Technical Advisory Council Meeting

July 30, 2020
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 Updegrove 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 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

<table>
<thead>
<tr>
<th>Member</th>
<th>Contact</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT&amp;T</td>
<td>Reuben Klein</td>
<td><a href="mailto:rk1518@att.com">rk1518@att.com</a></td>
</tr>
<tr>
<td>Baidu</td>
<td>Daxiang Dong</td>
<td><a href="mailto:dongdaxiang@baidu.com">dongdaxiang@baidu.com</a></td>
</tr>
<tr>
<td>Ericsson</td>
<td>Rani Yadav-Ranjan</td>
<td><a href="mailto:rani.yadav-ranjan@ericsson.com">rani.yadav-ranjan@ericsson.com</a></td>
</tr>
<tr>
<td>Huawei</td>
<td>Huang Zhipeng</td>
<td><a href="mailto:huangzhipeng@huawei.com">huangzhipeng@huawei.com</a></td>
</tr>
<tr>
<td>Nokia</td>
<td>Pantelis Monogioudis</td>
<td><a href="mailto:pantelis.monogioudis@nokia.com">pantelis.monogioudis@nokia.com</a></td>
</tr>
<tr>
<td>Tech Mahindra</td>
<td>Nikunj Nirmal</td>
<td><a href="mailto:nn006444@techmahindra.com">nn006444@techmahindra.com</a></td>
</tr>
<tr>
<td>Tencent</td>
<td>Bruce Tao</td>
<td><a href="mailto:brucetao@tencent.com">brucetao@tencent.com</a></td>
</tr>
<tr>
<td>Zilliz</td>
<td>Jun Gu</td>
<td><a href="mailto:jun.gu@zilliz.com">jun.gu@zilliz.com</a></td>
</tr>
<tr>
<td>ZTE</td>
<td>Wei Meng</td>
<td><a href="mailto:meng.wei2@zte.com.cn">meng.wei2@zte.com.cn</a></td>
</tr>
<tr>
<td>Acumos AI Project</td>
<td>Nat Subramaninan</td>
<td><a href="mailto:natarajan.subramanian@techmahindra.com">natarajan.subramanian@techmahindra.com</a></td>
</tr>
<tr>
<td>Angel Project</td>
<td>Bruce Tao</td>
<td><a href="mailto:brucetao@tencent.com">brucetao@tencent.com</a></td>
</tr>
<tr>
<td>ONNX Project</td>
<td>Jim Spohrer*</td>
<td><a href="mailto:spohrer@us.ibm.com">spohrer@us.ibm.com</a></td>
</tr>
</tbody>
</table>

* 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](https://github.com/lyft/amundsen)
- **Projects Level:** Incubation
- **Presenter(s):** Mark Grover and Tao Feng
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.

Website: lyft.github.io/amundsen
Repository: github.com/lyft/amundsen
Crunchbase: crunchbase.com/organization/lyft
LinkedIn: linkedin.com/company/lyft
Twitter: @lyft
First Commit: a year ago
Contributors: 40
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

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

Lots of queries like:

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

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

Increased database load

Interrupt heavy data culture
Evaluating solutions
Holy grail of solving for productivity

metadata

*noun* /ˈmedəˌdādə,ˈmedəˌdadə/

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

1. What kind of information?

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:**

<table>
<thead>
<tr>
<th>Data stores</th>
<th>People</th>
<th>Dashboard / Reports</th>
<th>Notebooks</th>
<th>Events / Schemas</th>
<th>Streams</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>Employees</td>
<td>Tableau</td>
<td>Jupyter</td>
<td>Schema registry</td>
<td>Kafka</td>
</tr>
<tr>
<td>HIVE</td>
<td></td>
<td>Looker</td>
<td>Segment</td>
<td></td>
<td>Amazon Kinesis</td>
</tr>
<tr>
<td>PostgreSQL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

**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

<table>
<thead>
<tr>
<th>Criteria / Products</th>
<th>Alation</th>
<th>Where Hows</th>
<th>Airbnb Data Portal</th>
<th>Cloudera Navigator</th>
<th>Apache Atlas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search based</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Lineage based</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Network based</td>
<td>❌</td>
<td>❌</td>
<td>✔️</td>
<td>❌</td>
<td>❌</td>
</tr>
<tr>
<td>Hive/Presto support</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Redshift support</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
</tr>
<tr>
<td>Open source (pref.)</td>
<td>❌</td>
<td>✔️</td>
<td>✔️</td>
<td>❌</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Meet Amundsen

First person to discover the South Pole - Norwegian explorer, Roald Amundsen
<table>
<thead>
<tr>
<th>Power user</th>
<th>Noob user</th>
<th>Manager</th>
</tr>
</thead>
</table>
| ● All info in their head  
● Get interrupted a lot due to questions | ● Lost  
● Ask “power users” a lot of questions | ● Dependencies landing on time  
● Communicating with stakeholders |
Search for data, or 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

<table>
<thead>
<tr>
<th>tag1</th>
<th>tag2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

My Bookmarks

| test_schema.test_table2 | 2nd test table | dynamo |

Popular Tables

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

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

**default.event_amundsenfrontend_user_action**

*Datasets • Hive*

### Description

User action event from Amundsen frontend

#### Request Description

#### Issues

No associated issues

#### Report an issue

### Date Range

- **From:** Aug 02, 2018
- **To:** Mar 02, 2020

#### Last Updated

Mar 05, 2020 7am PST

### Frequent Users

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

#### Read-only information, auto-generated

### Columns (17) | Dashboards (1)

<table>
<thead>
<tr>
<th>Column</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>event_id</td>
<td>string</td>
</tr>
<tr>
<td>Unique event identifier. Due to current assumptions in the pipeline, it's important that this be a version 4 (random) UUID.</td>
<td></td>
</tr>
<tr>
<td>ds</td>
<td>string</td>
</tr>
<tr>
<td>test</td>
<td>string</td>
</tr>
<tr>
<td>command</td>
<td>string</td>
</tr>
<tr>
<td>Action command type from user e.g: search, get_table-metadata, etc.</td>
<td></td>
</tr>
<tr>
<td>end_epoch_ms</td>
<td>bigint</td>
</tr>
<tr>
<td>end time in epoch ms</td>
<td></td>
</tr>
<tr>
<td>error</td>
<td>string</td>
</tr>
<tr>
<td>An error message or exception stacktrace</td>
<td></td>
</tr>
<tr>
<td>host_name</td>
<td>string</td>
</tr>
<tr>
<td>Sending host name</td>
<td></td>
</tr>
<tr>
<td>http_request_id</td>
<td>string</td>
</tr>
<tr>
<td>keyword_args_json</td>
<td>string</td>
</tr>
<tr>
<td>Json object contains key word arguments</td>
<td></td>
</tr>
<tr>
<td>occurred_at</td>
<td></td>
</tr>
</tbody>
</table>
See detailed descriptions and profile of the column

<table>
<thead>
<tr>
<th>Description</th>
<th>string</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is an editable test description for the first column. This also supports <strong>Markdown</strong>.</td>
<td></td>
</tr>
</tbody>
</table>

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

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>distinct values</td>
<td>8</td>
<td>min</td>
<td>aardvark</td>
</tr>
<tr>
<td>num nulls</td>
<td>500320</td>
<td>max</td>
<td>zebra</td>
</tr>
<tr>
<td>verified</td>
<td>230430</td>
<td></td>
<td></td>
</tr>
<tr>
<td>USERS</td>
<td>DESK_COUNT</td>
<td>DESK_COUNT</td>
<td>DIMENSION_SIDES</td>
</tr>
<tr>
<td>-------------</td>
<td>------------</td>
<td>------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Keegan Jones</td>
<td>true</td>
<td>120,000</td>
<td>120,000</td>
</tr>
<tr>
<td>Keegan Jones</td>
<td>true</td>
<td>120,000</td>
<td>120,000</td>
</tr>
<tr>
<td>Keegan Jones</td>
<td>true</td>
<td>120,000</td>
<td>120,000</td>
</tr>
<tr>
<td>Keegan Jones</td>
<td>true</td>
<td>120,000</td>
<td>120,000</td>
</tr>
<tr>
<td>Keegan Jones</td>
<td>true</td>
<td>120,000</td>
<td>120,000</td>
</tr>
<tr>
<td>Keegan Jones</td>
<td>true</td>
<td>120,000</td>
<td>120,000</td>
</tr>
<tr>
<td>Keegan Jones</td>
<td>true</td>
<td>120,000</td>
<td>120,000</td>
</tr>
<tr>
<td>Keegan Jones</td>
<td>true</td>
<td>120,000</td>
<td>120,000</td>
</tr>
<tr>
<td>Keegan Jones</td>
<td>true</td>
<td>120,000</td>
<td>120,000</td>
</tr>
<tr>
<td>Keegan Jones</td>
<td>true</td>
<td>120,000</td>
<td>120,000</td>
</tr>
<tr>
<td>Keegan Jones</td>
<td>true</td>
<td>120,000</td>
<td>120,000</td>
</tr>
<tr>
<td>Keegan Jones</td>
<td>true</td>
<td>120,000</td>
<td>120,000</td>
</tr>
</tbody>
</table>
See dashboards built on this data set
# Search for existing dashboards/reports

**Resource**
- **Datasets**: 60
- **Dashboards**: 2
- **People**: 0

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Mode</th>
<th>Last Successful Run</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPE</td>
<td>amundsen_dashboard_table_lineage</td>
<td></td>
<td>Jun 12, 2020</td>
</tr>
<tr>
<td></td>
<td><strong>Global Ops Analytics - Scratchpad</strong></td>
<td>Mode</td>
<td>May 25, 2020</td>
</tr>
<tr>
<td></td>
<td>Cloned copy of the report linked to in <a href="https://confluence.lyft.net/display/DATA/Amundsen/Search+Tutorial+as+of+S/">https://confluence.lyft.net/display/DATA/Amundsen/Search+Tutorial+as+of+S/</a>..</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

https://amundsen.lyft.net/dashboardmode_dashboard%3A%2F%2Fgold.bc0d9dfe007c2fe52d535ae0f838f838/
Dashboard detail page

Dashboard in DPE

**Description**
Add Description in Mode

**Owners**
Tao Feng

**Created**
Jun 01, 2020 11pm PDT

**Last Successful Run**
Jun 12, 2020 5pm PDT

**Last Run**
Jun 12, 2020 5pm PDT

**Recent View Count**
1

**Tables (5)**

- **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 completed.
  - Hive

- **hivemetastore.dbs**
  - Imported by sqoop on 2019/10/01 00:31:07
  - Hive

- **hivemetastore.tbsls**
  - Hive

- **default.event_security_audit**
  - The event that is emitted when logging a security audit event
  - Hive

Amundsen was last indexed on June 23rd 2020 at 5:30:49 pm
Search for co-workers!

Mark Grover
Product Manager • Data Tools & Productivity
User
Search for data owned and used by your peers!
Amundsen Architecture
1. Metadata Service
# Rides

May 25, 2012 – Mar 03, 2019

The source for all ride related data.

## Columns

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>users</strong> string</td>
<td>Dummy description. You can click here to edit.</td>
</tr>
<tr>
<td><strong>desk_count</strong> int</td>
<td>Dummy description. You can click here to edit.</td>
</tr>
<tr>
<td><strong>passenger</strong> string</td>
<td>Add Description</td>
</tr>
<tr>
<td><strong>ride_id</strong> string</td>
<td>Add Description</td>
</tr>
<tr>
<td><strong>driver_os</strong> string</td>
<td>Add Description</td>
</tr>
<tr>
<td><strong>driver_os_version</strong> string</td>
<td>Dummy description. You can click here to edit.</td>
</tr>
<tr>
<td><strong>driver_app_version</strong> string</td>
<td>Add Description</td>
</tr>
</tbody>
</table>
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

- 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

Sources

Dashboard Tools
Databuilder

Job

Task

Extractor → Transformer → Loader

Publisher
Databuilder in action

HiveTableMetadataExtractor

SQLAlchemyExtractor

NoopTransformer

FsNeo4jCSVLoader

Neo4jCsvPublisher

Neo4j

Neo4j serialized CSV files

produces

: data movement
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`)
4. Frontend service
Frontend Stacks

- TypeScript
- ReactJS
- webpack
- Redux
- Jest
- npm
- Babel
Core Infra high level architecture
**See details of the data set**

### Dataset Details

- **Name**: default.event_airflow_task_routed
- **Description**: Tracks the routing information for an Airflow task.
- **Tags**: None
- **Last Updated**: Jul 08, 2020 9pm PDT
- **Owners**: dp-exp@lyft.com

### Columns (16)

<table>
<thead>
<tr>
<th>Column</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>event_id</td>
<td>string</td>
</tr>
<tr>
<td>ds</td>
<td>string</td>
</tr>
<tr>
<td>dag_id</td>
<td>string</td>
</tr>
<tr>
<td>execution_date</td>
<td>string</td>
</tr>
<tr>
<td>http_request_id</td>
<td>string</td>
</tr>
<tr>
<td>occurred_at</td>
<td>timestamp</td>
</tr>
<tr>
<td>producer_service_name</td>
<td>string</td>
</tr>
<tr>
<td>queue</td>
<td>string</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Unique event identifier. Due to current assumptions in the pipeline, it's important that this be a version 4 (random) UUID.**
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

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
Community overview
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

- Search all Lyft data...

Popular Tables
- rides
- passengers
- contact\_trail
- dimension\_rides
- weekly\_active\_trails

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**
- Provide data & metadata only on a per need basis

**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

- Data Discovery & Trust
- Security / ACL
- Compliance (GDPR / CCPA / Financial)
- Data Monitoring / Ops
- Data Quality
- Cost Mgmt
- Data Maintenance

Metadata
2. Develop depth of metadata

Phase 1 (Complete)
Data sets

Phase 2 (In development)
Dashboards
People

Phase 3 (In Scoping)
Streams
Schemas
Workflows
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/
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 LF AI Project Portfolio
A Growing Developer Community

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
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

<table>
<thead>
<tr>
<th></th>
<th>Joined incubation</th>
<th>Unique contributors</th>
<th>Involved companies</th>
<th>GH Stars</th>
<th>Flow of commits</th>
<th>CII badge</th>
<th>Collaboration with other LF AI project</th>
<th>TSC Members</th>
</tr>
</thead>
</table>
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
LF AI PR/Comms

› Please follow LF AI on [Twitter](https://twitter.com) & [LinkedIn](https://www.linkedin.com) 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](https://opensource.linuxfoundation.org/display/LF/AI+Blog)

› Open call to publish project/committee updates or other relevant content on the [LF AI Blog](https://opensource.linuxfoundation.org/display/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.