Disrupting Data Discovery

Saturday, October 26th 2019
Alagappan Sethuraman | Engineering Manager, Lyft
Daniel Won | Software Engineer, Lyft

Agenda

- What is Data Discovery?
- Challenges in Data Discovery
- Introducing Amundsen
- Amundsen Architecture
- Impact and Future Work

What is Data Discovery?

Data is used to make informed decisions

Analysts

Data Scientists

Product Managers

General Managers

Engineers

Experimenters

Make data the heart of every decision

Data-driven decision making process:

1. Search & find data
2. Understand the data
3. Perform an analysis/visualization
4. Share insights and/or make a decision

What is Data Discovery?

Consider a data-driven decision making process:

1. Search & find data
2. Understand the data

Data Discovery

3. Perform an analysis/create a visualization
4. Share insights and/or make a decision

Challenges in Data Discovery

Hi! I’m a new Analyst!

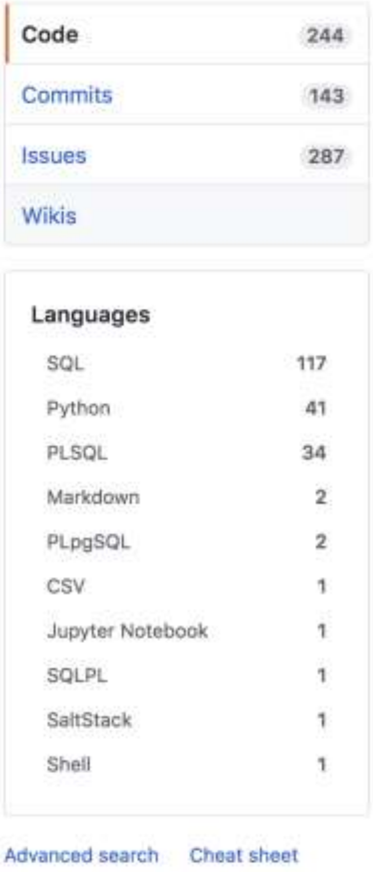
- My first project is predict the attendance for IDEAS conference
- Goal: Help the office team make a decision on number of chairs to provide?
- Idea: Let’s take a look into attendance from previous conferences… but where do I look?

Step 1: Search & find data

- Ask a friend/manager/coworker
- Ask in a wider Slack channel
- Search in the Github repos

We end up finding tables: `hosted_events`

that seems to be the right one



Code	244
Commits	143
Issues	287
Wikis	

Languages

SQL	117
Python	41
PLSQL	34
Markdown	2
PLpgSQL	2
CSV	1
Jupyter Notebook	1
SQLPL	1
SaltStack	1
Shell	1

Advanced search Cheat sheet

Step 2: Understand the data

- You find several columns that might be what you're looking for:
  - `booked`, `registered`, and `attendance`
- But you still have many questions such as:
  - Does `attendance` include staff?
  - What's the difference between `booked` and `registered?`
  - How accurate are these figures?

Step 2: Understand the data

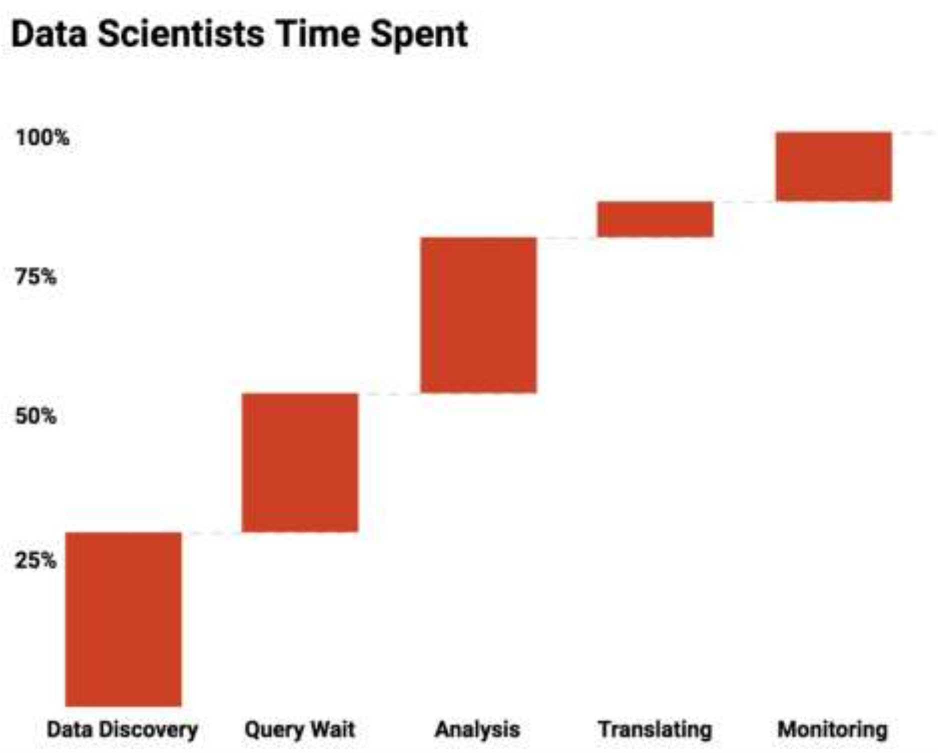
- Look for further documentation on these columns
  - Where does this documentation live?
- Ask an expert who knows this table
  - Who is an expert?
- Run some queries to try to figure it out at the risk of being wrong

```
SELECT * FROM schema.host_events

LIMIT 100;
```

Nearly 1/3 of Data Scientist time is spent in Data Discovery

- **Data discovery is a problem** because of the lack of understanding of what data exists, where, who owns it, & how to use it.
- Data Discovery provides little to no intrinsic value
- **Impactful work happens in Analysis**

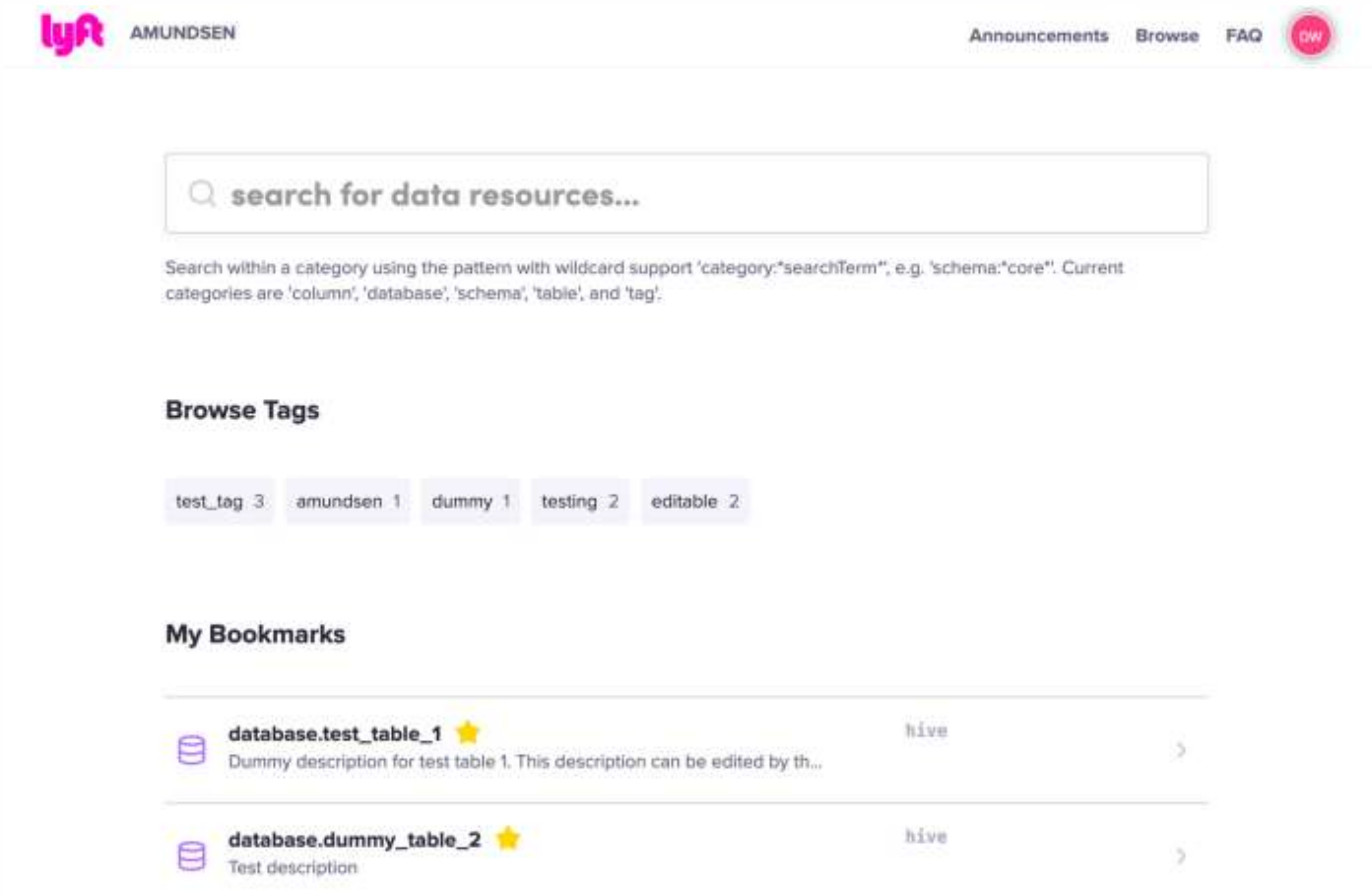


Introducing Amundsen

What is Amundsen?

- Built at Lyft, official launch in late 2018
- Inspired by Google Search, Airbnb Data Portal, and Apache Gobblin
- Named after Norwegian explorer Roald Amundsen
  - Led the first expedition to the South Pole
  - Led the first expedition through the Northwest Passage

Home Page



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

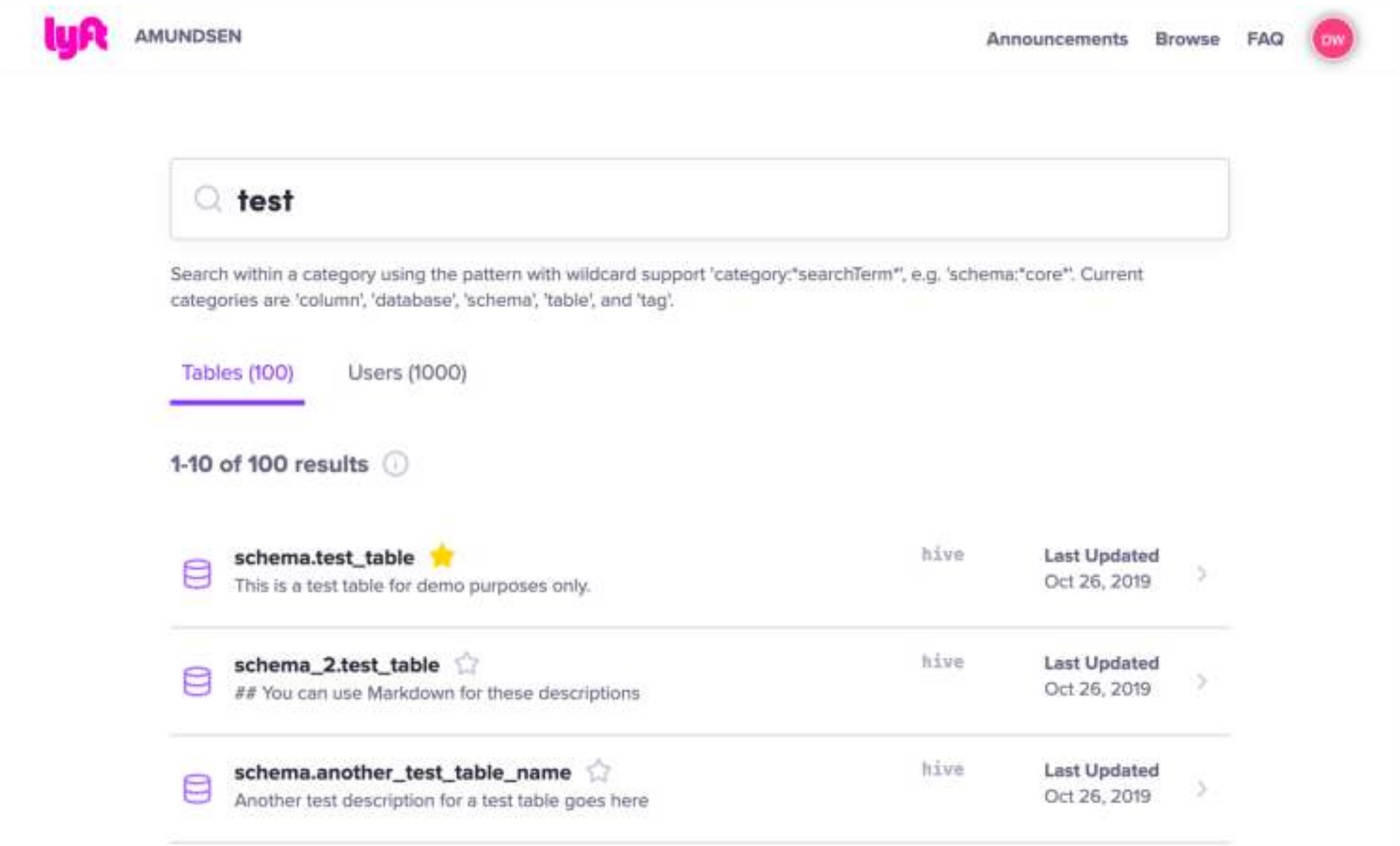
test_tag 3 amundsen 1 dummy 1 testing 2 editable 2

My Bookmarks

database.test_table_1 ★ hive
Dummy description for test table 1. This description can be edited by th...

database.dummy_table_2 ★ hive
Test description

Search



test

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'.

Tables (100) Users (1000)

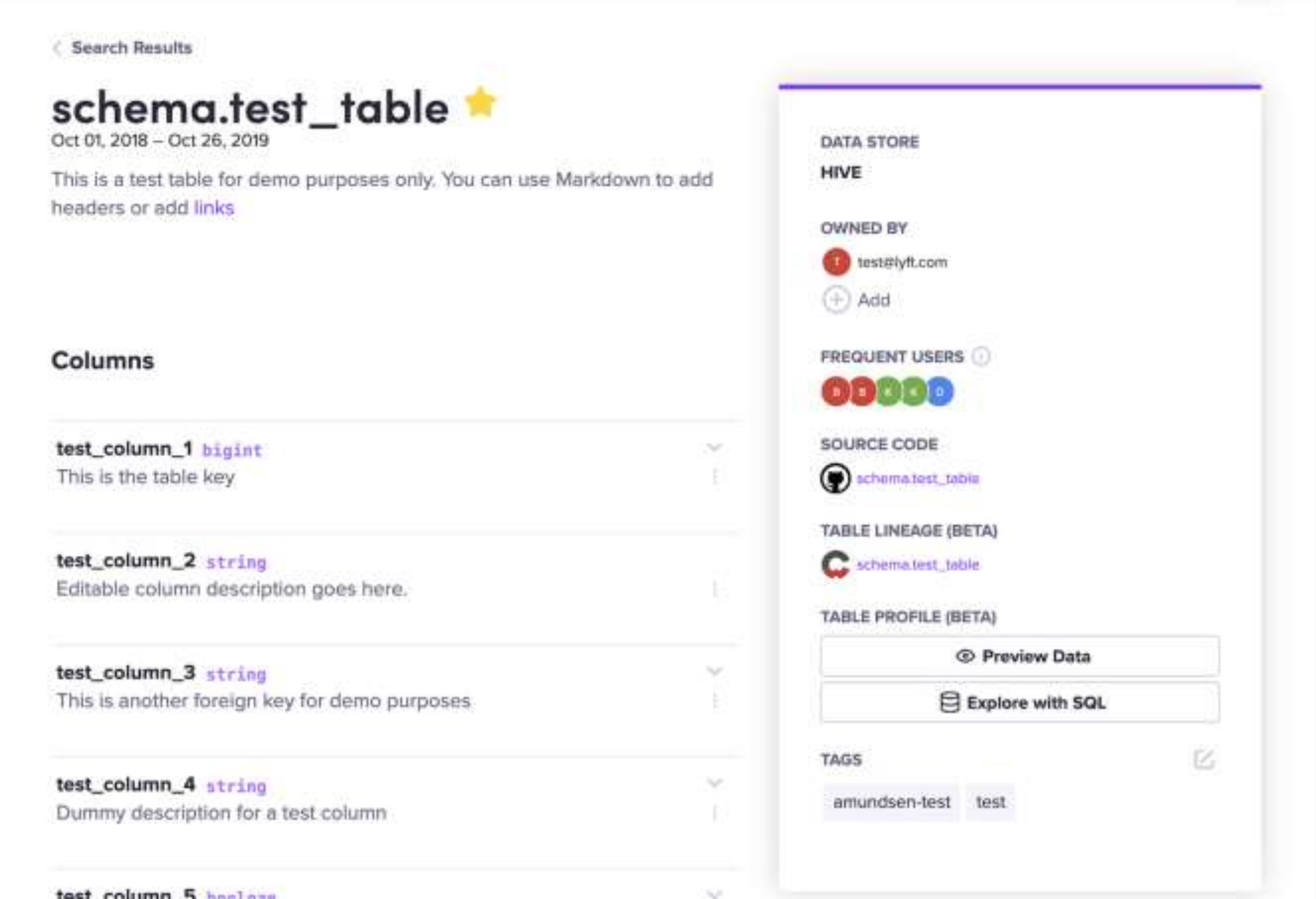
1-10 of 100 results

schema.test_table ★
This is a test table for demo purposes only.
hive
Last Updated
Oct 26, 2019

schema_2.test_table ☆
You can use Markdown for these descriptions
hive
Last Updated
Oct 26, 2019

schema.another_test_table_name ☆
Another test description for a test table goes here
hive
Last Updated
Oct 26, 2019

Resource Metadata



‹ Search Results

schema.test_table ★

Oct 01, 2018 – Oct 26, 2019

This is a test table for demo purposes only. You can use Markdown to add headers or add links

Columns

test_column_1 `bigint`
This is the table key

test_column_2 `string`
Editable column description goes here.

test_column_3 `string`
This is another foreign key for demo purposes

test_column_4 `string`
Dummy description for a test column

test_column_5 `boolean`

DATA STORE
HIVE

OWNED BY
test@lyft.com
Add

FREQUENT USERS

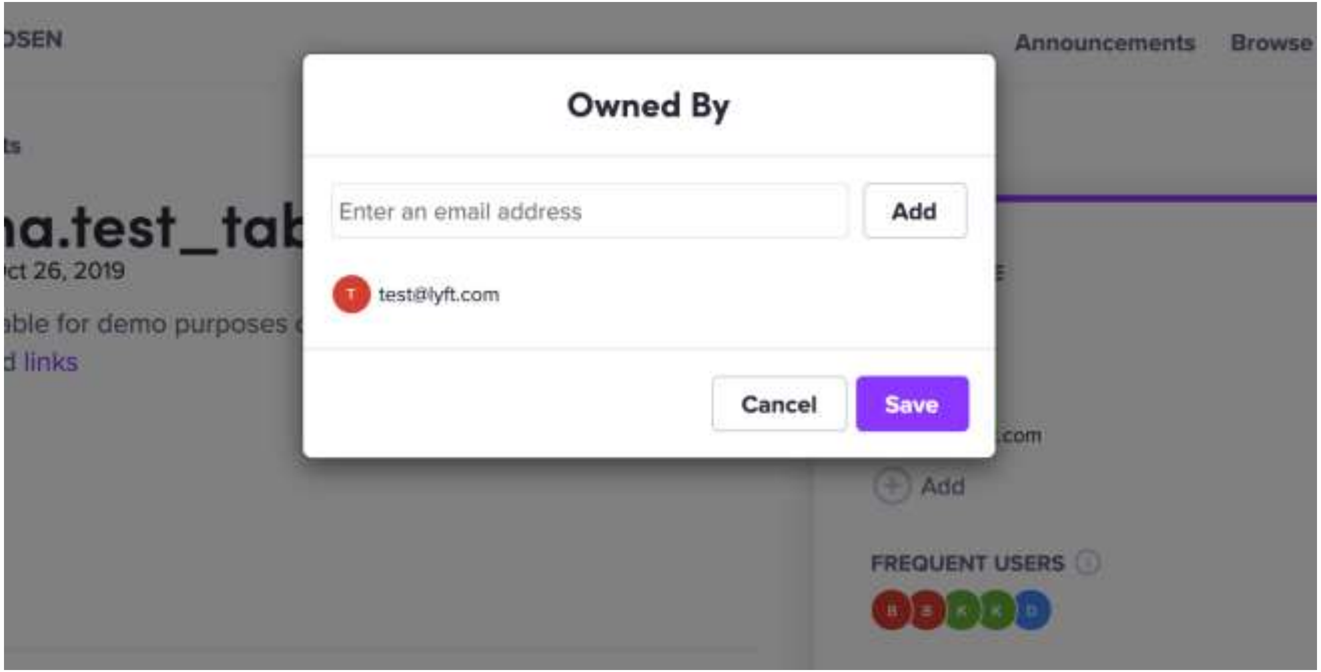
SOURCE CODE
schema.test_table

TABLE LINEAGE (BETA)
schema.test_table

TABLE PROFILE (BETA)
Preview Data
Explore with SQL

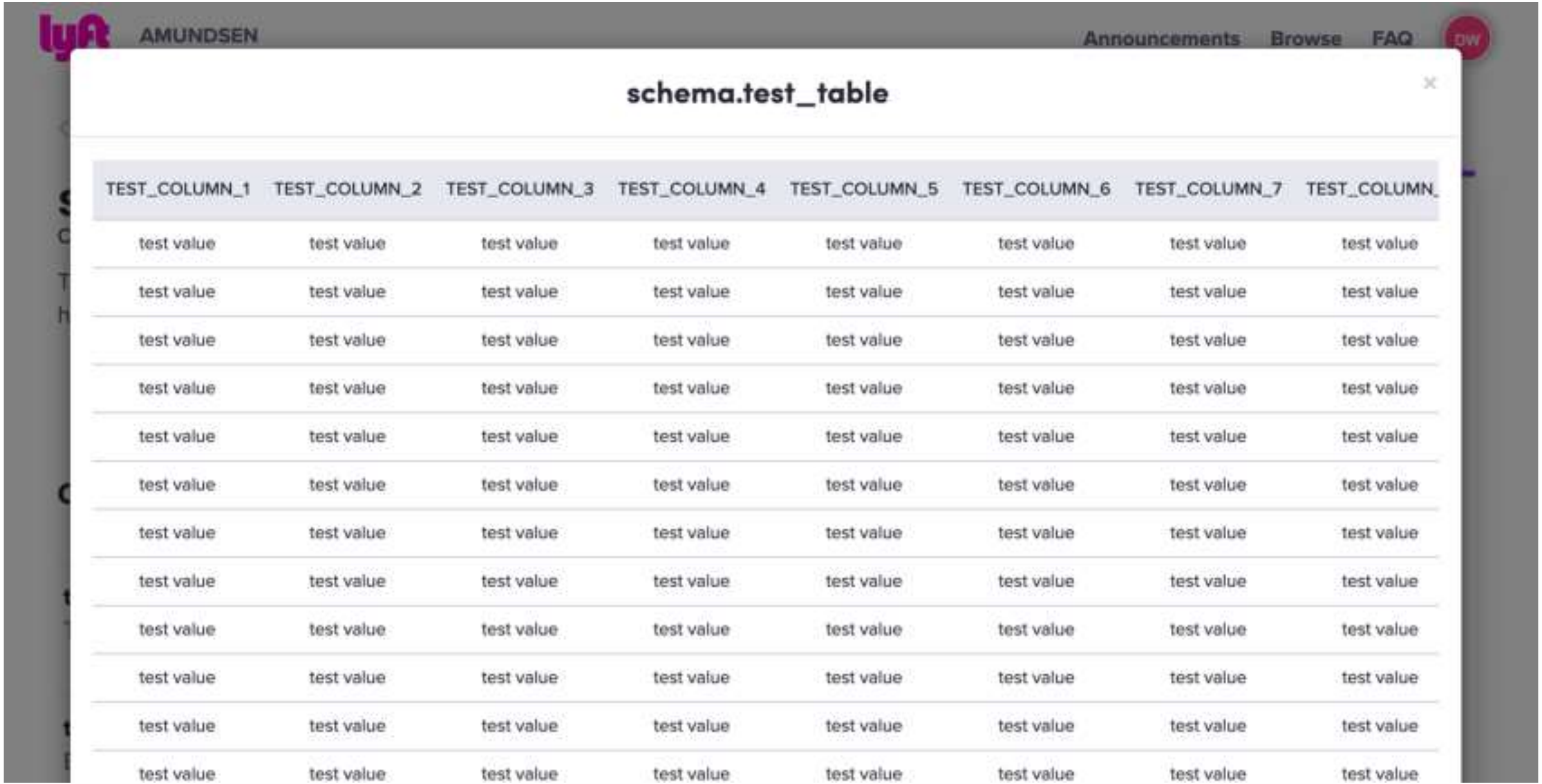
TAGS
amundsen-test
test

Resource Ownership



test@lyft.com

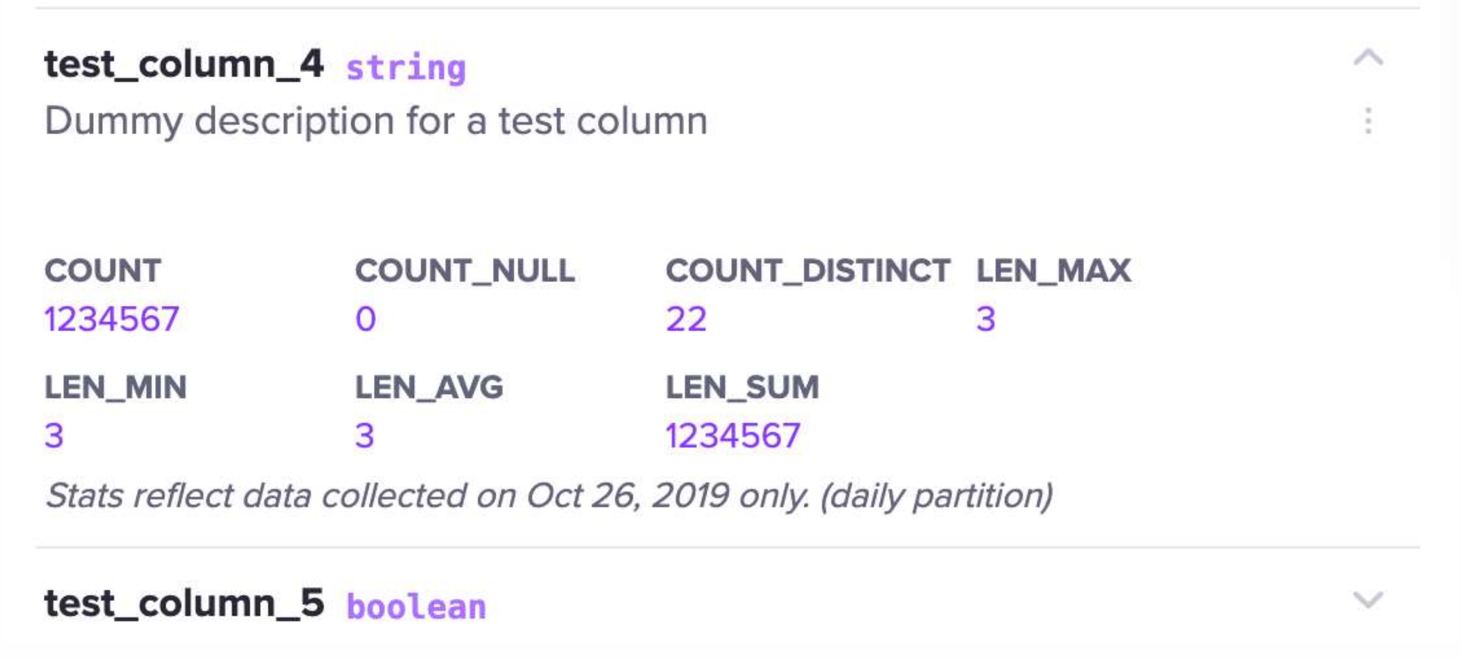
Data Preview



schema.test_table

TEST_COLUMN_1	TEST_COLUMN_2	TEST_COLUMN_3	TEST_COLUMN_4	TEST_COLUMN_5	TEST_COLUMN_6	TEST_COLUMN_7	TEST_COLUMN_
test value	test value	test value	test value	test value	test value	test value	test value
test value	test value	test value	test value	test value	test value	test value	test value
test value	test value	test value	test value	test value	test value	test value	test value
test value	test value	test value	test value	test value	test value	test value	test value
test value	test value	test value	test value	test value	test value	test value	test value
test value	test value	test value	test value	test value	test value	test value	test value
test value	test value	test value	test value	test value	test value	test value	test value
test value	test value	test value	test value	test value	test value	test value	test value
test value	test value	test value	test value	test value	test value	test value	test value
test value	test value	test value	test value	test value	test value	test value	test value
test value	test value	test value	test value	test value	test value	test value	test value
test value	test value	test value	test value	test value	test value	test value	test value

Computed Column Statistics



test_column_4 `string`

Dummy description for a test column

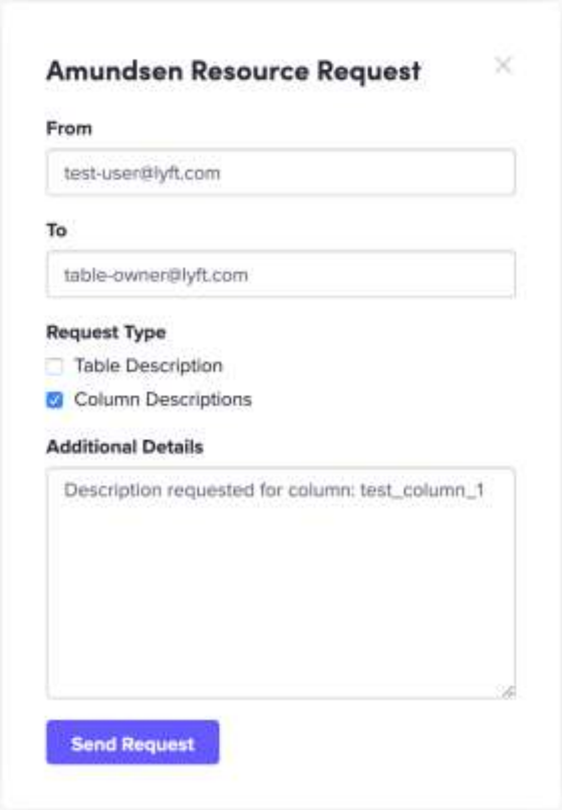
COUNT	COUNT_NULL	COUNT_DISTINCT	LEN_MAX
1234567	0	22	3
LEN_MIN	LEN_AVG	LEN_SUM	
3	3	1234567	

Stats reflect data collected on Oct 26, 2019 only. (daily partition)

test_column_5 `boolean`

Disclaimer: these stats are arbitrary.

Requesting Descriptions



Amundsen Resource Request

From

test-user@lyft.com

To

table-owner@lyft.com

Request Type

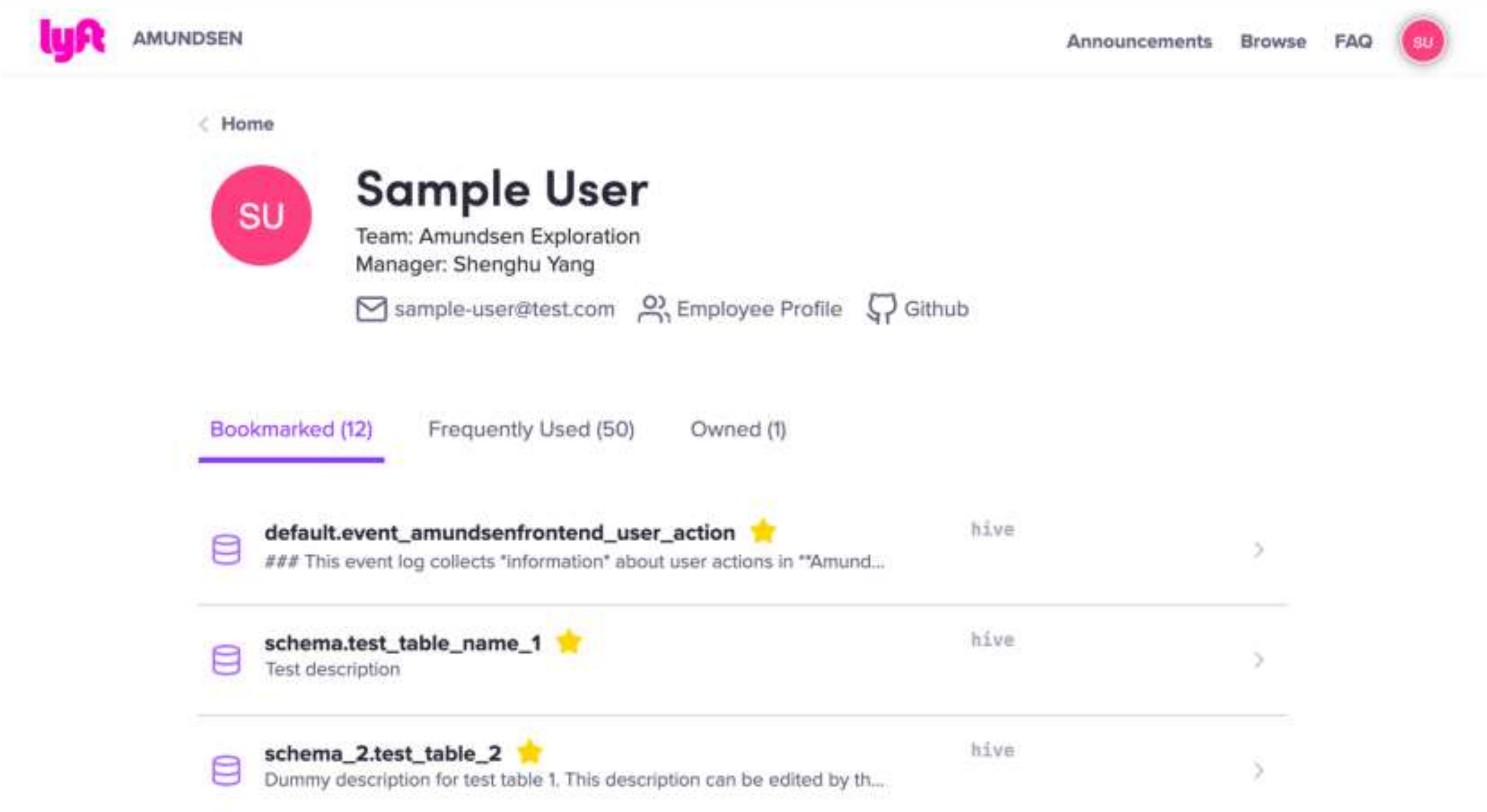
□ Table Description

☑ Column Descriptions

Additional Details

Description requested for column: test_column_1

User Profile



Sample User

Team: Amundsen Exploration
Manager: Shenghu Yang

sample-user@test.com Employee Profile Github

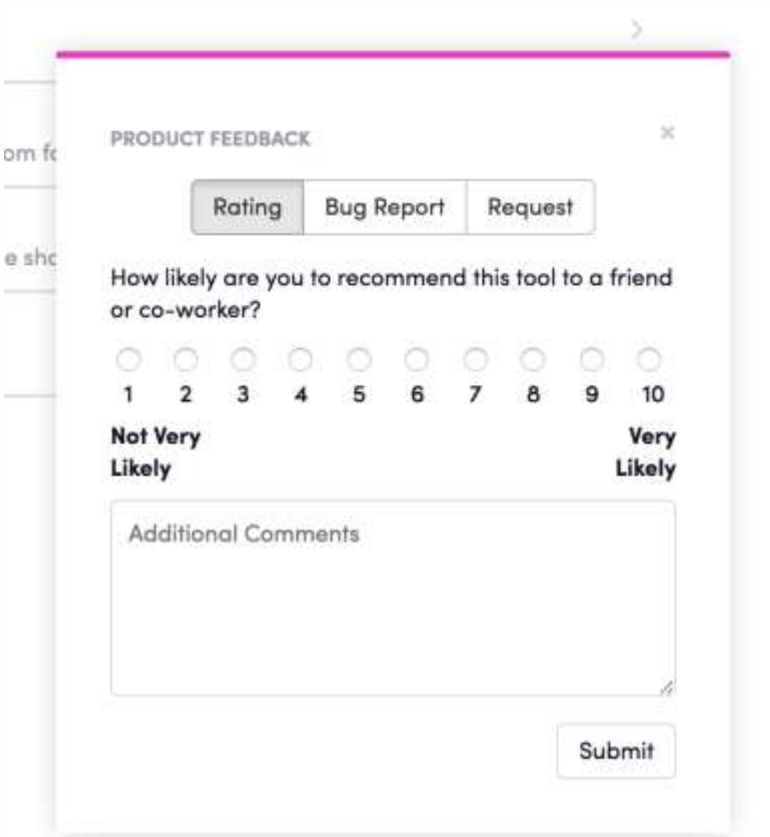
Bookmarked (12) Frequently Used (50) Owned (1)

default.event_amundsenfrontend_user_action ★ hive
This event log collects *information* about user actions in **Amund...

schema.test_table_name_1 ★ hive
Test description

schema_2.test_table_2 ★ hive
Dummy description for test table 1. This description can be edited by th...

In-Application User Feedback



PRODUCT FEEDBACK

How likely are you to recommend this tool to a friend or co-worker?

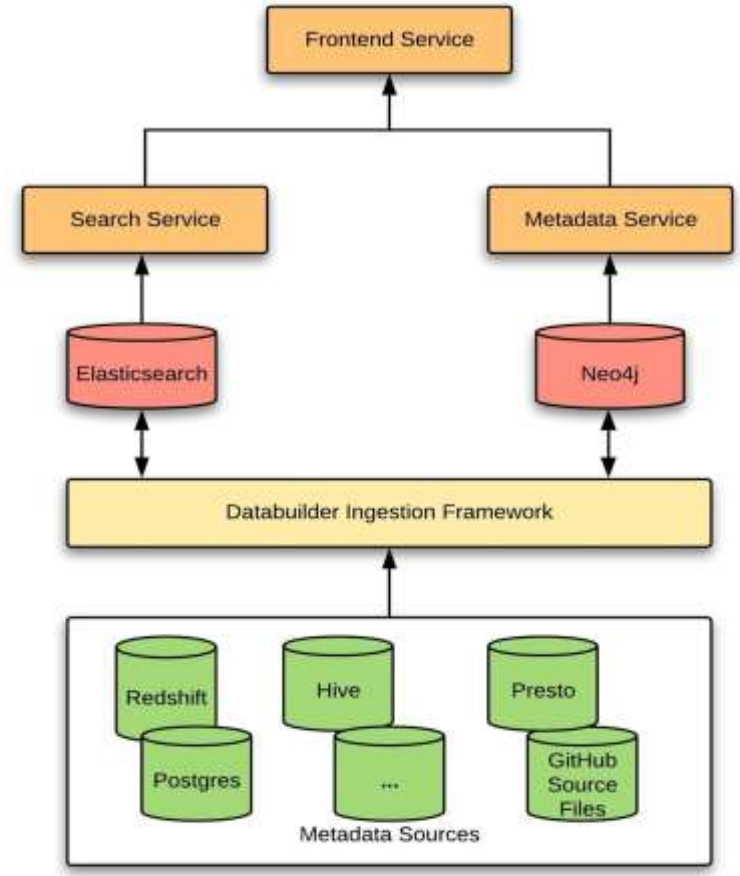
○	○	○	○	○	○	○	○	○	○
1	2	3	4	5	6	7	8	9	10

Not Very Likely **Very Likely**

Additional Comments

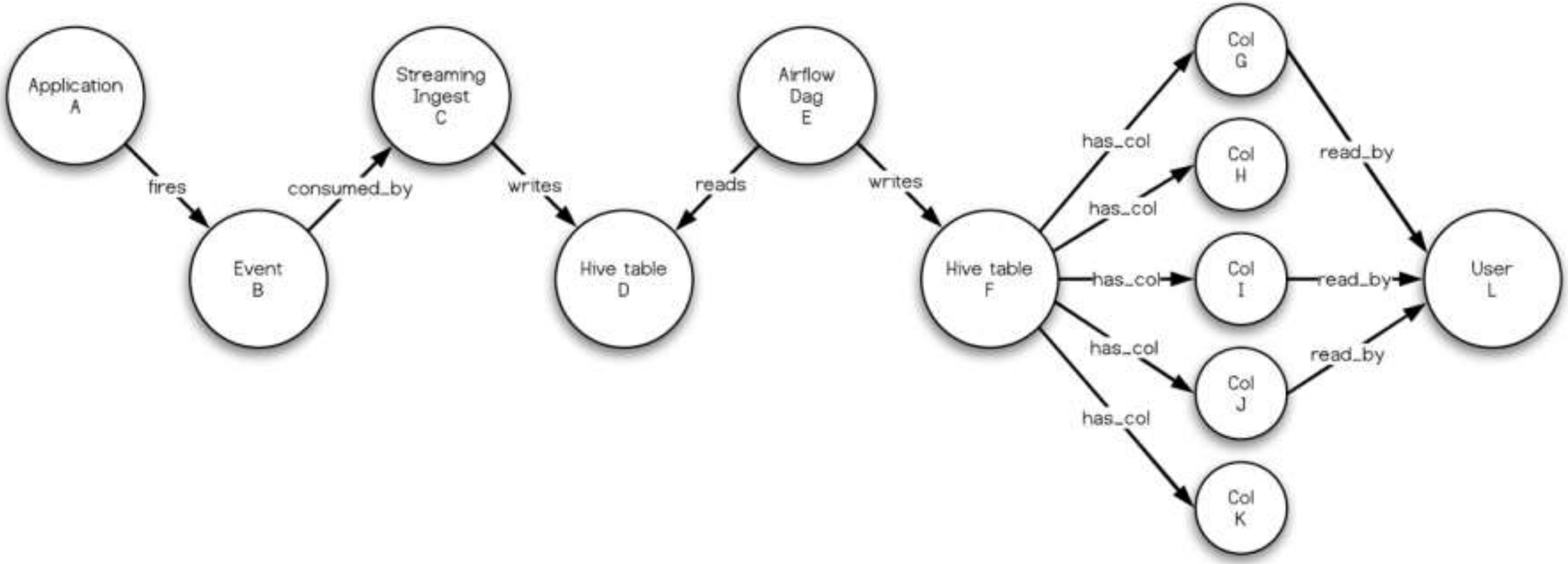
Amundsen Architecture

Amundsen Architecture

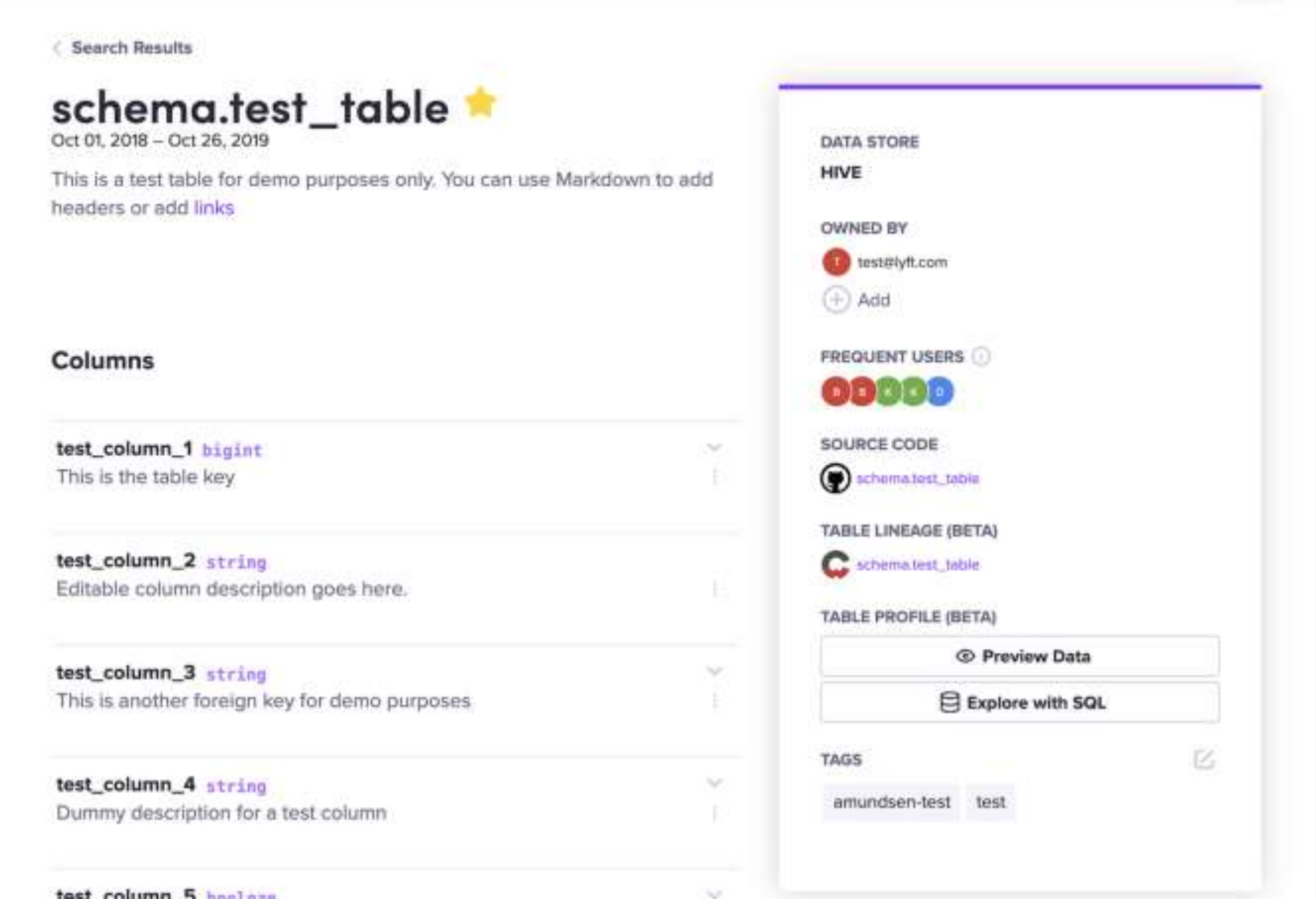


Why choose a graph database?

Why Graph database? (1/2)



View Resource Metadata



‹ Search Results

schema.test_table ★

Oct 01, 2018 – Oct 26, 2019

This is a test table for demo purposes only. You can use Markdown to add headers or add links

Columns

test_column_1 `bigint`
This is the table key

test_column_2 `string`
Editable column description goes here.

test_column_3 `string`
This is another foreign key for demo purposes

test_column_4 `string`
Dummy description for a test column

test_column_5 `boolean`

DATA STORE
HIVE

OWNED BY
test@lyft.com
⊕ Add

FREQUENT USERS

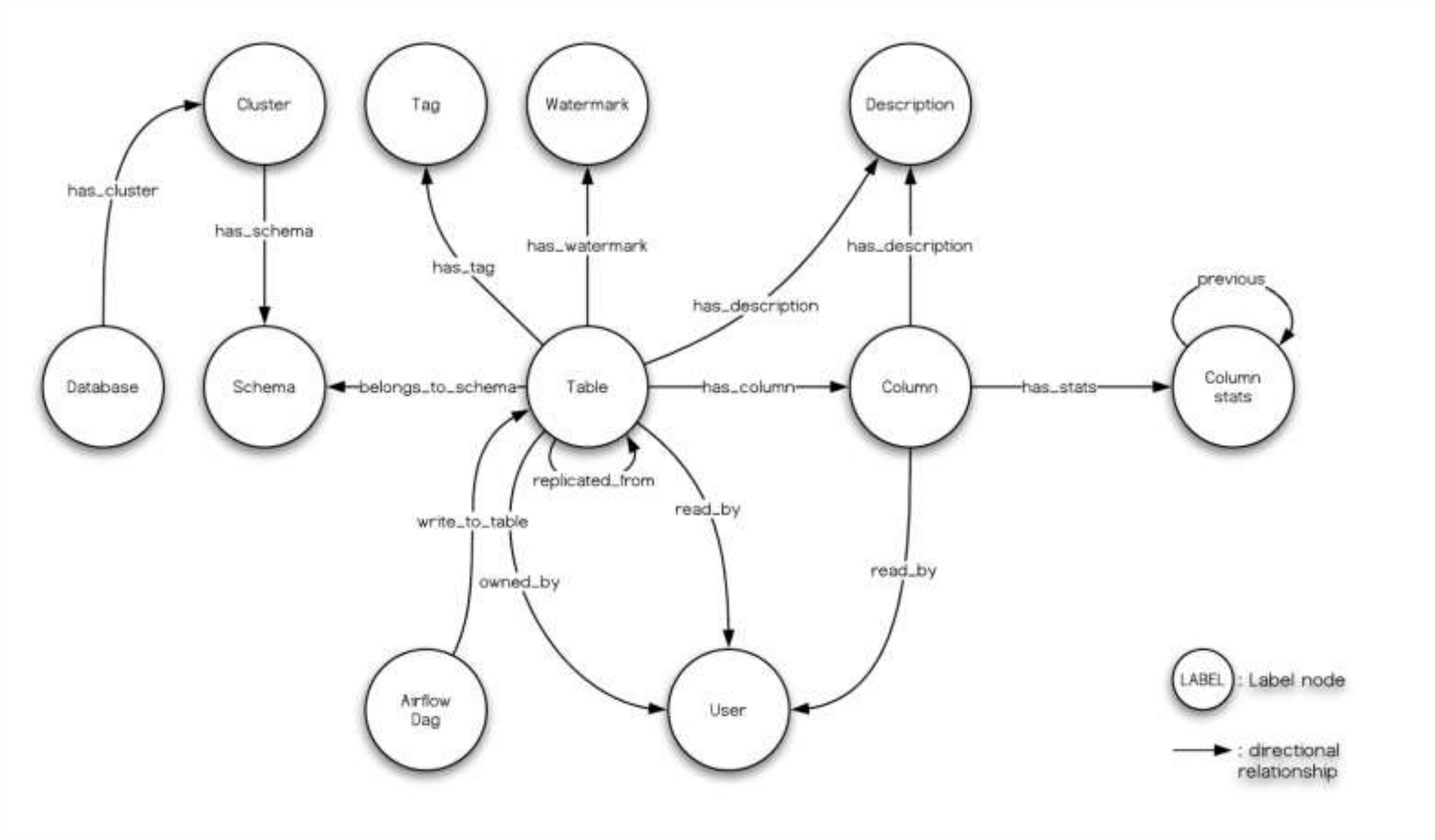
SOURCE CODE
schema.test_table

TABLE LINEAGE (BETA)
schema.test_table

TABLE PROFILE (BETA)

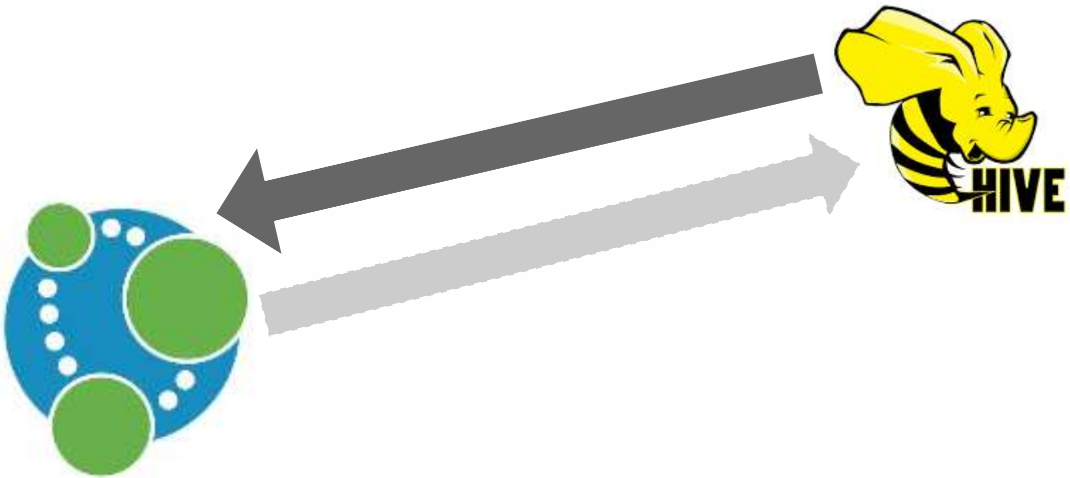
TAGS
amundsen-test
test

Why Graph database? (2/2)

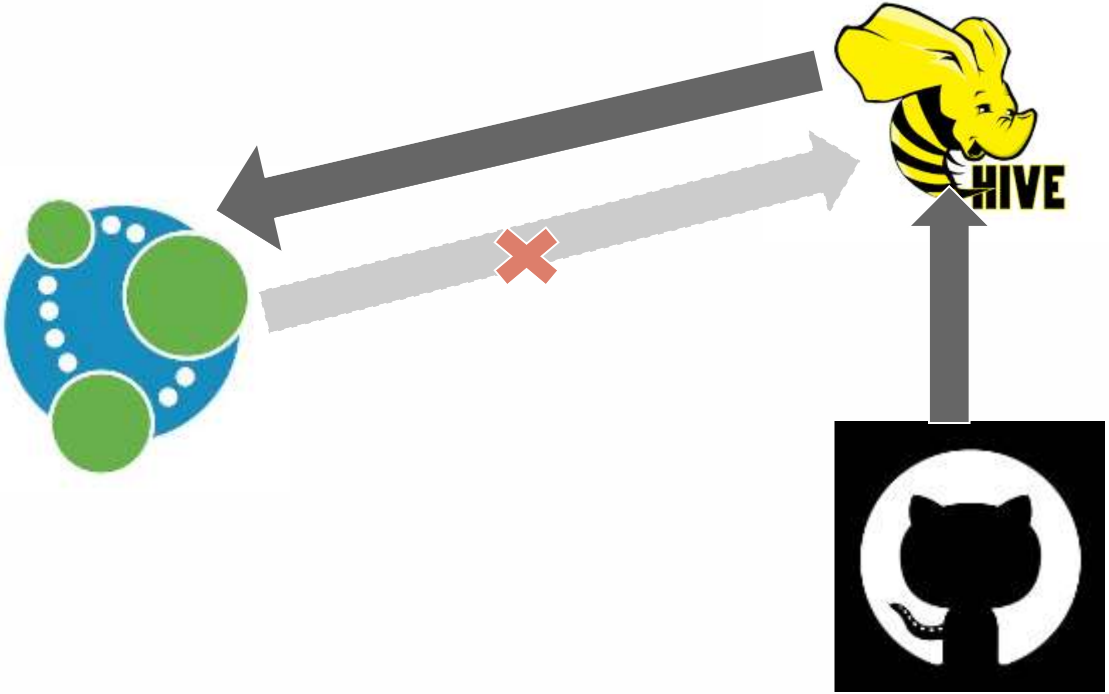


Neo4j is the source of truth for editable metadata

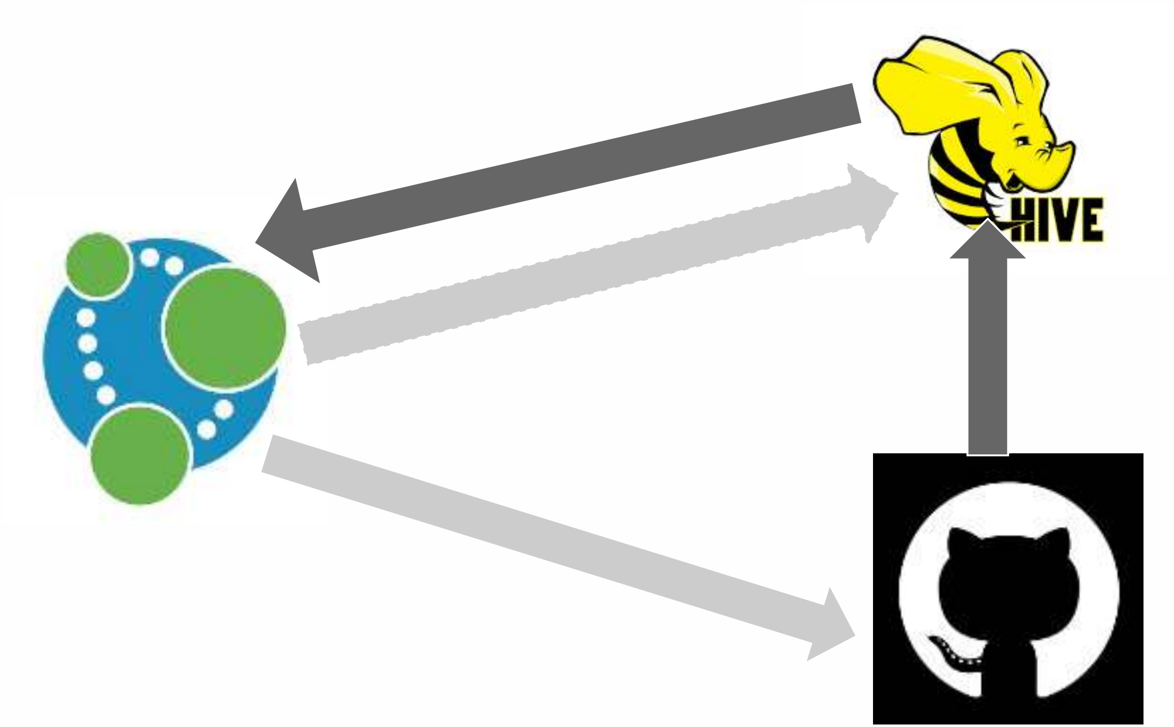
Why not propagate the editabled metadata back to source



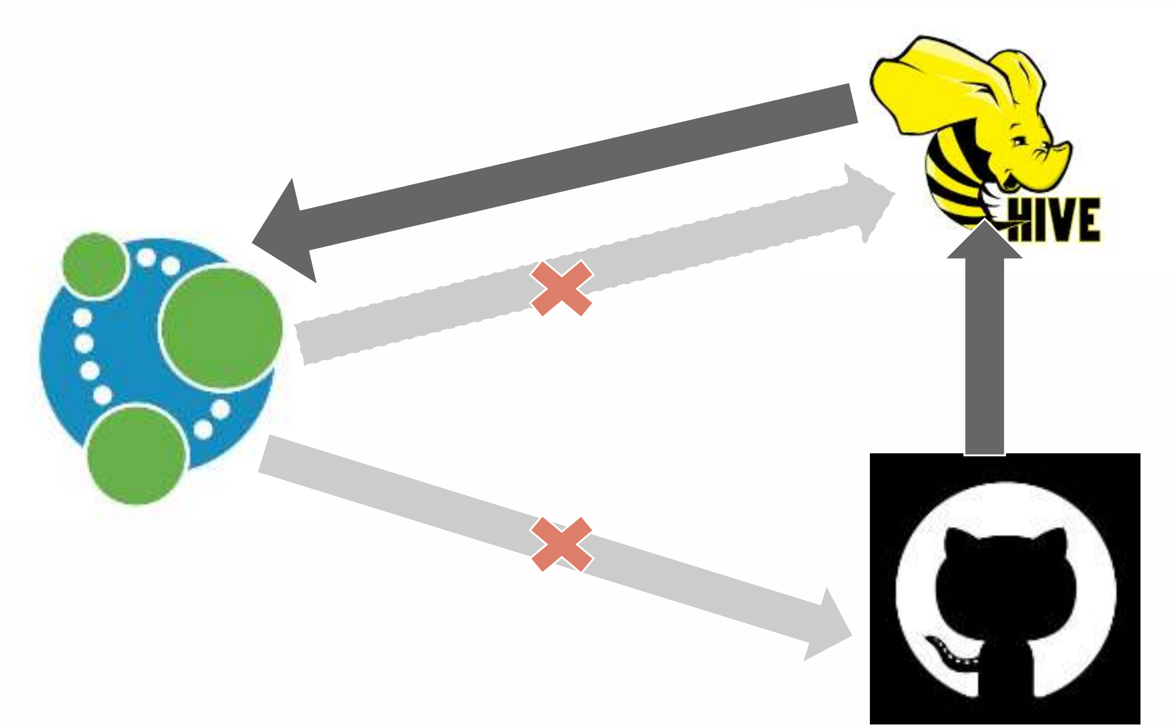
Why not propagate the editabled metadata back to source



Why not propagate the editabled metadata back to source



Why not propagate the editabled metadata back to source



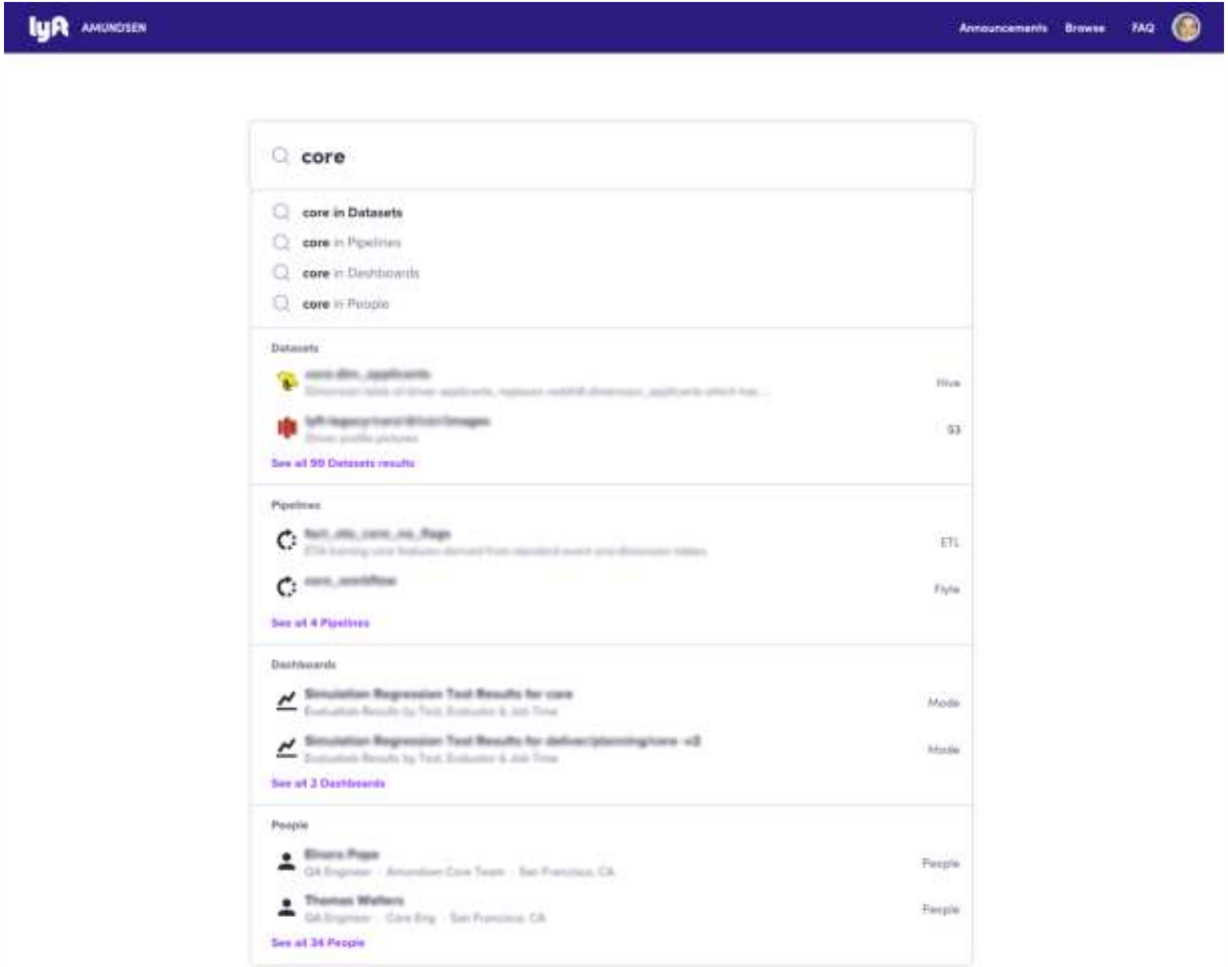
Impact at Lyft

Amundsen’s Impact at Lyft

- Deployed at Lyft for over 1 year
- Over 700 Weekly Active Users
- 90% penetration among Data Scientists
- Reduced mean time to discovery by 75%
- Also used by Data Eng, Software Eng, PMs, Ops, Marketing Managers, and more

Future Work

Search Preview



core

core in Datasets

core in Pipelines

core in Dashboards

core in People

Datasets

core.dim_applicants
Dimension table of driver applicants, replaces redshift dimension_applicants which has ...
Hive

lyft-legacy/[illegible]/images
Driver profile pictures
S3

See all 99 Datasets results

Pipelines

fact_ats_core_ns_flags
ETA training core features derived from standard event and dimension tables
ETL

core_workflow
Flyte

See all 4 Pipelines

Dashboards

Simulation Regression Test Results for core
Evaluation Results by Test, Evaluator & Job Time
Mode

Simulation Regression Test Results for deliver/planning/core -v2
Evaluation Results by Test, Evaluator & Job Time
Mode

See all 2 Dashboards

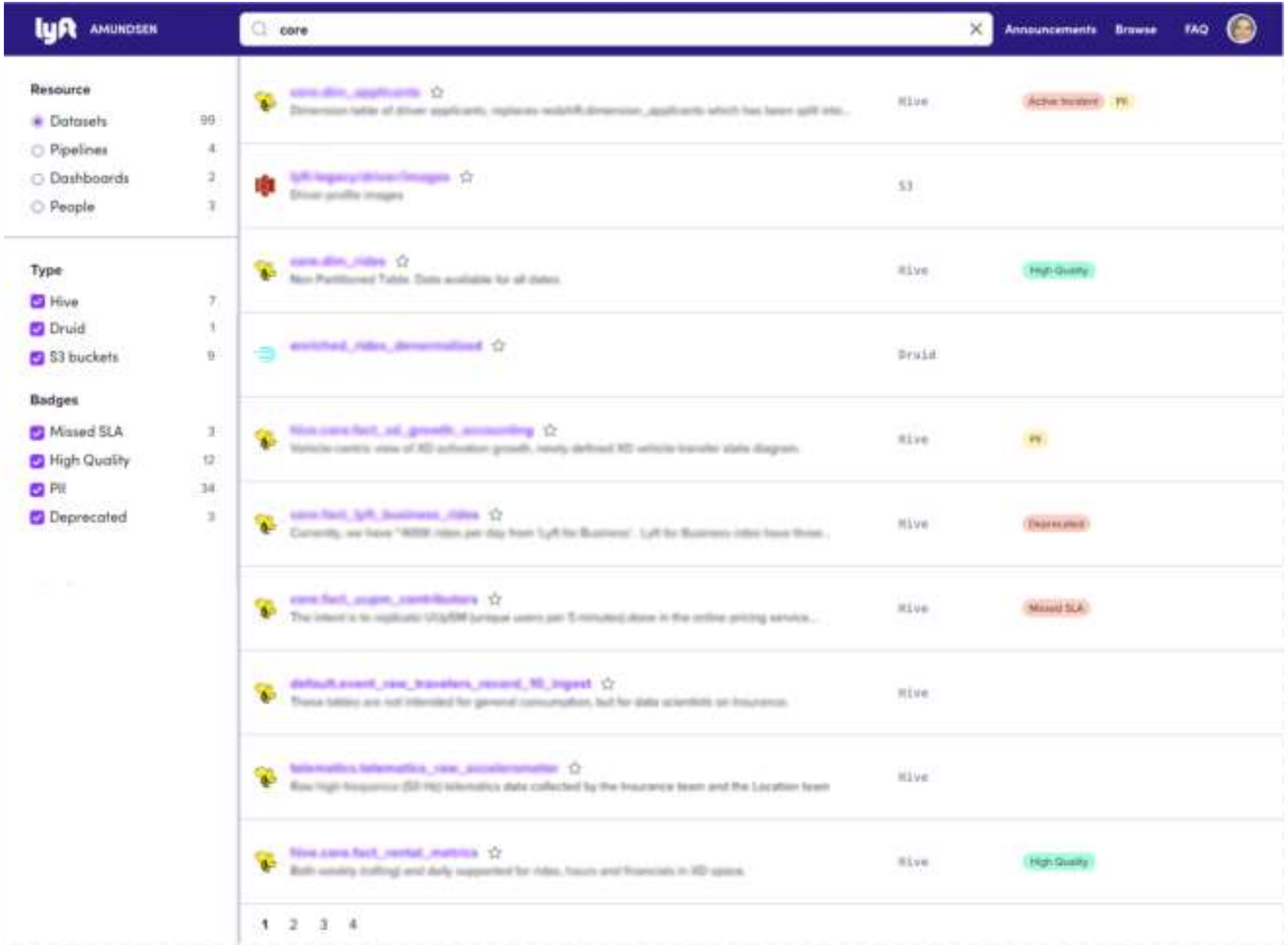
People

Elnora Pope
QA Engineer · Amundsen Core Team · San Francisco, CA
People

Thomas Walters
QA Engineer · Core Eng · San Francisco, CA
People

See all 34 People

Advanced Search



core

Resource

- ● Datasets 99
- ○ Pipelines 4
- ○ Dashboards 2
- ○ People 3

Type

- ☑ Hive 7
- ☑ Druid 1
- ☑ S3 buckets 9

Badges

- ☑ Missed SLA 3
- ☑ High Quality 12
- ☑ PII 34
- ☑ Deprecated 3

core.dim_applicants ☆
Dimension table of driver applicants, replaces notlyft.dimension_applicants which has been split into...
Hive
Active Incident PII

lyft-legacy/driver/images ☆
Driver profile images
S3

core.dim_rides ☆
Non Partitioned Table. Data available for all dates.
Hive
High Quality

enriched_rides_denormalized ☆
Druid

hive.core.fact_xd_growth_accounting ☆
Vehicle-centric view of XD activation growth, newly defined XD vehicle transfer state diagram.
Hive
PII

core.fact_lyft_business_rides ☆
Currently, we have ~80XK rides per day from 'Lyft for Business'. Lyft for Business rides have three...
Hive
Deprecated

core.fact_uupm_contributors ☆
The intent is to replicate UUpSM (unique users per 5-minutes) done in the online pricing service...
Hive
Missed SLA

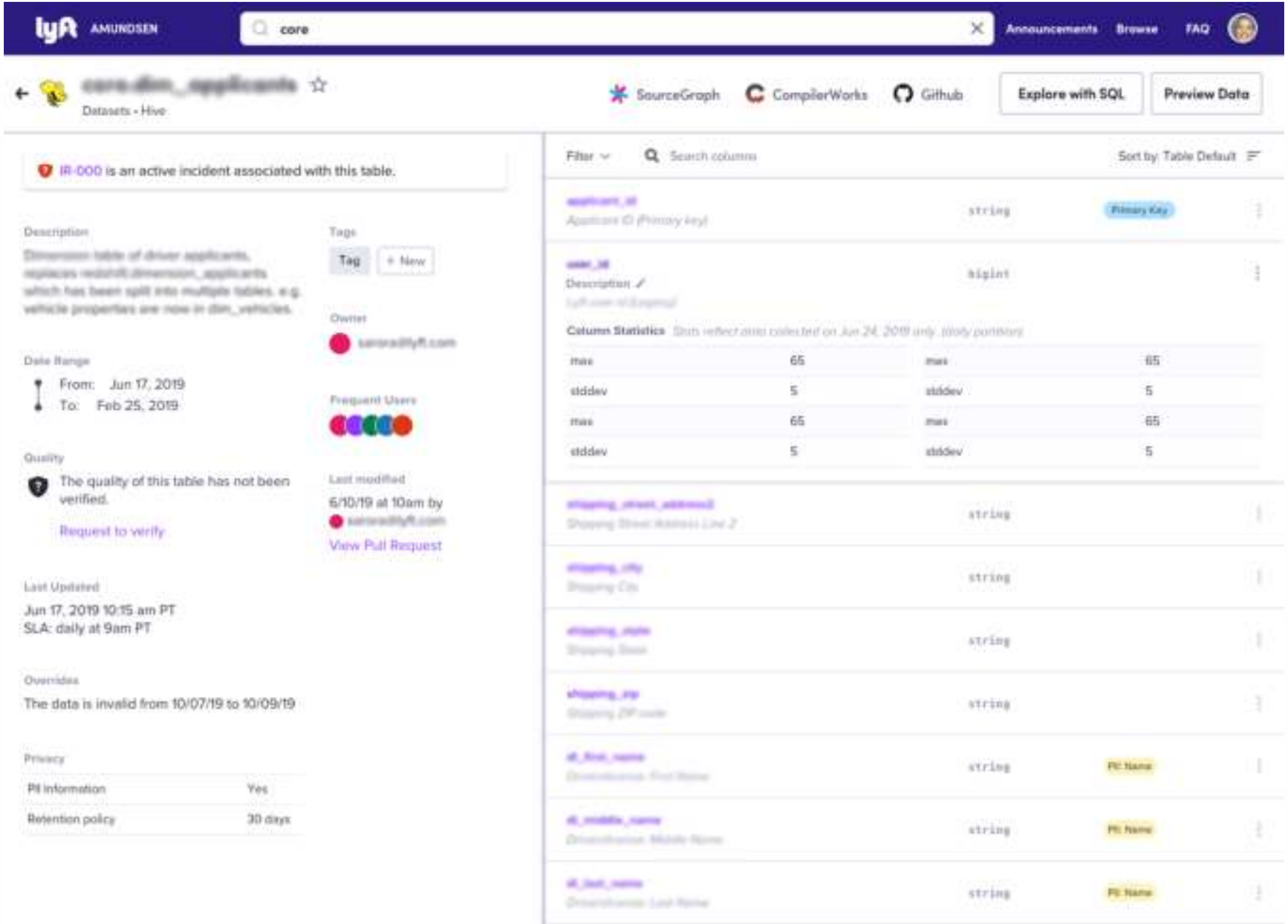
default.event_raw_travelers_record_50_ingest ☆
These tables are not intended for general consumption, but for data scientists on Insurance.
Hive

telematics.telematics_raw_accelerometer ☆
Raw high frequency (50 Hz) telematics data collected by the Insurance team and the Location team
Hive

hive.core.fact_rental_metrics ☆
Both weekly (rolling) and daily supported for rides, hours and financials in XD space.
Hive
High Quality

1 2 3 4

More Metadata



core.dim_applicants ☆

Datasets • Hive

IR-000 is an active incident associated with this table.

Description

Dimension table of driver applicants, replaces redshift.dimension_applicants which has been split into multiple tables. e.g. vehicle properties are now in dim_vehicles.

Tags

Tag + New

Owner

● [illegible]

Date Range

From: Jun 17, 2019
To: Feb 25, 2019

Frequent Users

Quality

The quality of this table has not been verified.

Request to verify

Last modified

6/10/19 at 10am by
● [illegible]
View Pull Request

Last Updated

Jun 17, 2019 10:15 am PT
SLA: daily at 9am PT

Overrides

The data is invalid from 10/07/19 to 10/09/19

Privacy

PII information	Yes
Retention policy	30 days

applicant_id
Applicant ID (Primary key) — string — Primary Key

user_id — bigint
Description
Lyft user id [illegible]

Column Statistics Stats reflect data collected on Jun 24, 2019 only. (daily partition)

max	65	max	65
stddev	5	stddev	5
max	65	max	65
stddev	5	stddev	5

shipping_street_address2
Shipping Street Address Line 2 — string

shipping_city
Shipping City — string

shipping_state
Shipping State — string

shipping_zip
Shipping ZIP code — string

dl_first_name
Driverlicense: First Name — string — PII: Name

dl_middle_name
Driverlicense: Middle Name — string — PII: Name

dl_last_name
Driverlicense: Last Name — string — PII: Name

We're Open Source

Amundsen is Open Source!

- github.com/lyft/amundsen
- 200+ github stars, 10+ companies contributing back
- Slack channel 250+ people from 30+ companies
- Presented at conferences in San Francisco, Barcelona, Vilnius, Moscow, LA, NYC by Lyft employees and community

Community Overview

Contributors

- ING
- Square
- EVERFI
- B&O BANG & OLUFSEN
- cool blue
- PerkinElmer
- trivadis
- lnc
- edmunds

Active community

- ORACLE
- workday.
- we
- INTERCOM
- AXA
- one medical
- sas
- BRANDLESS™
- upwork™
- HOVER
- stuart
- City of Hope
- kik·
- VOX MEDIA
- babylon
- M6
- vacasa
- SODA
- iRobot

Thank You

Data @ Lyft

Alagappan Sethuraman | /in/alagappanut
Daniel Won | /in/danwon

Project Code @ github.com/lyft/amundsen

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