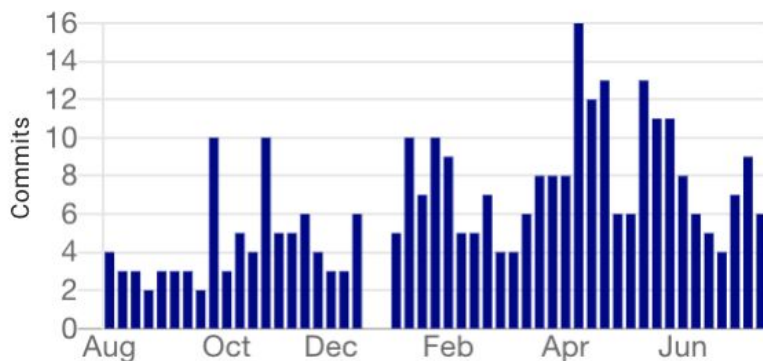
Open Source Software

License Apache License 2.0

No CII Best Practices

Tweet 97

Smarty 60%
Shell 40%



| | | | |
|--------------|----------------------------------|---------------|-------------|
| Website | lyft.github.io/amundsen | | |
| Repository | github.com/lyft/amundsen 1,099 | | |
| Crunchbase | crunchbase.com/organization/lyft | | |
| LinkedIn | linkedin.com/company/lyft | | |
| Twitter | @lyft | Latest Tweet | this week |
| First Commit | a year ago | Latest Commit | this week |
| Contributors | 40 | Headcount | 1,001-5,000 |
| Headquarters | San Francisco, California | | |
| Market Cap | \$9.33B | | |

Tweets by @lyft



Amundsen @ LF AI

Mark Grover | mgrover@lyft.com

Tao Feng | tfeng@lyft.com

(Representing Amundsen team at Lyft)



Why donate Amundsen?

Neutral holding ground

- Vendor-neutral, Not for profit

Growing community

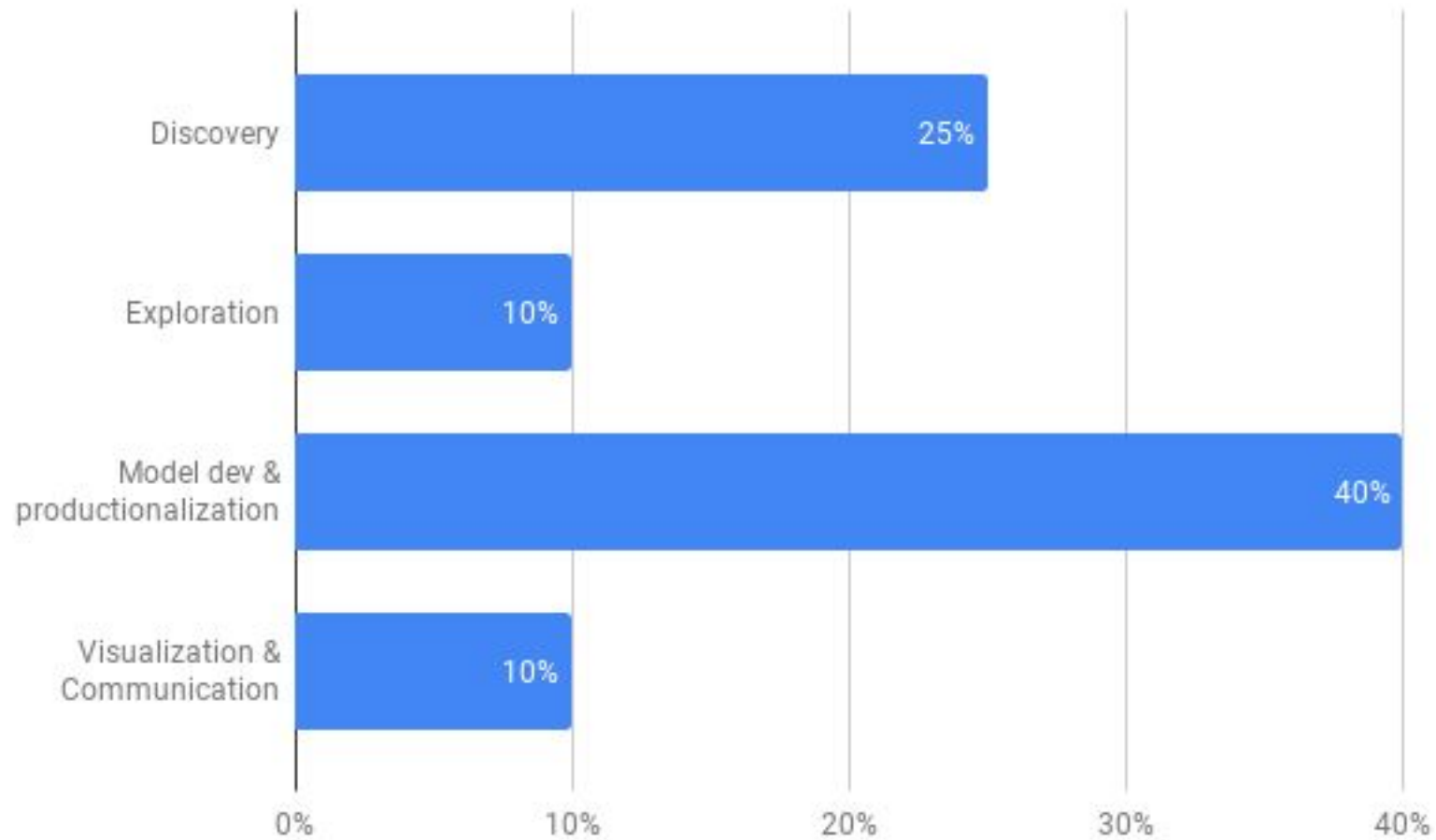
- Increase contributors by converting new & existing users
- Opportunities to collaborate with other hosted projects
- Increase users by broader outreach through the foundation

Open Governance model

- open governance + open source license
- Distills trust in the running & management of the project
- Neutral management of projects' assets by the foundation

Problem

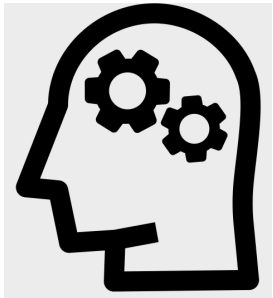
Lots of wasted tech & biz users time



Analyst/DS workflow and time spent on each step

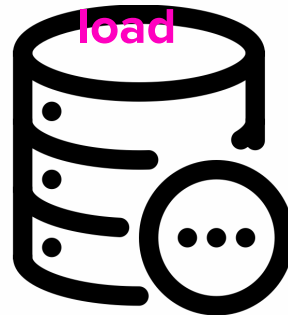
Lack of productivity had many side effects

Lots of unknowns



- Does data exist?
- Prior work?
- Source of truth?
- Who owns it?
- Who uses it?

Increased database load



Lots of queries like:

```
SELECT
*
FROM
default.my_table
WHERE ds='2018-01-01'
LIMIT 100;
```

Interrupt heavy data culture



- No way to know & understand trusted data
- Created channels & oncalls for data questions

Evaluating solutions

Holy grail of solving for productivity

metadata

noun /'medə,dādə,'medə,dadə/

:a **set of data** that describes and gives information about other **data**.

1. What kind of information?

The diagram consists of a central definition of metadata. Two arrows point from the words 'set of data' and 'data' in the definition to two separate boxes. The left box contains the question '1. What kind of information?' and the right box contains '2. About what data?'.

2. About what data?

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?

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:

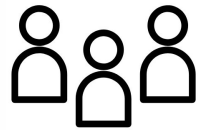
Data stores



PostgreSQL



People



Employees

Dashboard / Reports



looker

Notebooks



Events / Schemas

Schema registry


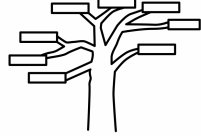

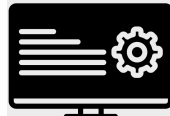
Segment

Streams



TODAY

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> |
















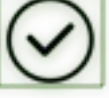



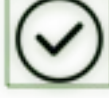










Other requirements

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

Solution space

- Vendors - Alation, Collibra
- Existing open source projects (e.g. [Apache Atlas](#), [Marquez](#))
- LinkedIn's data portal - Wherehows & DataHub ([blog](#), [code](#))
- Twitter's data discovery ([blog](#))
- Netflix's metacat ([code](#), [blog](#))
- Airbnb's data portal ([blog](#), [video](#))
- Big Query SQL Web UI & catalog ([blog](#))
- Goods: Organizing Google's Datasets ([paper](#))
- Data Warehousing and Analytics Infrastructure at Facebook ([paper](#))
- Ground (RISE Lab): <https://rise.cs.berkeley.edu/projects/ground/>

Compared various existing solutions/open source projects

| Criteria / Products | Alation | Where Hows | Airbnb Data Portal | Cloudera Navigator | Apache Atlas |
|---------------------|--|---|---|---|---|
| Search based |  |  |  |  |  |
| Lineage based |  |  |  |  |  |
| Network based |  |  |  |  |  |
| Hive/Presto support |  |  |  |  |  |
| Redshift support |  |  |  |  |  |
| Open source (pref.) |  |  |  |  |  |

Meet Amundsen

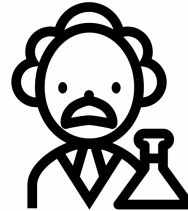
First person to discover the South Pole -
Norwegian explorer, Roald Amundsen

Not all Data Scientists are created equal



Power user

- All info in their head
- Get interrupted a lot due to questions



Noob user


- Lost
- Ask “power users” a lot of questions



Manager

- Dependencies landing on time
- Communicating with stakeholders

Search for data, or browse



AMUNDSEN Announcements Browse 

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  | dynamo |
| 2nd test table | |

Popular Tables

| | |
|---|------|
|  test_schema.test_table1  | Hive |
| 1st test table | |

Amundsen was last indexed on December 18th 2019 at 2:07:17 pm

Search for datasets

AMUNDSEN

Announcements

Browse

T

🔍 **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](#) [FAQ](#) [?](#) [MG](#)

[←](#)  **default.event_amundsenfrontend_user_action** ★
Datasets • Hive

 Lineage  github [Preview](#) [Explore](#)

Description

User action event from Amundsen frontend
[Request Description](#)

Issues

No associated issues

[Report an issue](#)

Date Range

From: Aug 03, 2018
To: Mar 02, 2020






Tags

is testing hello test tracking
a6n amundsen


Last Updated

Mar 05, 2020 7am PST

Owners

-  mgrover@lyft.com
-  jinchang@lyft.com
-  tfeng@lyft.com
-  ttannis@lyft.com
-  dwon@lyft.com

Frequent Users



Read-only information, auto-generated

Columns (17)

[Dashboards \(1\)](#)

| | | |
|--|--------|---|
| event_id Unique event identifier. Due to current assumptions in the pipeline, it's important that this be a version 4 (random) UUID. | string | ⋮ |
| ds test | string | ⋮ |
| command Action command type from user e.g: search, get_table_metadata, etc. | string | ⋮ |
| end_epoch_ms end time in epoch ms | bigint | ⋮ |
| error an error message or exception stacktrace | string | ⋮ |
| host_name Sending host name | string | ⋮ |
| http_request_id | string | ⋮ |
| keyword_args_json json object contains key word arguments | string | ⋮ |
| occurred_at | | |

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 |
|----------|--------|

See dashboards built on this data set

 **AMUNDSEN**

[Announcements](#) [Browse](#) [FAQ](#)  

 **default.event_amundsenfrontend_user_action** ★
Datasets • Hive

 Lineage  github [Preview](#) [Explore](#)

Description

User action event from Amundsen frontend
[Request Description](#)

Issues

No associated issues
[Report an issue](#)

Date Range

From: Aug 03, 2018
To: Mar 02, 2020

Tags

a6n hello test tracking-event
tracking amundsen

Last Updated

Mar 05, 2020 7am PST


Frequent Users

T T S W Y

Owners

- M mgrover@lyft.com
- J jinchang@lyft.com
- T tfeng@lyft.com
- T ttannis@lyft.com
- D dwon@lyft.com

Columns (17) Dashboards (1)

| | | |
|---|------|-------------------------------------|
|  General Amundsen Weekly Active Users ☆ | Mode | Last Successful Run Jul 06, 2020 |
|---|------|-------------------------------------|

Search for existing dashboards/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

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**

Tags
+ New

Created
Jun 01, 2020 11pm PDT






Last Successful Run
Jun 12, 2020 5pm PDT

Last Updated
Jun 12, 2020 5pm PDT

Last Run
Jun 12, 2020 5pm PDT
Succeeded

Recent View Count
1

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

| | tsdb_line | schema | table_name | part_name |
|---|---------------------|---------|------------------------------------|--------------------|
| 1 | 2020-06-01 21:02:47 | default | event_amundsenforstend_user_action | dt=2020-06-01,h=13 |
| 2 | 2020-06-01 21:02:43 | default | event_amundsenforstend_user_action | dt=2020-06-01,h=18 |
| 3 | 2020-06-01 21:02:49 | default | event_amundsenforstend_user_action | dt=2020-06-01,h=17 |
| 4 | 2020-06-01 21:02:37 | default | event_amundsenforstend_user_action | dt=2020-06-01,h=16 |
| 5 | 2020-06-01 21:02:33 | default | event_amundsenforstend_user_action | dt=2020-06-01,h=15 |
| 6 | 2020-06-01 19:22:04 | default | event_amundsenforstend_user_action | dt=2020-06-01,h=9 |
| 7 | 2020-06-01 16:53:44 | default | event_amundsenforstend_user_action | dt=2020-06-01,h=14 |

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

Search for co-workers!

lyft AMUNDSEN

mark grover

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

Amundsen Architecture

1. Metadata Service

Amundsen table detail page

Rides




May 25, 2012 – Mar 03, 2019

The source for all ride related data.

Columns

| | |
|--|---|
| users <code>string</code> | ▼ |
| Dummy description. You can click here to edit. | |
| desk_count <code>int</code> | ▼ |
| Dummy description. You can click here to edit. | |
| passenger <code>string</code> | ▼ |
| Add Description | |
| ride_id <code>string</code> | ▼ |
| Add Description | |
| driver_os <code>string</code> | ▼ |
| Add Description | |
| driver_os_version <code>string</code> | ▼ |
| Dummy description. You can click here to edit. | |
| driver_app_version <code>string</code> | ▼ |
| Add Description | |

OWNED BY

-  test@lyft.com
-  default-user@lyft.com
-  Add

FREQUENT USERS

-     

GENERATED BY

-  rides/rides



SOURCE CODE

-  rides.rides

TABLE LINEAGE (BETA)

-  rides.rides

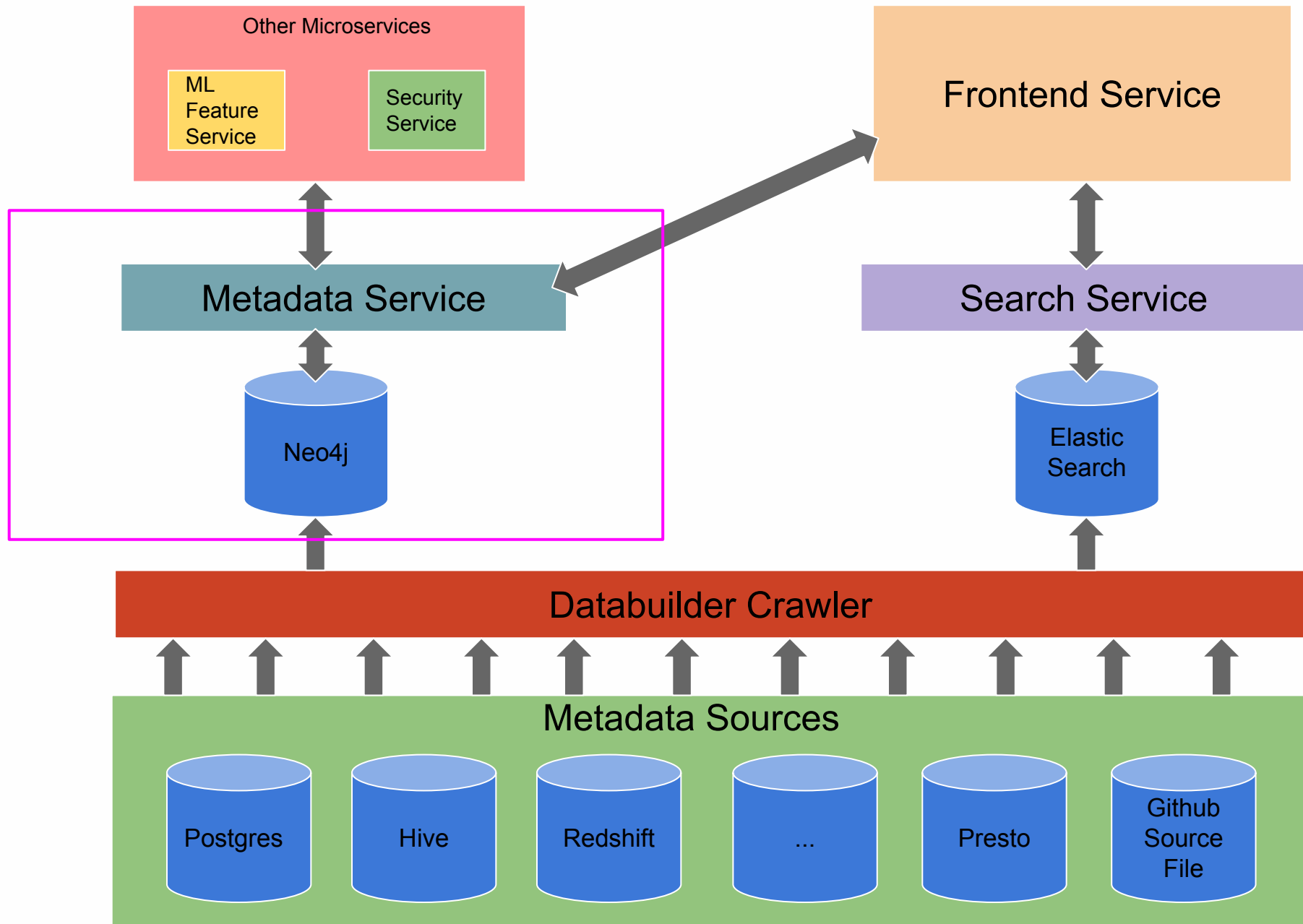
TABLE PROFILE (BETA)

-  Preview Data
-  Explore with SQL

TAGS

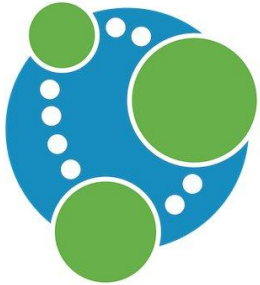
- driver
- passenger
- events





Metadata Service

- A thin proxy layer to interact with graph database
 - Currently Neo4j is the default option for graph backend engine
 - Work with the community to support Apache Atlas



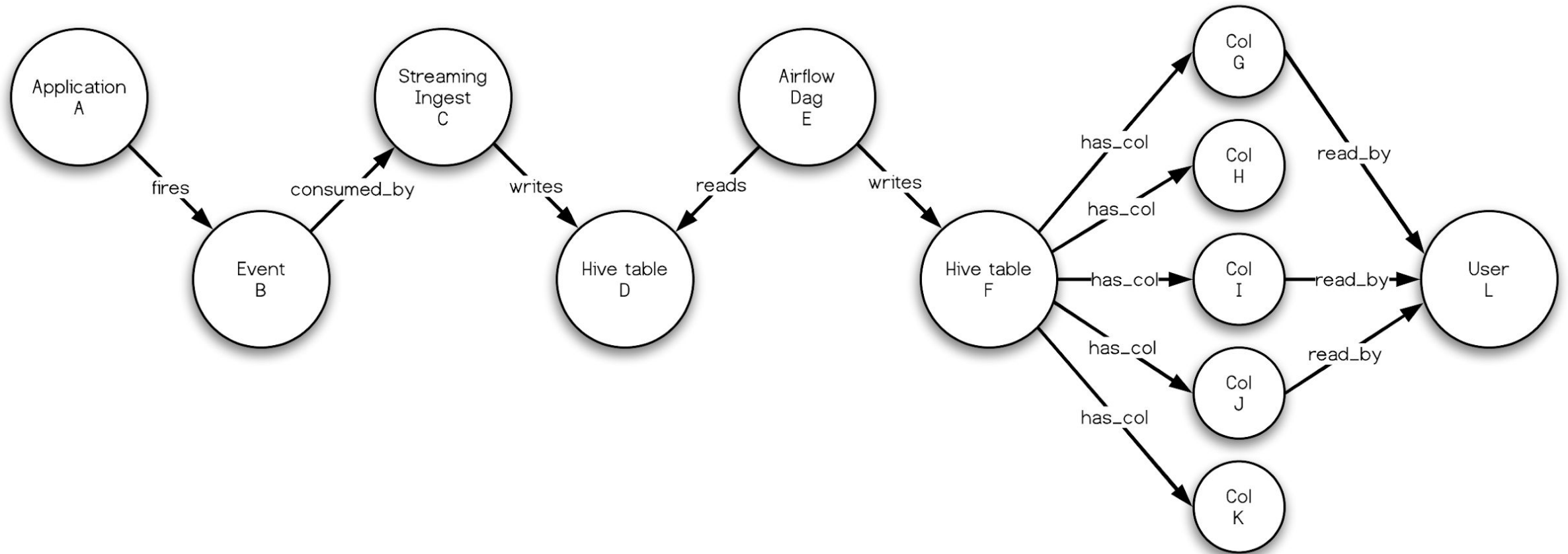
Apache **Atlas**

- Support Rest API for other services pushing / pulling metadata directly

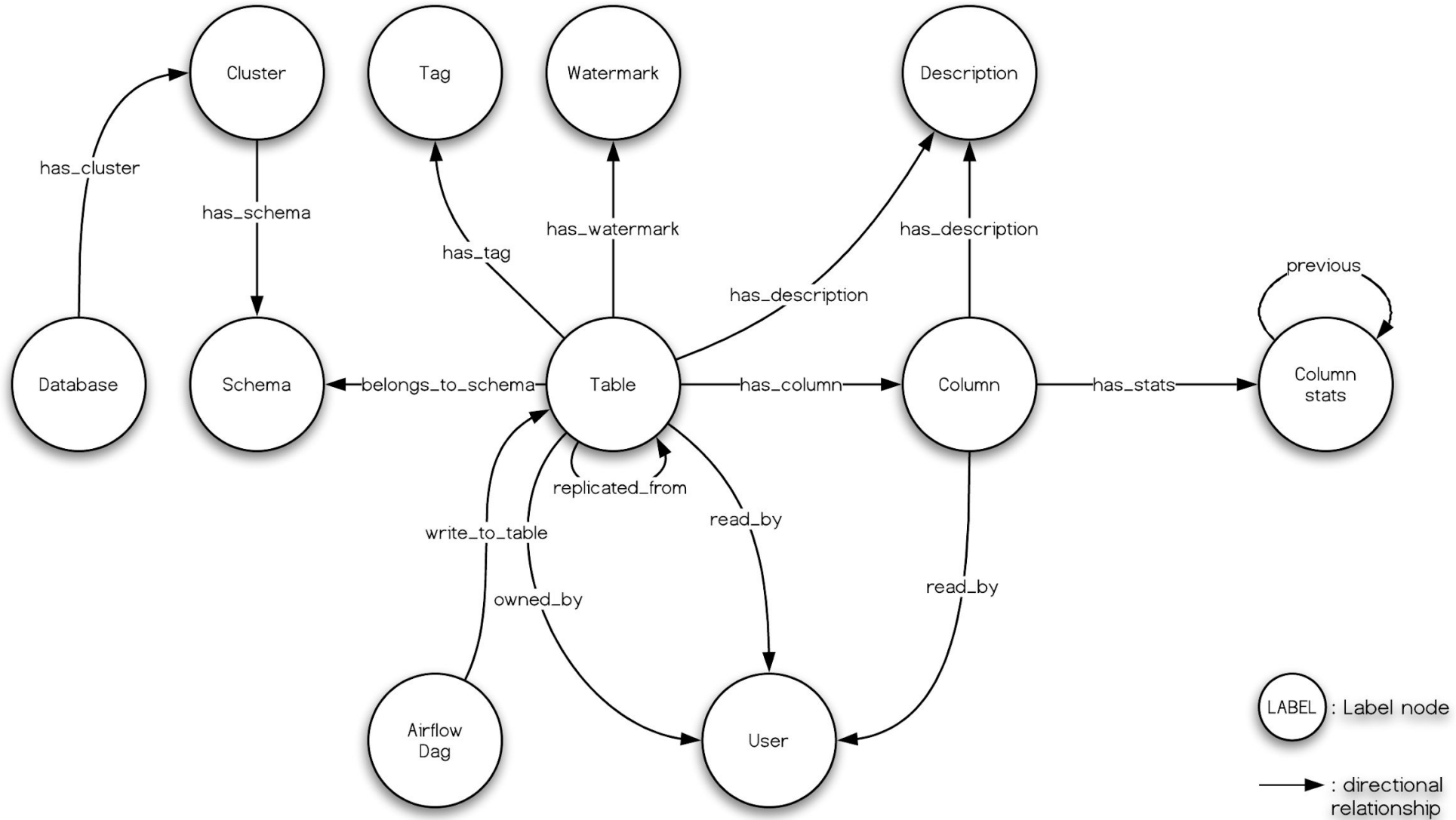
Challenge #1

Choosing the right
metadata model

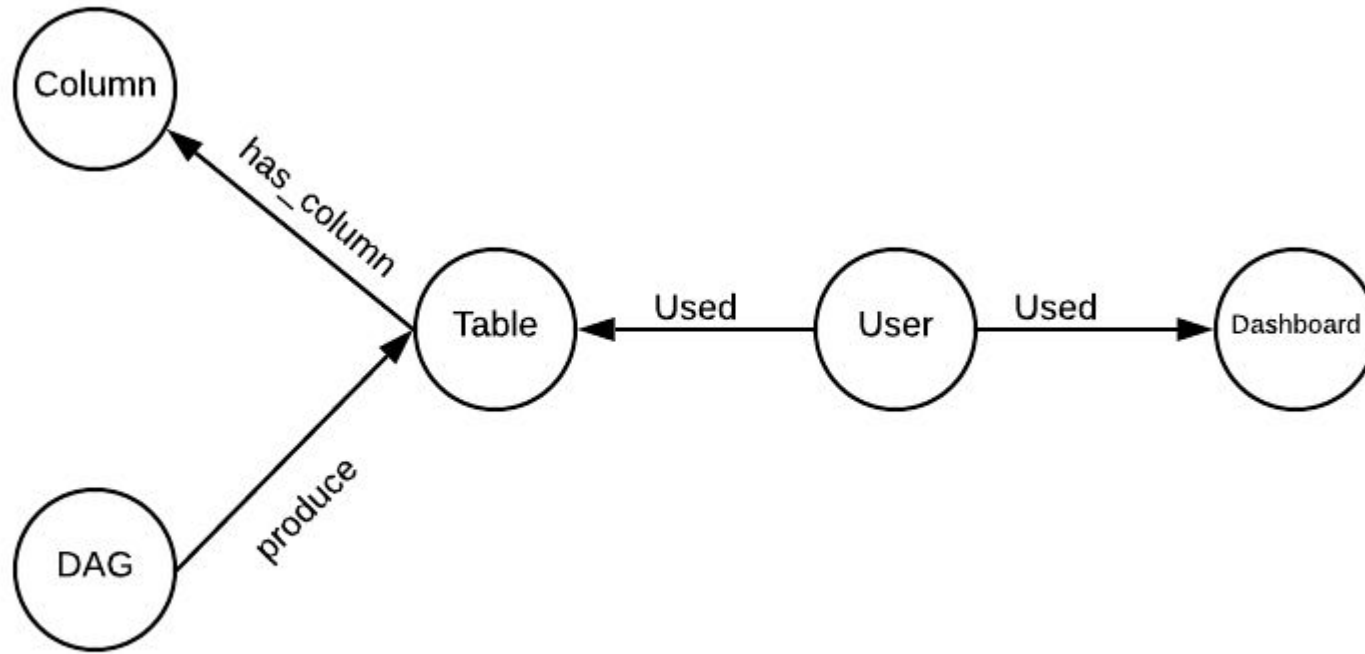
Logical map of the world



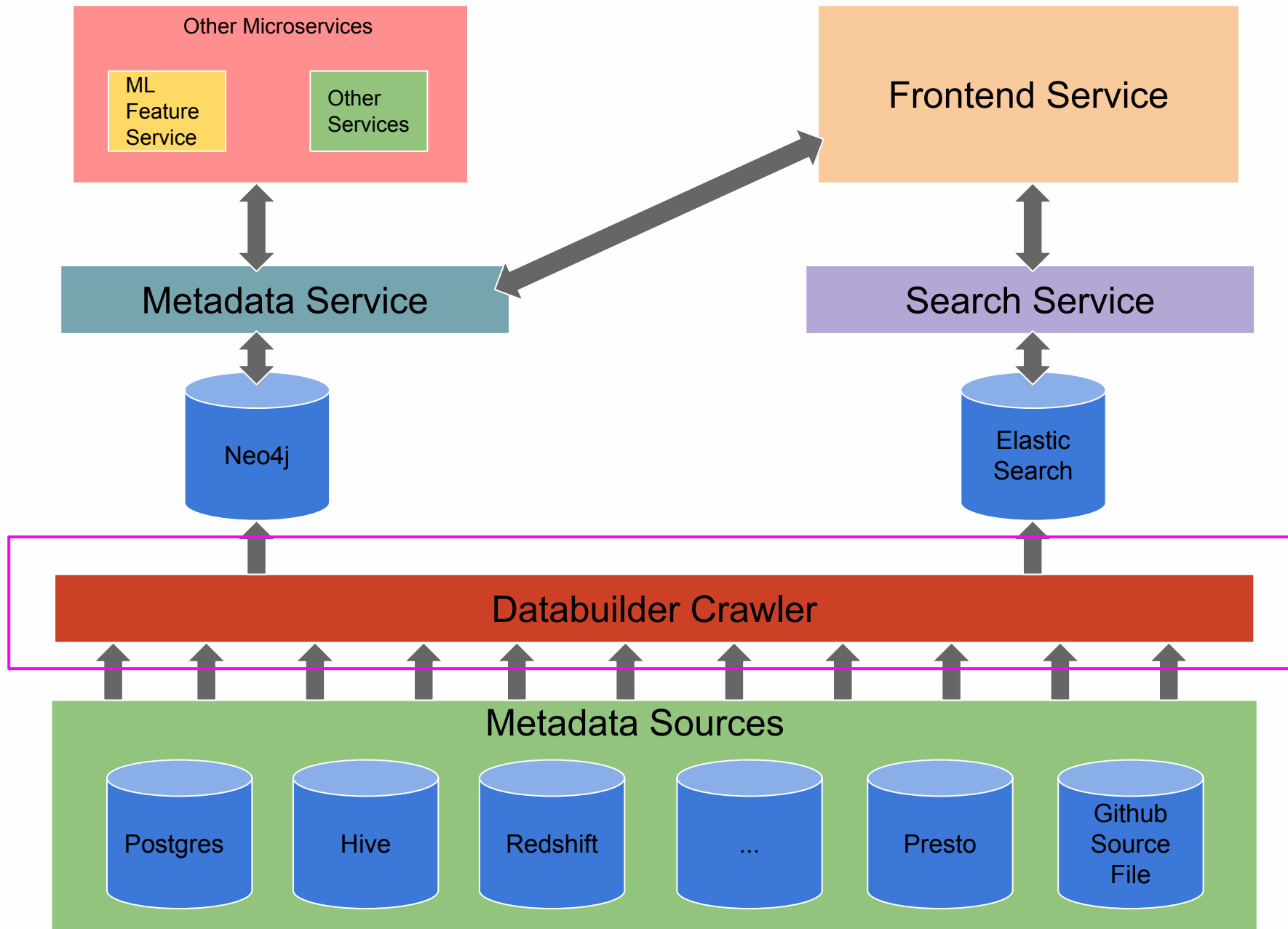
Current graph model



Graph makes it easy to extend to more data resources



2. Databuilder



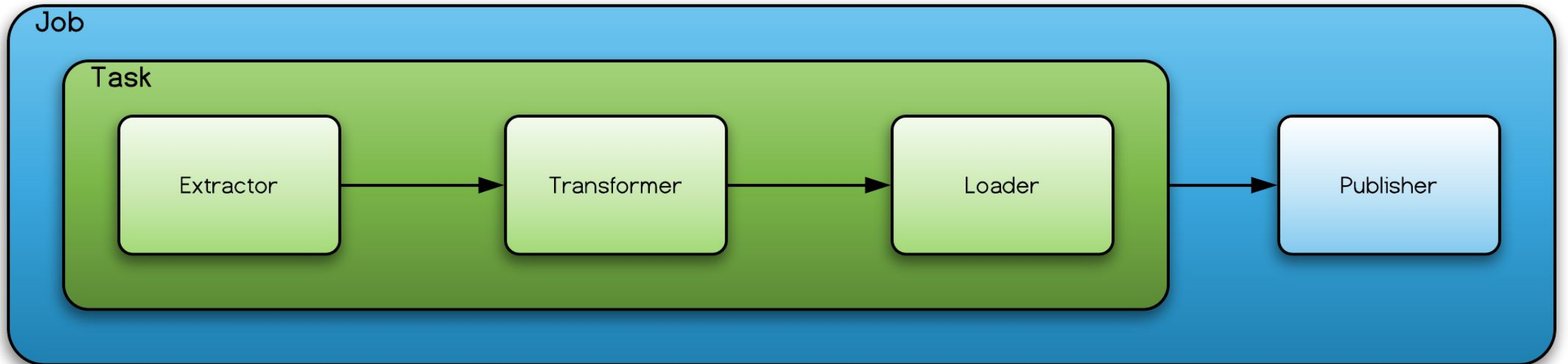
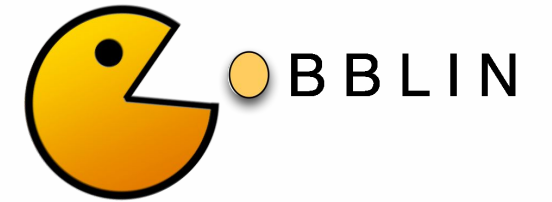
Challenge #2

Various forms of metadata

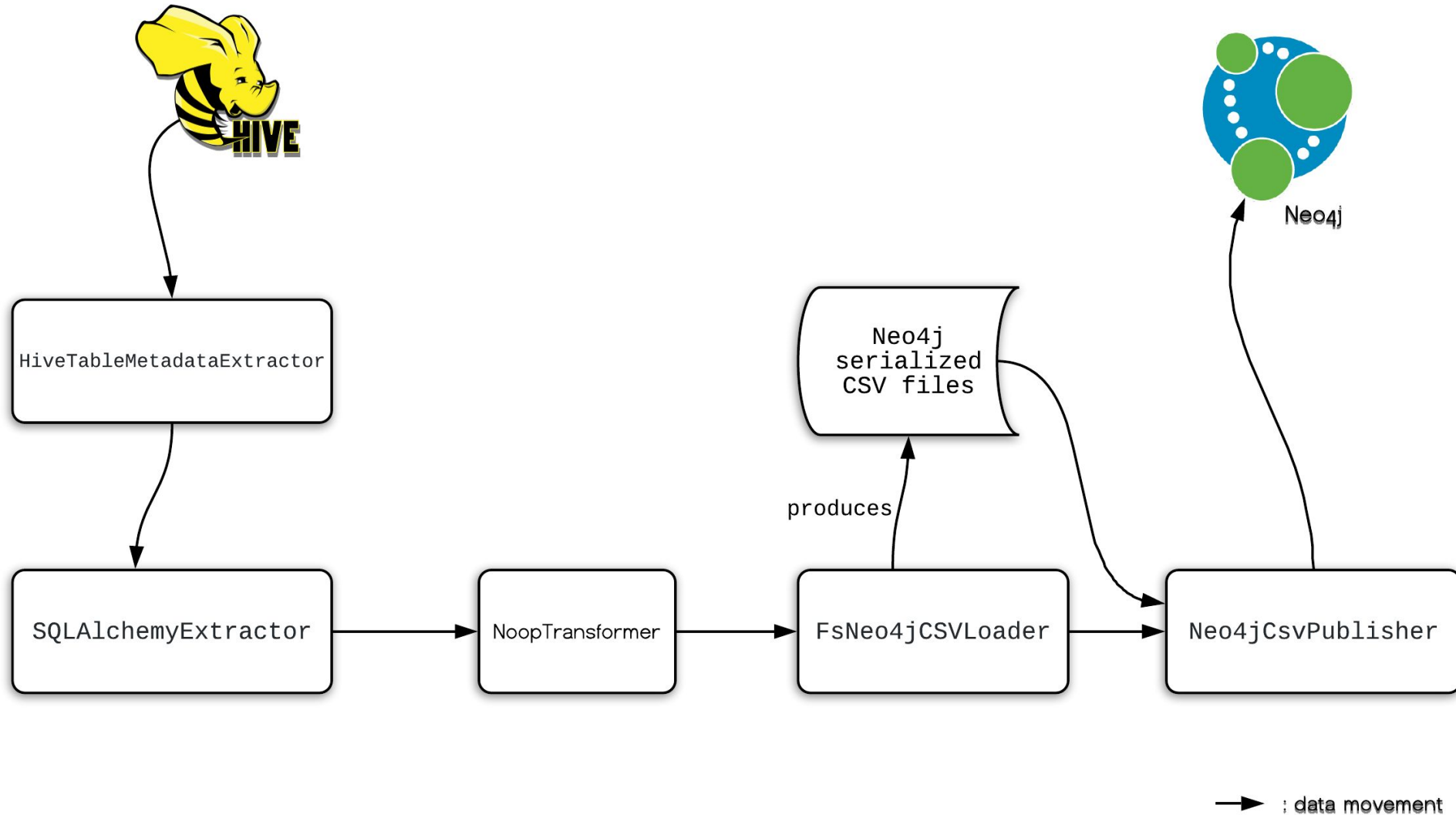
Metadata Sources @ Lyft



Databuilder



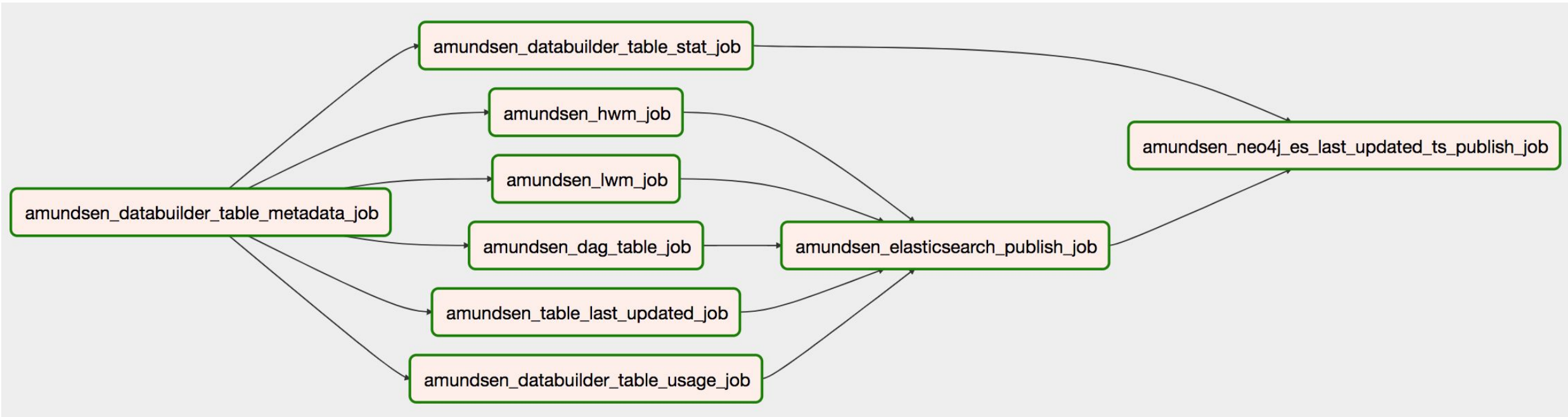
Databuilder in action



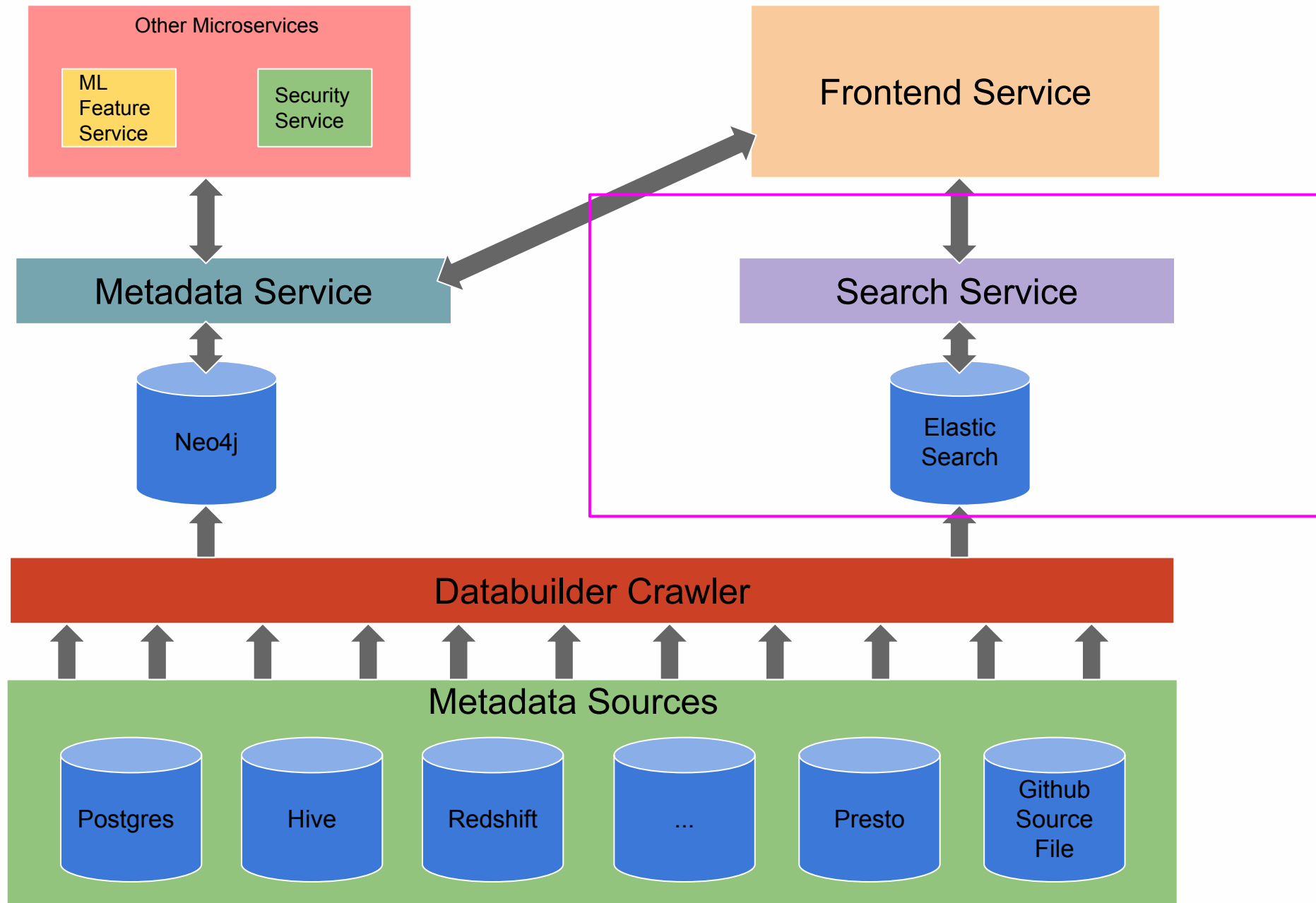
How is the databuilder orchestrated?



Amundsen uses Apache Airflow to orchestrate Databuilder jobs



3. Search service



3. Search Service



- A thin proxy layer to interact with the search backend
 - Currently it supports Elasticsearch as the search backend.
- Support different search patterns
 - **Normal** Search: match records based on relevancy
 - **Category** Search: match records first based on data type, then relevancy
 - **Wildcard** Search

Challenge #3

How to make the search
result more relevant?

How to make the search result more relevant?

- Define a search quality metric
 - Click-Through-Rate (CTR) over top 5 results

- Search behaviour instrumentation is important

- Couple of improvements:
 - Boost the **exact table** ranking
 - Support **wildcard** search (e.g. `event_*`)
 - Support **category** search (e.g. `column: is_line_ride`)

Repeated search in 30 sec (for analysis)

| occurred_at | user_value | search_term | search_result |
|----------------------------|----------------------|-------------|--|
| 2019-10-01 23:32:35.663 | [REDACTED]g@lyft.com | [REDACTED] | |
| 2019-10-01 23:23:33.123 | [REDACTED]z@lyft.com | [REDACTED] | t.dimension_ride_invoices |
| 2019-10-01 23:23:32.970 | [REDACTED]z@lyft.com | [REDACTED] | t.dimension_ride_invoices |
| 2019-10-01 23:12:53.286 | [REDACTED]g@lyft.com | [REDACTED] | ns |
| 2019-10-01 23:01:32.179 | [REDACTED]n | [REDACTED] | dsr |
| 2019-10-01 22:50:08.910 | [REDACTED]:om | [REDACTED] | |
| 2019-10-01 22:46:28.131 | [REDACTED]n | [REDACTED] | _dynamo_incremental_wallet_charge_accounts |
| 2019-10-01 22:23:59.170 | [REDACTED]um | [REDACTED] | s |
| 2019-10-01 22:23:51.928 | [REDACTED]um | [REDACTED] | ension rides |

4. Frontend service

Frontend Stacks



ReactJS



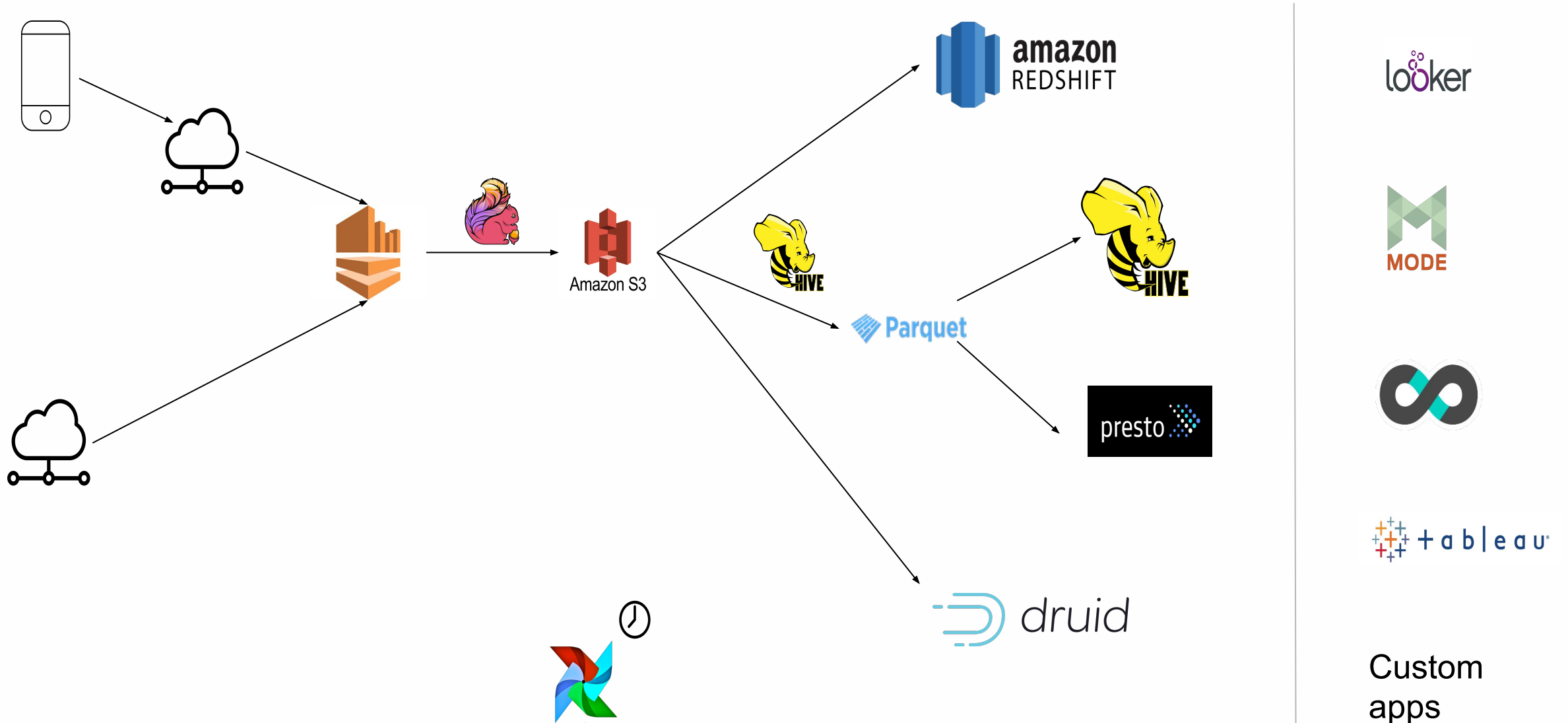
webpack




Redux







Core Infra high level architecture



See details of the data set

Announcements Browse FAQ ? MG

<  **default.event_airflow_task_routed** ☆
Datasets • Hive

   Preview Explore

Description

AirflowTaskRouted tracks the routing information for an Airflow task
[Request Description](#)

Issues

No associated issues

[Report an issue](#)

Date Range

📅 From: May 15, 2020
📅 To: Jul 06, 2020


Tags

+ New



Last Updated

Jul 08, 2020 9pm PDT

Owners

 dp-exp@lyft.com

Frequent Users

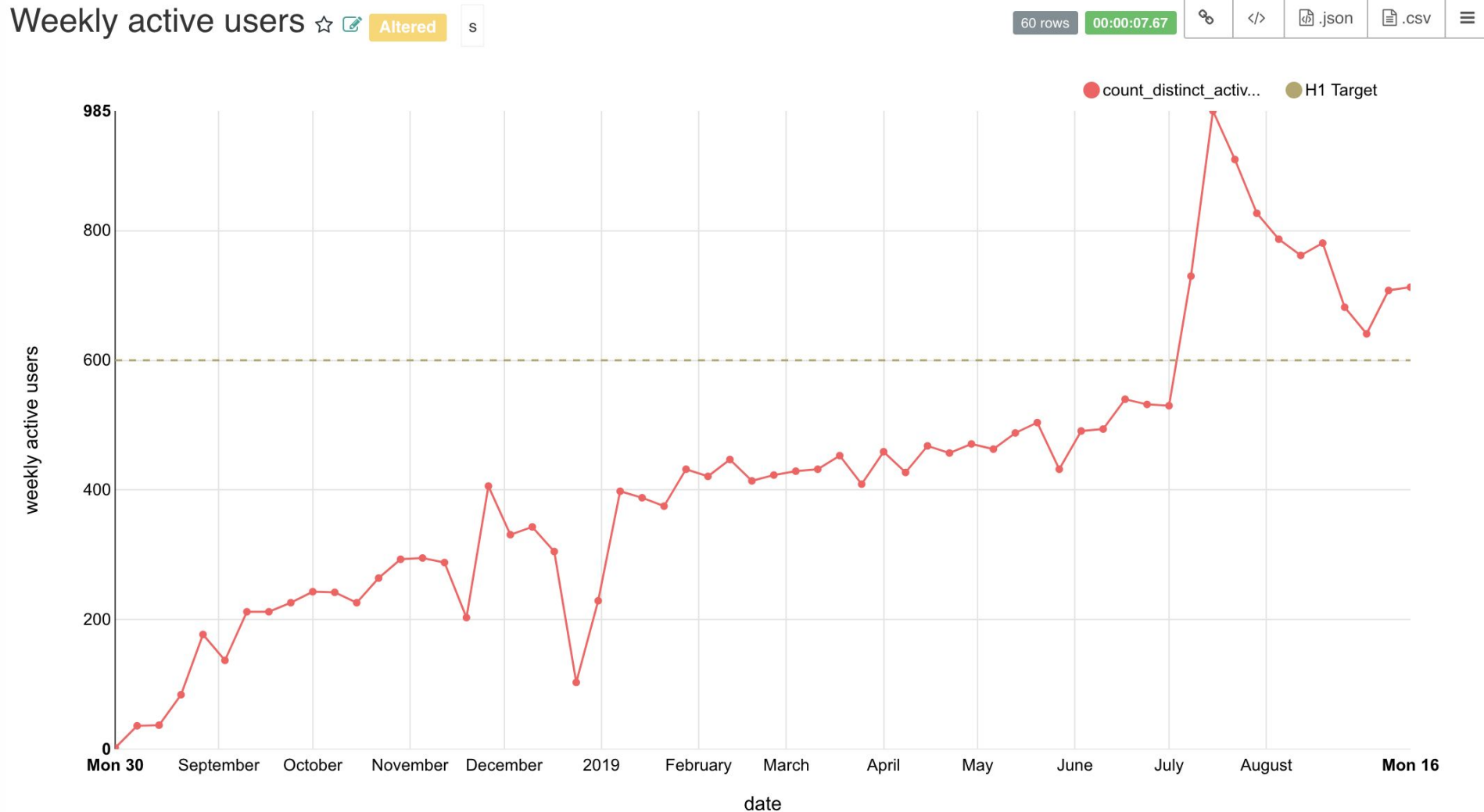
 

Columns (16) **Dashboards (1)**

| | | | |
|------------------------------|---|-----------|---|
| event_id | Unique event identifier. Due to current assumptions in the pipeline, it's important that this be a version 4 (random) UUID. | string | ⋮ |
| ds | | string | ⋮ |
| dag_id | | string | ⋮ |
| execution_date | | string | ⋮ |
| http_request_id | | string | ⋮ |
| occurred_at | Point in time when this event occurred. | timestamp | ⋮ |
| producer_service_name | The name of the service that sent this event | string | ⋮ |
| queue | | string | ⋮ |

Impact

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



“This is God’s work” - George X, ex-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” - Bomee P, DS, Lyft

Roles of Amundsen users at Lyft

Weekly Amundsen user roles (Rolling 7 days) ☆ **Altered**

15 rows

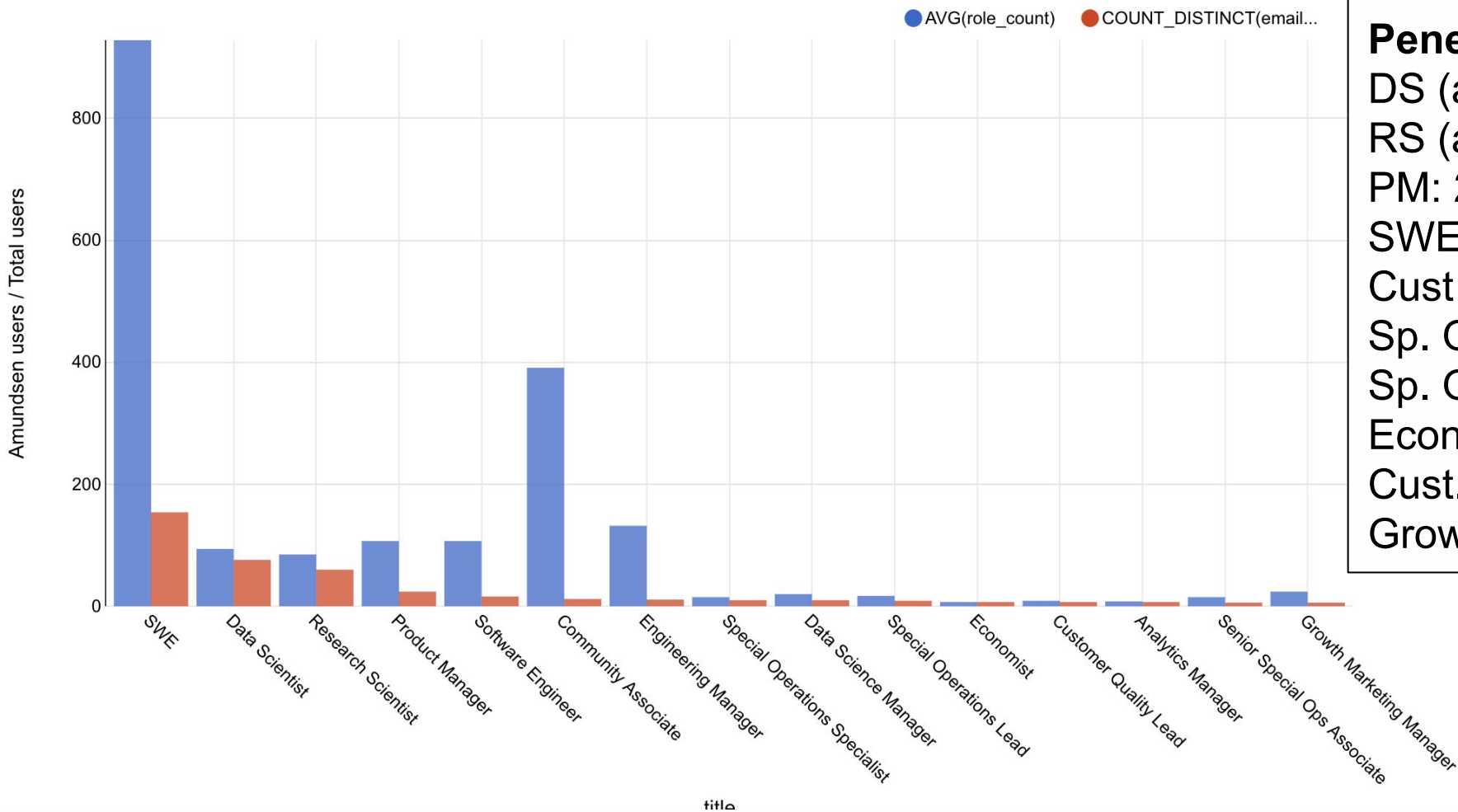
cached

00:00:00.15

</>

.json

.csv



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)

Amundsen Open Source

700

Community
members

150+

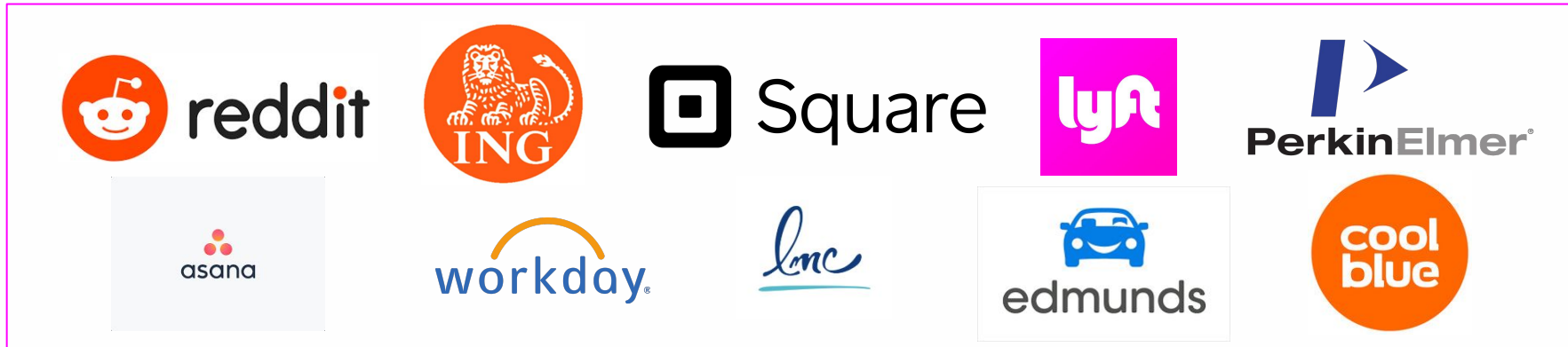
Companies in
the community

20+

Companies using
in production

Amundsen Open Source Community

Prominent users



Active community



Community overview

Contributors



SQUARESPACE

BANG & OLUFSEN



PerkinElmer®

UiO
University of Oslo

intuit.

Remitly

trivadis

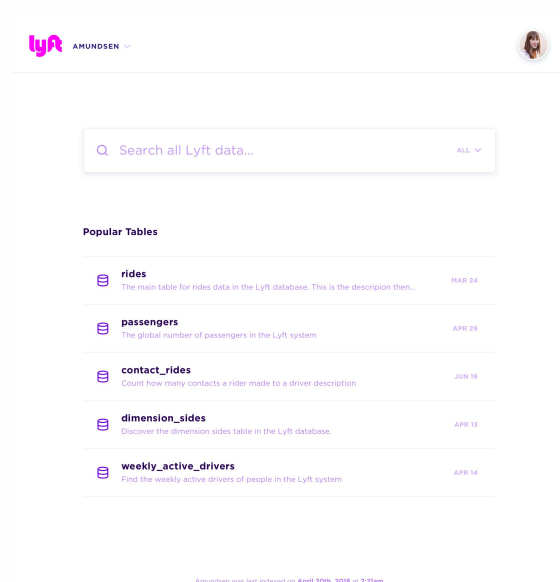
Recent Contributions from the community

- BigQuery integration (Coolblue)
- PostgreSQL and Redshift integration (Everfi)
- Security improvements and Apache Atlas integration (ING)
- Snowflake integration (LMC)
- Toolbar on landing page (In progress, Workday)
- Integrating with Delta analytics platform (In progress, Databricks)
- Talks by ING & Coolblue at conferences in Barcelona, Vilnius & Moscow

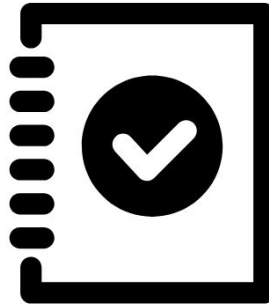
Ongoing trade-offs

Discovery vs. Curation

Discovery



Curation



Guidelines on:

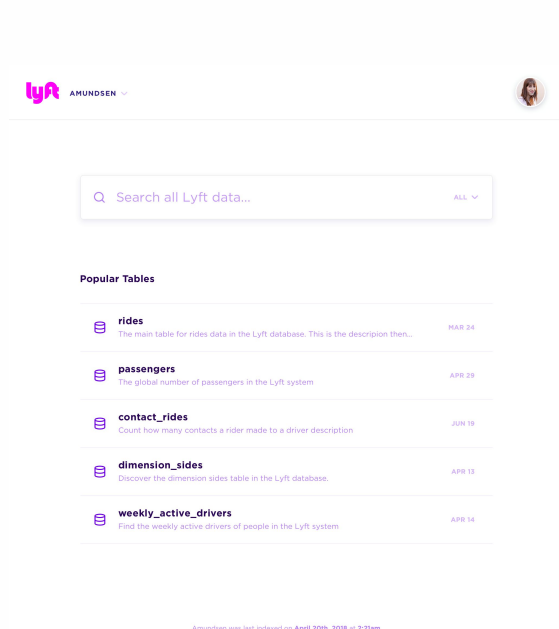
- Where to store data
- How to name tables, dashes, columns, etc.

Challenges

- How much do we push for guidelines vs. just make it discoverable?

Discovery vs. Security

Discovery



Security



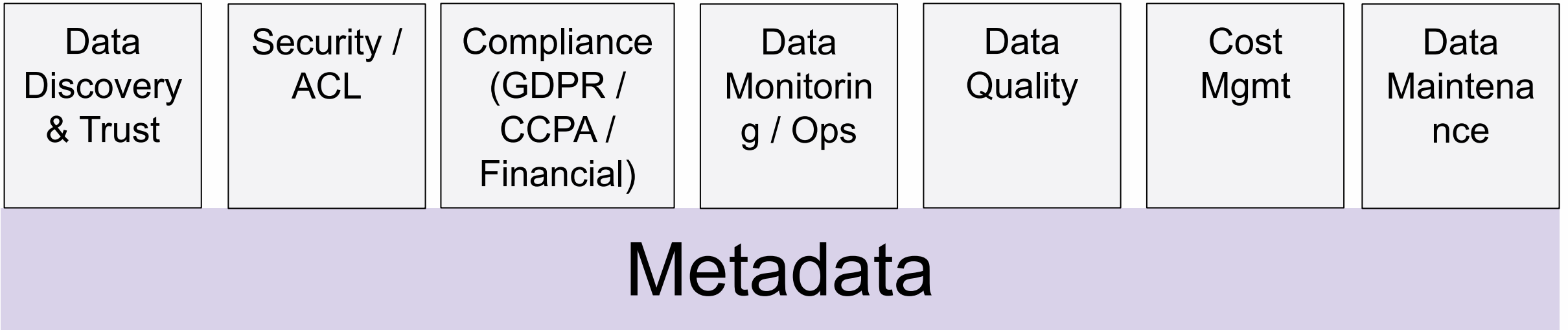
- Provide data & metadata only on a per need basis

Challenges

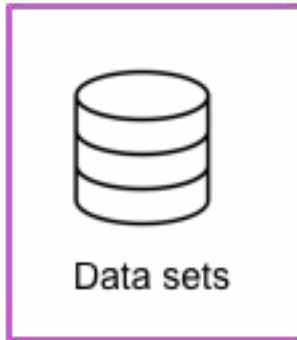
- Do we hide the existence of a data set?
- Do expose metadata regardless of whether the data scientist has access to the data?

Future

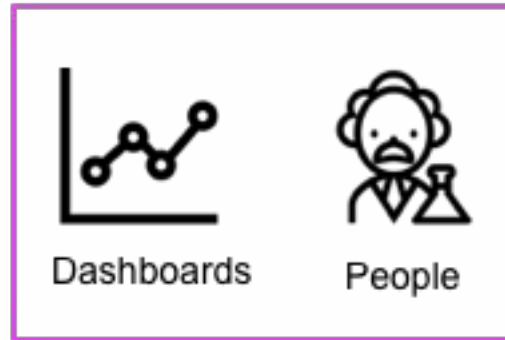
Develop breadth of applications



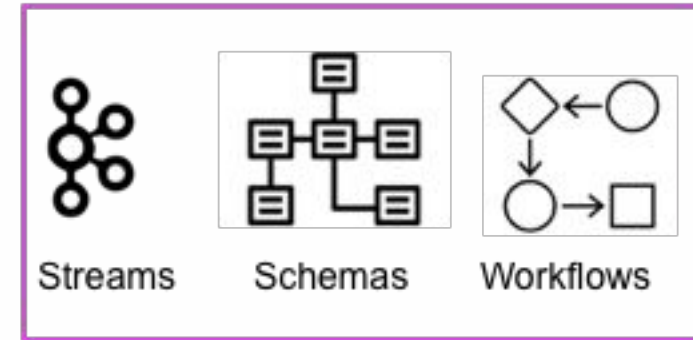
2. Develop depth of metadata



Phase 1
(Complete)



Phase 2
(In development)



Phase 3
(In Scoping)

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.)

Summary

Summary

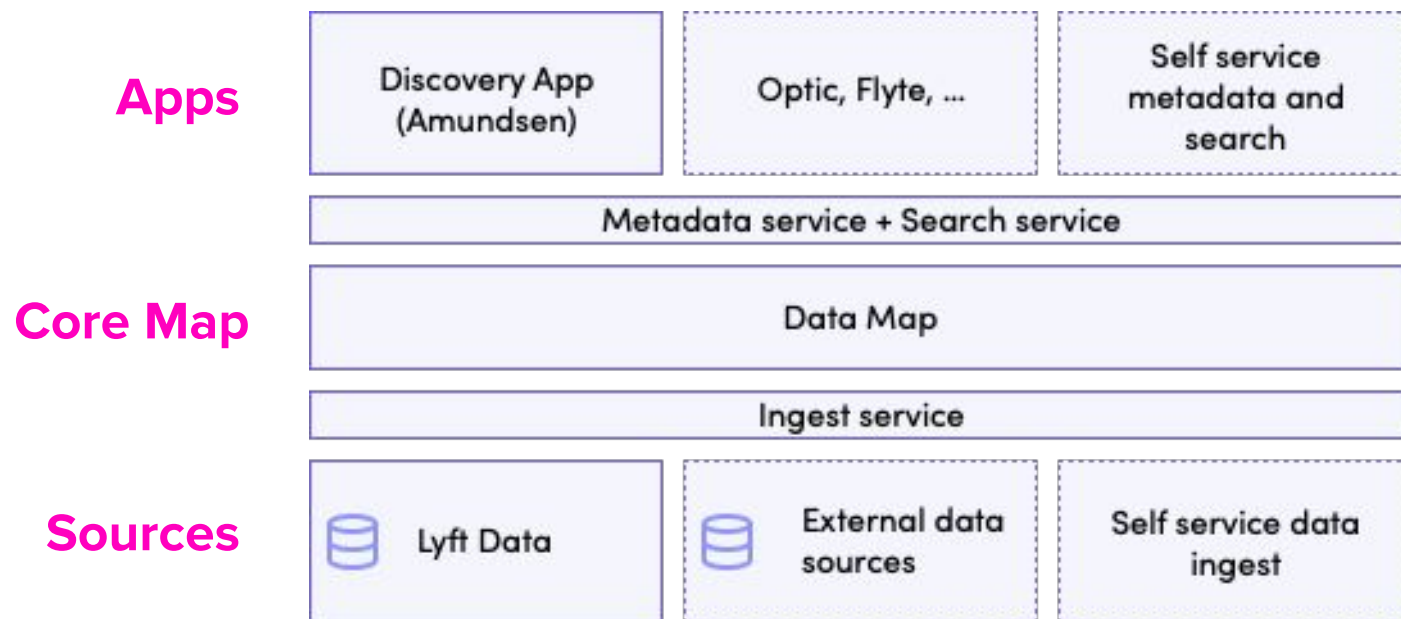
- Data Discovery is a huge pain
- Amundsen helps solve for data discovery
- HUGE opportunity for metadata driven applications

Thanks!

Mark Grover | @mark_grover

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

Amundsen's Architecture



We are building a **rich, comprehensive and actionable map of Lyft's data universe**

Apps are built on top, fuelled by the map. They are easy to build in partnership with product teams, with less and less support from the Data Map team. It can be an UX, like for discovery or an integration with our APIs

Data sources cover all Lyft's data (data storage, processes, users, jobs/tasks...), including payed vendor data. New data sources can be easily pushed to the Map by product teams, with minimal support from the Data Map team

TAC Vote on Project Proposal: Amundsen

Proposed Resolution:

The TAC approves the Amundsen Project as an Incubation project of the LF AI Foundation

Next Steps

LF AI staff will work with Amundsen to onboard the project leading to the announcement of the project joining LF AI

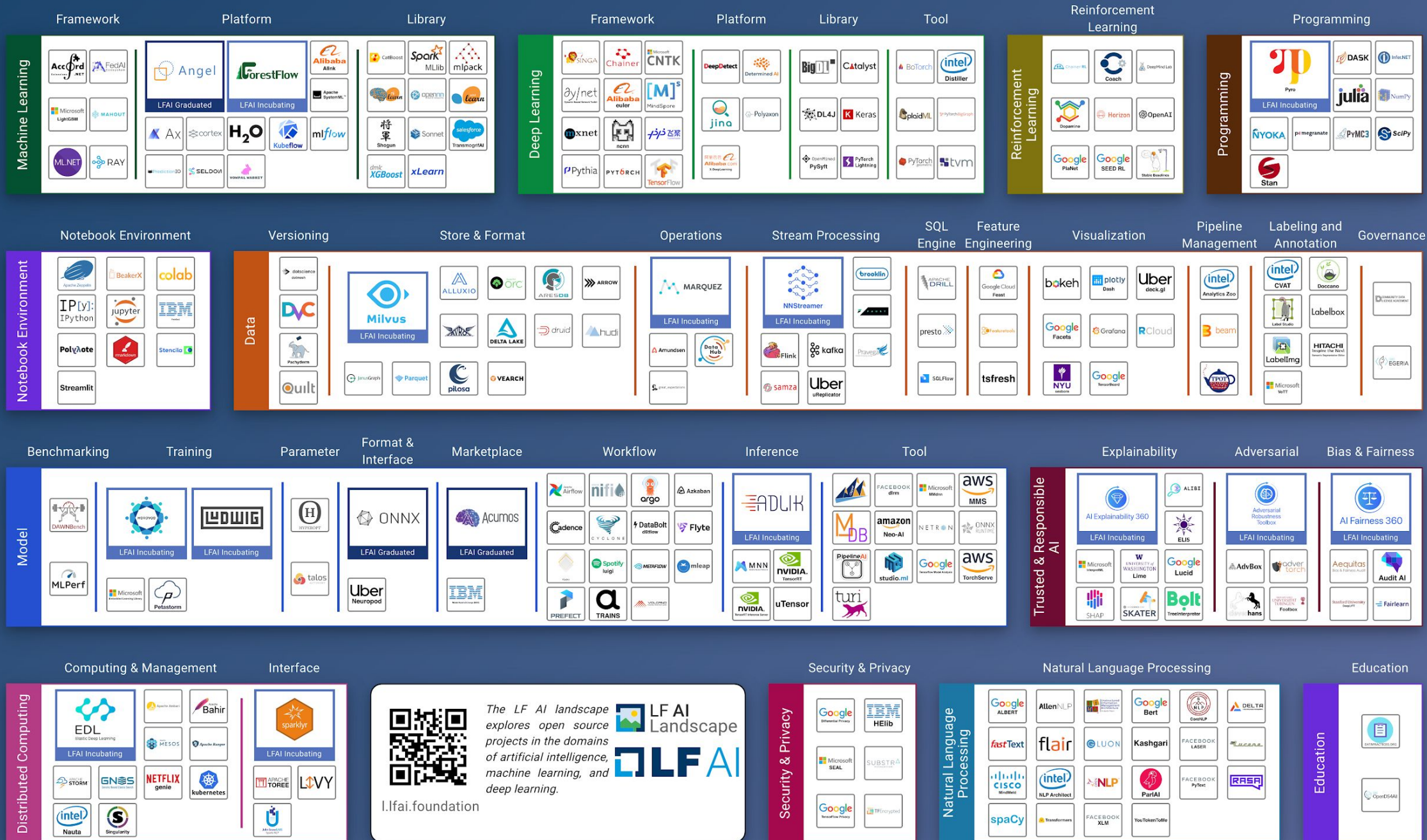
Explore potential integrations between the project and other LF AI projects

Integrate the project with LF AI operations

LF AI General Updates

A Growing Landscape

253 projects



The LF AI landscape explores open source projects in the domains of artificial intelligence, machine learning, and deep learning.

LF AI Landscape

LFAI

lfaifoundation.org

A Growing LF AI Project Portfolio


Graduated LFAI Projects (3)



Acumos ★ 10
LF Artificial Intelligence Foundation



Angel-ML ★ 5,904
LF Artificial Intelligence Foundation




ONNX ★ 8,703
LF Artificial Intelligence Foundation


Incubating LFAI Projects (13)




Adlik ★ 116
LF Artificial Intelligence Foundation



Adversarial Robustness Toolkit ★ 1,599
LF Artificial Intelligence Foundation




AI Explainability 360 Toolkit ★ 589
LF Artificial Intelligence Foundation




AI Fairness 360 Toolkit ★ 1,003
LF Artificial Intelligence Foundation




Elastic Deep Learning (EDL) ★ 88
LF Artificial Intelligence Foundation



ForestFlow ★ 33
LF Artificial Intelligence Foundation



Horovod ★ 9,597
LF Artificial Intelligence Foundation




Ludwig ★ 6,848
LF Artificial Intelligence Foundation




Marquez ★ 322
LF Artificial Intelligence Foundation



Milvus ★ 3,713
LF Artificial Intelligence Foundation



NNStreamer ★ 247
LF Artificial Intelligence Foundation



Pyro ★ 6,321
LF Artificial Intelligence Foundation



sparklyr ★ 745
LF Artificial Intelligence Foundation

A Growing Developer Community

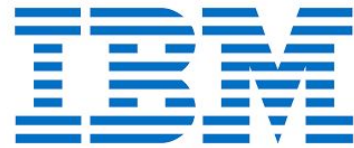
| |
|-----------------------------------|
| Acumos |
| Adlik |
| Adversarial Robustness 360 |
| AI Explainability 360 |
| AI Fairness 360 |
| Angel |
| EDL |
| ForestFlow |
| Horovod |
| Ludwig |
| Marquez |
| Milvus |
| nnstreamer |
| ONNX |
| Pyro |
| Sparklyr |

You are currently tracking **7,310,710 lines of code**, committed by **1,119 developers**, from **46 known organizations**, working in **76 repos**, on **16 projects** over the last **6 years, 6 months, and 14 days**.

Companies hosting projects in LF AI



FACEBOOK



Looking to host a project with LF AI

Hosted project stages and life cycle:

<https://lfai.foundation/project-stages-and-lifecycle/>

Offered services for hosted projects:

<https://lfai.foundation/services-for-projects/>

Contact:

Jim Spohrer (TAC Chair) and Ibrahim Haddad (ED, LF AI)

Promoting Upcoming Project Releases

We promote project releases via a blog post and on LF AI [Twitter](#) and/or [LinkedIn](#) social channels

For links to details on upcoming releases for LF AI hosted projects visit the [Technical Project Releases wiki](#)

If you are an LF AI hosted project and would like LF AI to promote your release, reach out to pr@lfai.foundation to coordinate in advance (min 2 wks) of your expected release date.

Project Graduation Opportunities

To be scheduled on a future TAC call

| | Joined incubation | Unique contributors | Involved companies | GH Stars | Flow of commits | CII badge | Collaboration with other LF AI project | TSC Members |
|----------------|-------------------|---------------------|---|----------|-----------------|-----------|--|--|
| Horovod | Dec 2018 | 90 | IBM MSFT FB Uber Nvidia Databricks AMD Intel H2O.ai UBC Independent | 9.5 K | ✓ | ✓ | ✓ | Uber Amazon IBM NVidia Microsoft Intel Independent |

Note on quorum

As LF AI is growing, we now have 12 voting members on the TAC.

TAC representative - please ensure you attend the bi-weekly calls or email Jacqueline/Ibrahim to designate an alternate representative when you can not make it.

We need to ensure quorum on the calls especially when we have items to vote on.

Updates from the Outreach Committee

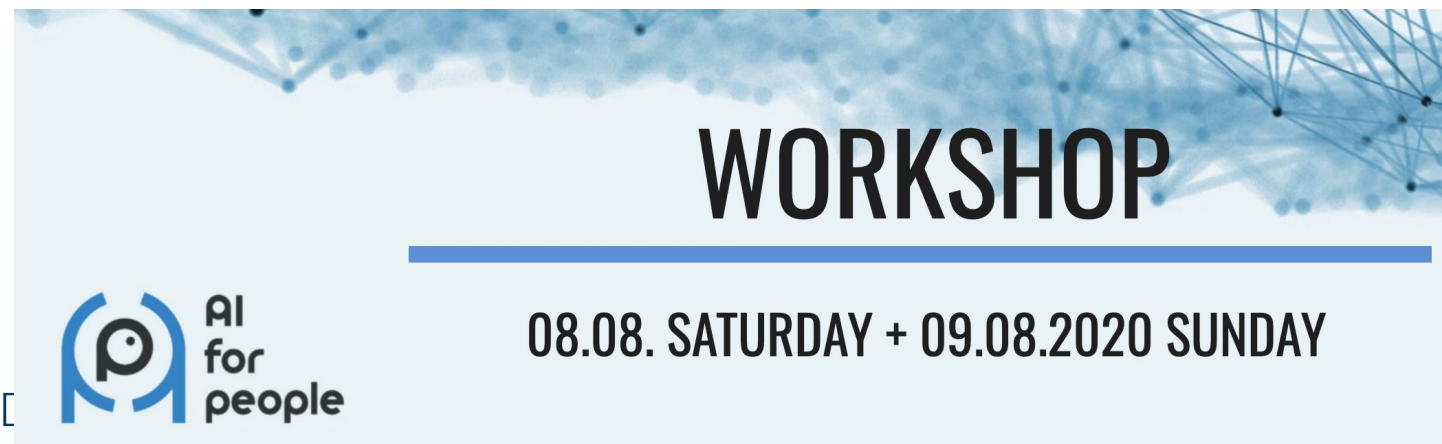
Events

- › Upcoming Events
 - › Visit the [LF AI Events Calendar](#) or the [LF AI 2020 Events wiki](#) for a list of all events
 - › To participate visit the [LF AI 2020 Events wiki page](#) or email info@lfai.foundation
- › Please consider holding virtual events
 - › To discuss participation, please email events@lfai.foundation

Upcoming Events




CLOUD NATIVE + OPEN SOURCE
.....
Virtual Summit China 2020
.....



WORKSHOP

08.08. SATURDAY + 09.08.2020 SUNDAY



AI
for
people

LF AI PR/Comms

- › Please follow LF AI on [Twitter](#) & [LinkedIn](#) and help amplify news via your social networks - Please retweet and share!
 - › Also watch for news updates via the tac-general mail list
 - › View recent announcement on the [LF AI Blog](#)
- › Open call to publish project/committee updates or other relevant content on the [LF AI Blog](#)
- › To discuss more details on participation or upcoming announcements, please email pr@lfai.foundation

Call to Participate in Ongoing Efforts

Trusted AI

- › **Leadership:**
Animesh Singh (IBM), Souad Ouali (Orange), and Jeff Cao (Tencent)
- › **Goal:** Create policies, guidelines, tooling and use cases by industry
- › **Github:**
<https://github.com/lfai/trusted-ai>
- › **Wiki:**
<https://wiki.lfai.foundation/display/DL/Trusted+AI+Committee>
- › **To participate:**
<https://lists.lfai.foundation/g/trustedai-committee/>
- › **Next call:** Bi-weekly on Thursdays at 7am PT, subscribe to group calendar on wiki
<https://wiki.lfai.foundation/pages/viewpage.action?pageId=12091895>

ML Workflow & Interop

- › **Leadership:**
Huang “Howard” Zhipeng (Huawei)
- › **Goal:**
Define an ML Workflow and promote cross project integration
- › **Wiki:**
<https://wiki.lfai.foundation/display/DL/ML+Workflow+Committee>
- › **To participate:**
<https://lists.lfai.foundation/g/mlworkflow-committee>
- › **Next call:** Every 4 weeks on Thursdays at 7:00 am PT, subscribe to group calendar on wiki
<https://wiki.lfai.foundation/pages/viewpage.action?pageId=18481242>

Launching an effort to create AI Ethics Training

Initial developed course by the LF: Ethics in AI and Big Data - published on edX platform:
<https://www.edx.org/course/ethics-in-ai-and-big-data>

The goal is to build 2 more modules and package all 3 as a professional certificate - a requirement for edX

- › The LF would cover the cost of the production and promotion
- › The course would be offered for free
- › The credit of the course will go to content creator and their organizations
- › Initial interested parties: IBM, AI for People, Montreal AI Ethics Institute, Ethical ML Institute
- › **To participate:**
<https://lists.lfai.foundation/g/aiethics-training>

Upcoming TAC Meetings

Upcoming TAC Meetings

- › **August 13:** Guest Presentations
 - › [OpenPower](#) - James Kulina (Executive Director)
 - › [ModelDB](#) - Conrado Silva Miranda (Verta.ai)

- › **August 24:** To be announced

Please send agenda topic requests to tac-general@lists.lfai.foundation

TAC Meeting Details

- › To subscribe to the TAC Group Calendar, visit the wiki: <https://wiki.lfai.foundation/x/XQB2>
- › Join from PC, Mac, Linux, iOS or Android: <https://zoom.us/j/430697670>
- › Or iPhone one-tap:
 - › US: +16465588656,,430697670# or +16699006833,,430697670#
- › Or Telephone:
 - › Dial(for higher quality, dial a number based on your current location):
 - › US: +1 646 558 8656 or +1 669 900 6833 or +1 855 880 1246 (Toll Free) or +1 877 369 0926 (Toll Free)
- › Meeting ID: 430 697 670
- › International numbers available: <https://zoom.us/u/achYtcw7uN>

Open Discussion

Legal Notices

- › The Linux Foundation, The Linux Foundation logos, and other marks that may be used herein are owned by The Linux Foundation or its affiliated entities, and are subject to The Linux Foundation's Trademark Usage Policy at <https://www.linuxfoundation.org/trademark-usage>, as may be modified from time to time.
- › Linux is a registered trademark of Linus Torvalds. Please see the Linux Mark Institute's trademark usage page at <https://lmi.linuxfoundation.org> for details regarding use of this trademark.
- › Some marks that may be used herein are owned by projects operating as separately incorporated entities managed by The Linux Foundation, and have their own trademarks, policies and usage guidelines.
- › TWITTER, TWEET, RETWEET and the Twitter logo are trademarks of Twitter, Inc. or its affiliates.
- › Facebook and the "f" logo are trademarks of Facebook or its affiliates.
- › LinkedIn, the LinkedIn logo, the IN logo and InMail are registered trademarks or trademarks of LinkedIn Corporation and its affiliates in the United States and/or other countries.
- › YouTube and the YouTube icon are trademarks of YouTube or its affiliates.
- › All other trademarks are the property of their respective owners. Use of such marks herein does not represent affiliation with or authorization, sponsorship or approval by such owners unless otherwise expressly specified.
- › The Linux Foundation is subject to other policies, including without limitation its Privacy Policy at <https://www.linuxfoundation.org/privacy> and its Antitrust Policy at <https://www.linuxfoundation.org/antitrust-policy>, each as may be modified from time to time. More information about The Linux Foundation's policies is available at <https://www.linuxfoundation.org>.
- › Please email legal@linuxfoundation.org with any questions about The Linux Foundation's policies or the notices set forth on this slide.