Meeting of the Technical Advisory Council (TAC)

July 15th, 2021

LF AI & DATA

Antitrust Policy

- 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 & Data Useful Links

- Web site: lfaidata.foundation
- Wiki: wiki.lfaidata.foundation
- GitHub: github.com/lfaidata
- Landscape: https://landscape.lfaidata.foundation or https://l.lfaidata.foundation
- Mail Lists: https://lists.lfaidata.foundation
- Slack: https://slack.lfaidata.foundation
- Youtube: https://www.youtube.com/channel/UCfasaeqXJBCAJMNO9HcHfbA
- LF AI Logos: https://github.com/lfaidata/artwork/tree/master/lfaidata
- LF AI Presentation Template: https://drive.google.com/file/d/1eiDNJvXCqSZHT4Zk_-czASlz2GTBRZk2/view?usp=sharing
-
- Events Page on LF AI Website: https://lfaidata.foundation/events/
- Events Calendar on LF AI Wiki (subscribe available): https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=12091544
- Event Wiki Pages: https://wiki.lfaidata.foundation/display/DL/LF+AI+Data+Foundation+Events

Agenda

- Roll Call (2 mins)
- Approval of Minutes from June 17th (3 mins)
- Invited Presentation - (35 minutes)
  - Tony Project - Tony Project https://github.com/linkedin/TonY (Kequi Hu - khu@linkedin.com);
- Annual Project Review (15 minutes)
  - EDL - https://github.com/lfai/proposing-projects/blob/master/proposals/edl.adoc (zhouti@baidu.com)
- Annual Project Review
- LF AI General Updates
- Open Discussion

TAC Voting Members

* = still need backup specified on wiki

Board Member	Contact Person	Email
AT&T	Anwar Atfab*	anwar@research.att.com
Baidu	Ti Zhou	zhouti@baidu.com
Ericsson	Rani Yadav-Ranjan*	rani.yadav-ranjan@ericsson.com
Huawei	Huang Zhipeng	huangzhipeng@huawei.com
IBM	Susan Malaika	malaika@us.ibm.com
Nokia	Jonne Soininen	jonne.soininen@nokia.com
OPPO	Jimin Jia*	jiajimin@oppo.com
SAS	Nancy Rausch	nancy.rausch@sas.com
Tech Mahindra	Amit Kumar	Kumar_Amit@techmahindra.com
Tencent	Bruce Tao	brucetao@tencent.com
Zilliz	Jun Gu	jun.gu@zilliz.com
ZTE	Wei Meng	meng.wei2@zte.com.cn
Graduate Project	**Contact Person**	**Email**
Acumos	Nat Subramanian	natarajan.subramanian@techmahindra.com
Angel	Bruce Tao	brucetao@tencent.com
Egeria	Mandy Chessell	mandy_chessell@uk.ibm.com
Horovod	Travis Addair*	taddair@uber.com
Milvus	Xiaofan Luan	xiaofan.luan@zilliz.com
ONNX	Jim Spohrer (Chair of TAC)	spohrer@us.ibm.com
Pyro	Fritz Obermeyer*	fritz.obermeyer@gmail.com

Approval of June 17th, 2021 Minutes

Draft minutes from the June 17th TAC call were previously distributed to the TAC members via the mailing list

Proposed Resolution:

- That the minutes of the June 17th meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.

Invited Presentation - Tony Project

LinkedIn - Tony Project
https://github.com/linkedin/TonY
(contact Keqiu Hu- khu@linkedin.com);

LF AI & DATA

TonY @ LF AI

Keqiu Hu khu@linkedin.com
07/15/2021

Why contribute TonY to Linux Foundation

Neutral holding ground
- vendor-neutral, not for profit.

Open governance model
- Transparent and open governance model
- Instill trust in contributors and adopters in the management of the project
- Neutral management of projects' assets by the foundation

Growing community
- Increase visibility of project through LF ecosystem
- Increase contributors by converting new & existing users
- Opportunities to collaborate with other hosted projects

Agenda

1. AI & Big Data @ LI
2. Training at Scale Challenges
3. How TonY Helps?
4. Next Steps

LinkedIn's Vision

Create economic opportunity
for every member of the global workforce

750M
Members

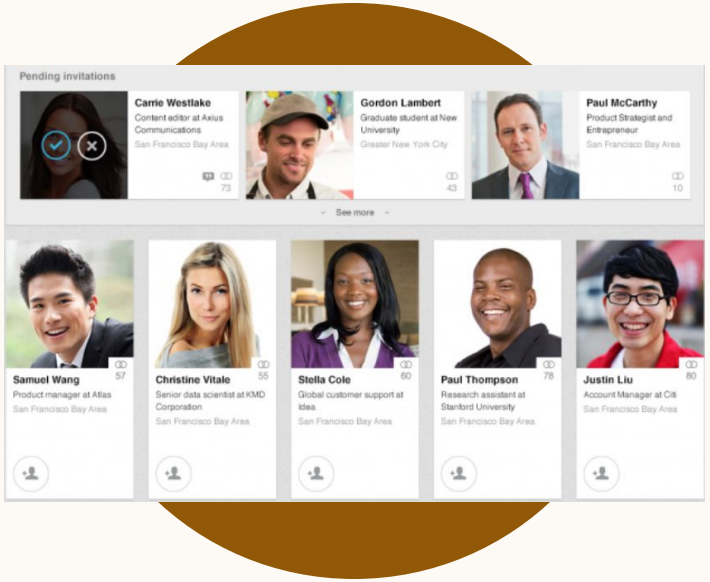
57M
Companies

40M
Jobs

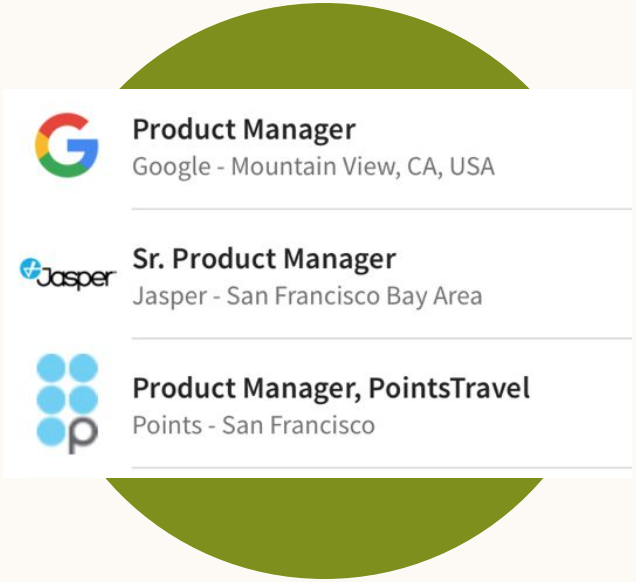
38K
Skills

120K
Schools

AI @ LinkedIn



People You May Know



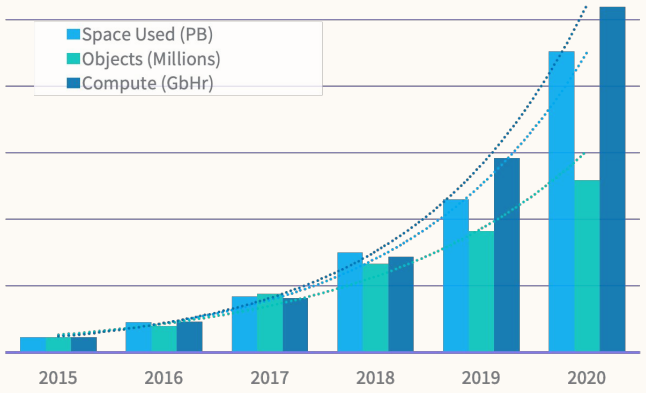
Job Recommendations

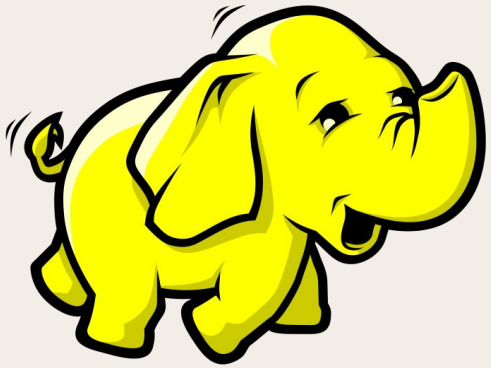
News Feed

Big Data Platform @ LinkedIn

Our stack
- Hadoop as big data OS
- Spark + TonY/TF + Hive for compute engines
- Azkaban for workflow management

Hadoop Cluster
- Single cluster 10K+ nodes, total 20K+ nodes.
- 2.2 PB compute memory
- 1 exabyte storage
- 5K+ users
- 3K concurrent apps.
- 2x YoY growth in capacity





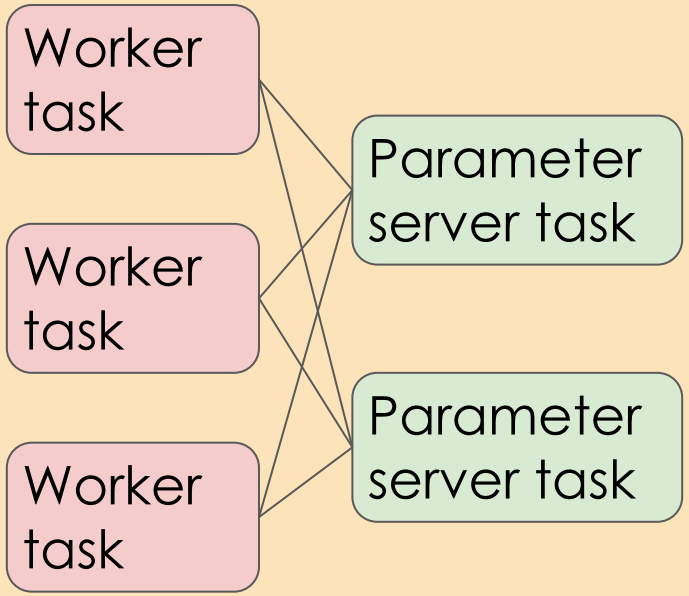
Deep Learning ? Big Data Platform

TensorFlow / PyTorch

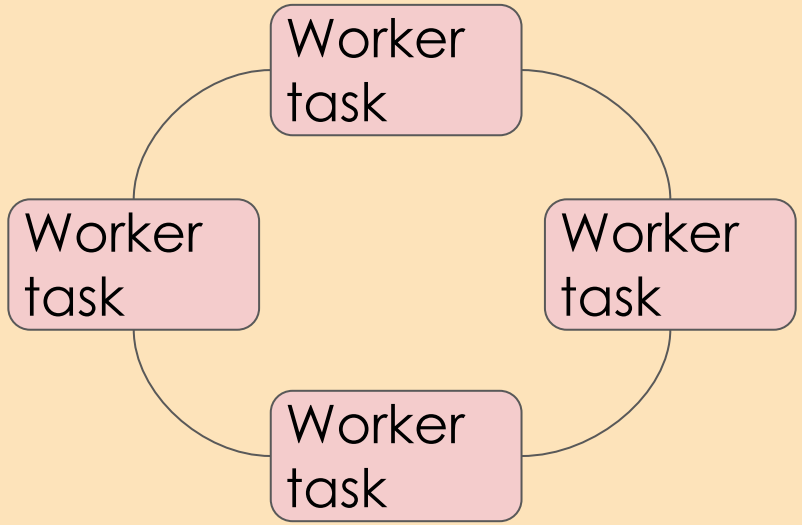
Hadoop / Kubernetes

Distributed TensorFlow

Worker/Parameter Server Model



Ring All-Reduce Model



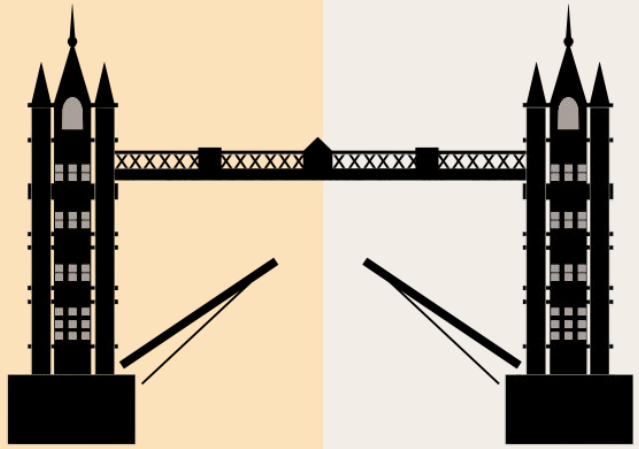
Early days: how AI engineers did training

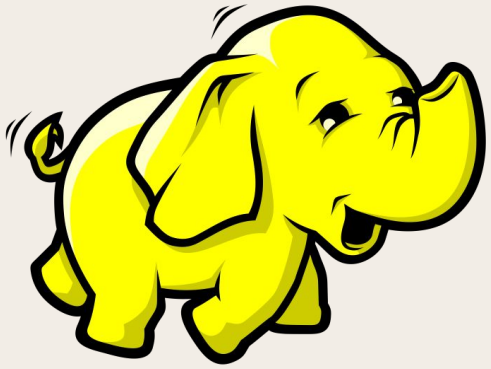
- Copy code and dependencies to each host
- Manually specify host and port of each process
- Customize arguments for each process

```
# On ps0.example.com:
$ python trainer.py \
     --ps_hosts=ps0.example.com:2222,ps1.example.com:2222 \
     --worker_hosts=worker0.example.com:2222,worker1.example.com:2222 \
     --job_name=ps --task_index=0
# On ps1.example.com:
$ python trainer.py \
     --ps_hosts=ps0.example.com:2222,ps1.example.com:2222 \
     --worker_hosts=worker0.example.com:2222,worker1.example.com:2222 \
     --job_name=ps --task_index=1
# On worker0.example.com:
$ python trainer.py \
     --ps_hosts=ps0.example.com:2222,ps1.example.com:2222 \
     --worker_hosts=worker0.example.com:2222,worker1.example.com:2222 \
     --job_name=worker --task_index=0
# On worker1.example.com:
$ python trainer.py \
     --ps_hosts=ps0.example.com:2222,ps1.example.com:2222 \
     --worker_hosts=worker0.example.com:2222,worker1.example.com:2222 \
      --job_name=worker --task_index=1
```

Challenges of scaling up training

- Managing code and dependencies
- Orchestrating distributed training
- Resource contention (especially for GPUs)
- Managing an ML workflow (data preparation, training, deployment)
- Fault tolerance

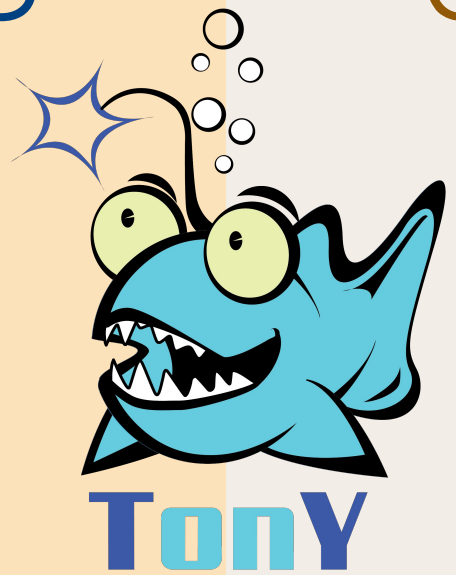




Deep Learning

Big Data Platform

TensorFlow / PyTorch



Hadoop YARN

Using TonY

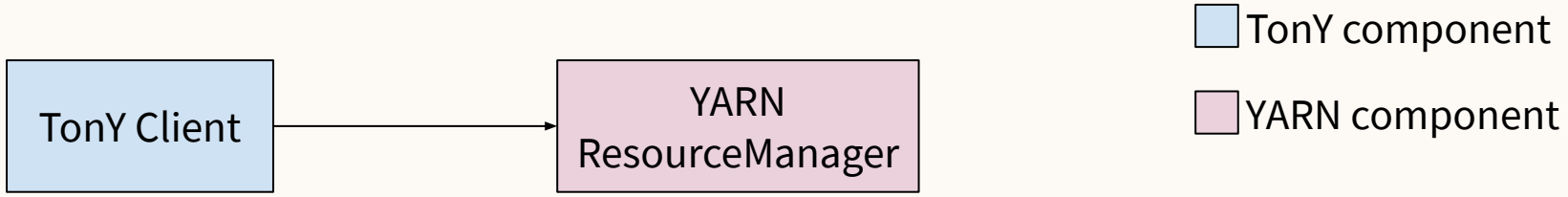
CLI

```
java -cp `hadoop
classpath`:tony-cli-0.3.7-all.jar \
com.linkedin.tony.cli.ClusterSubmitter \
--python_venv=venv.zip \
--src_dir=src \
--executes='venv/Python my_model.py' \
--conf_file=config.xml
```

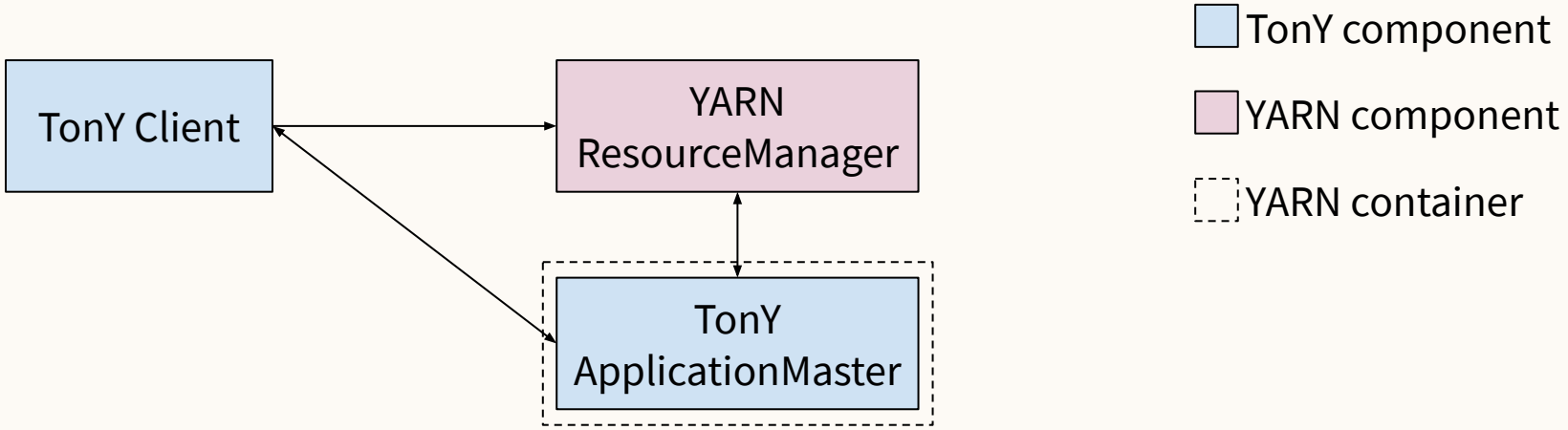
Conf File

```
<configuration>
  <property>
    <name>tony.worker.instances</name>
    <value>3</value>
  </property>
  <property>
    <name>tony.worker.gpus</name>
    <value>1</value>
  </property>
  <property>
    <name>tony.ps.instances</name>
    <value>1</value>
  </property>
</configuration>
```

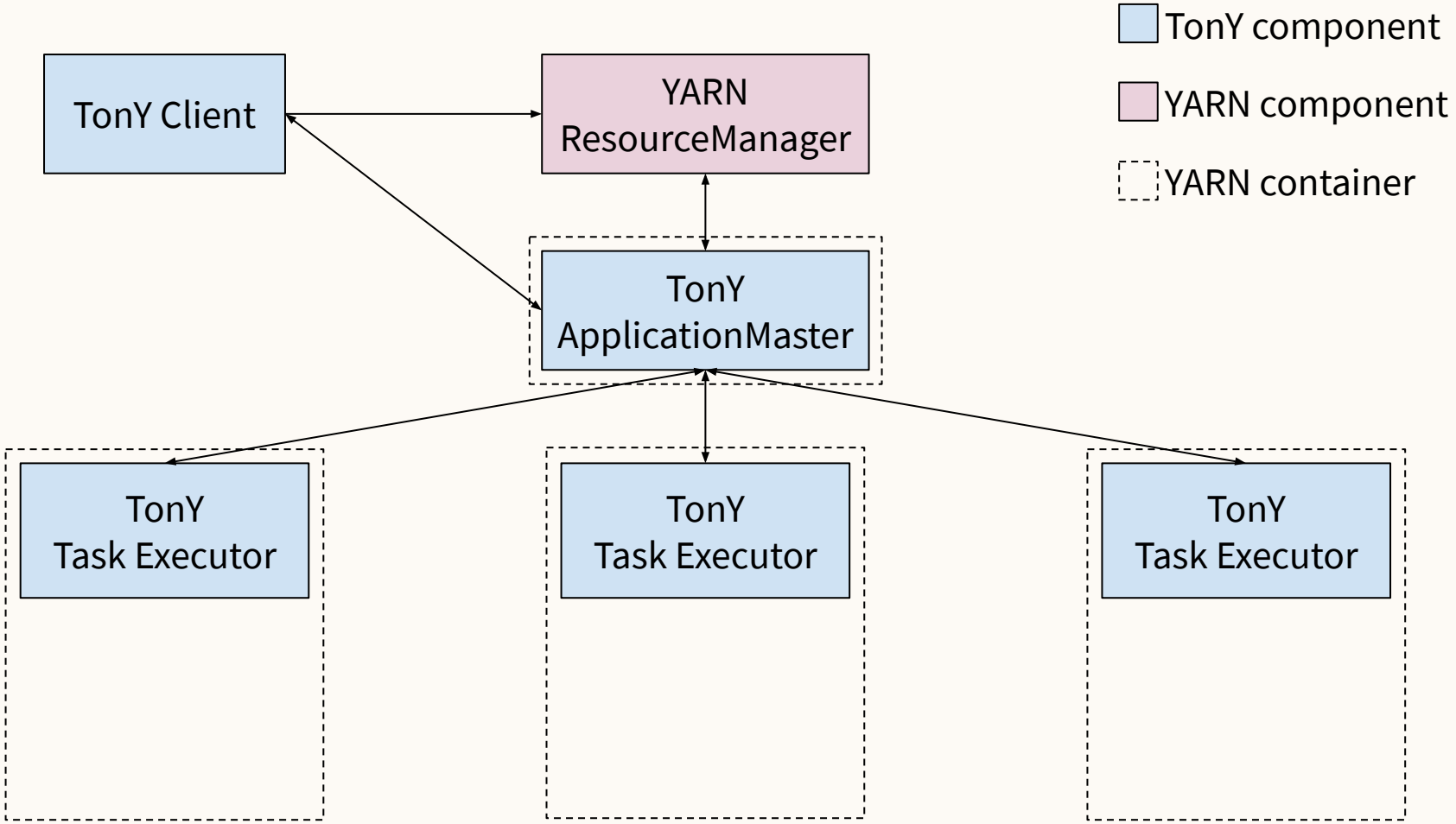
TonY Under the Hood



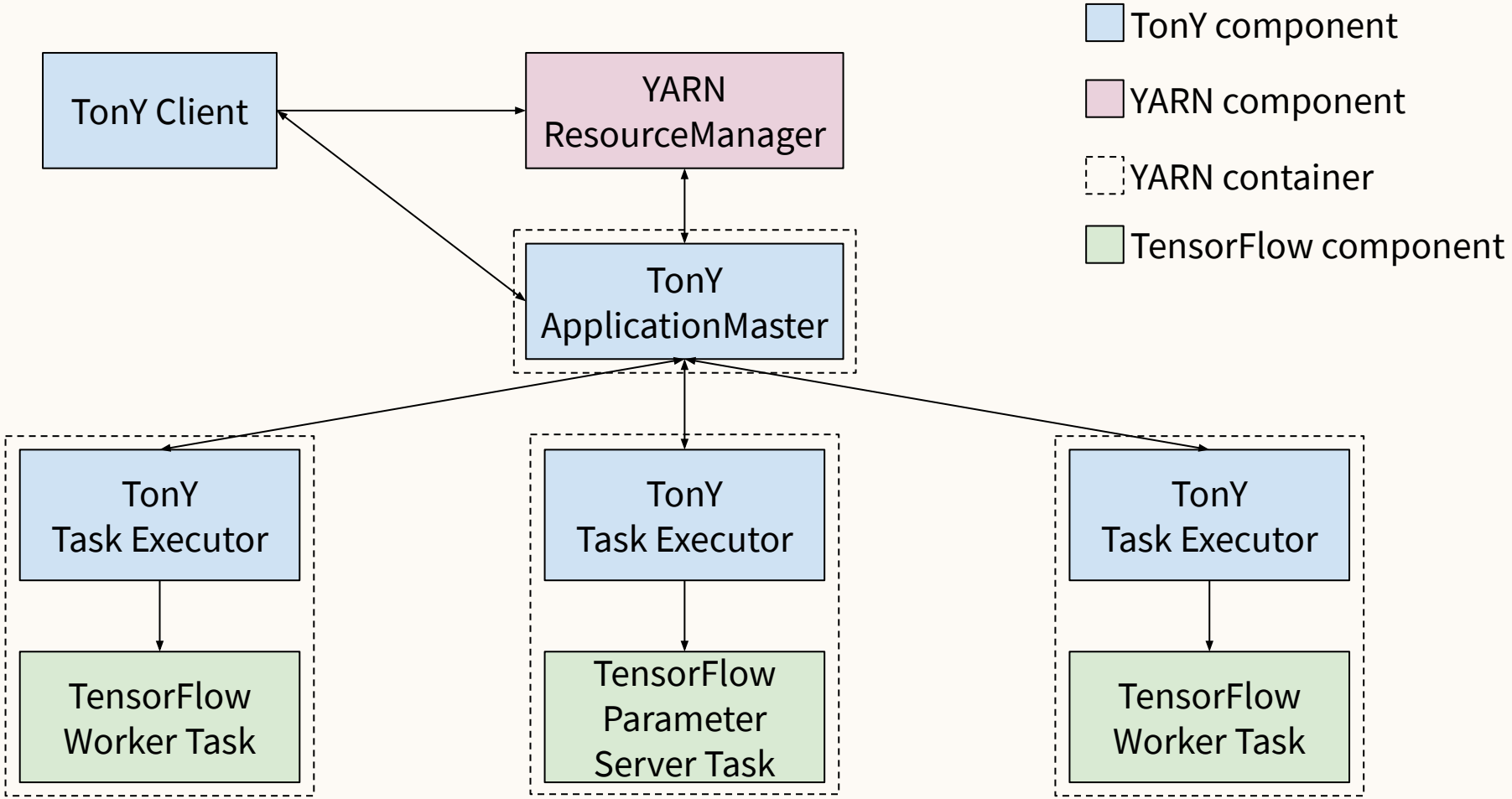
TonY Under the Hood



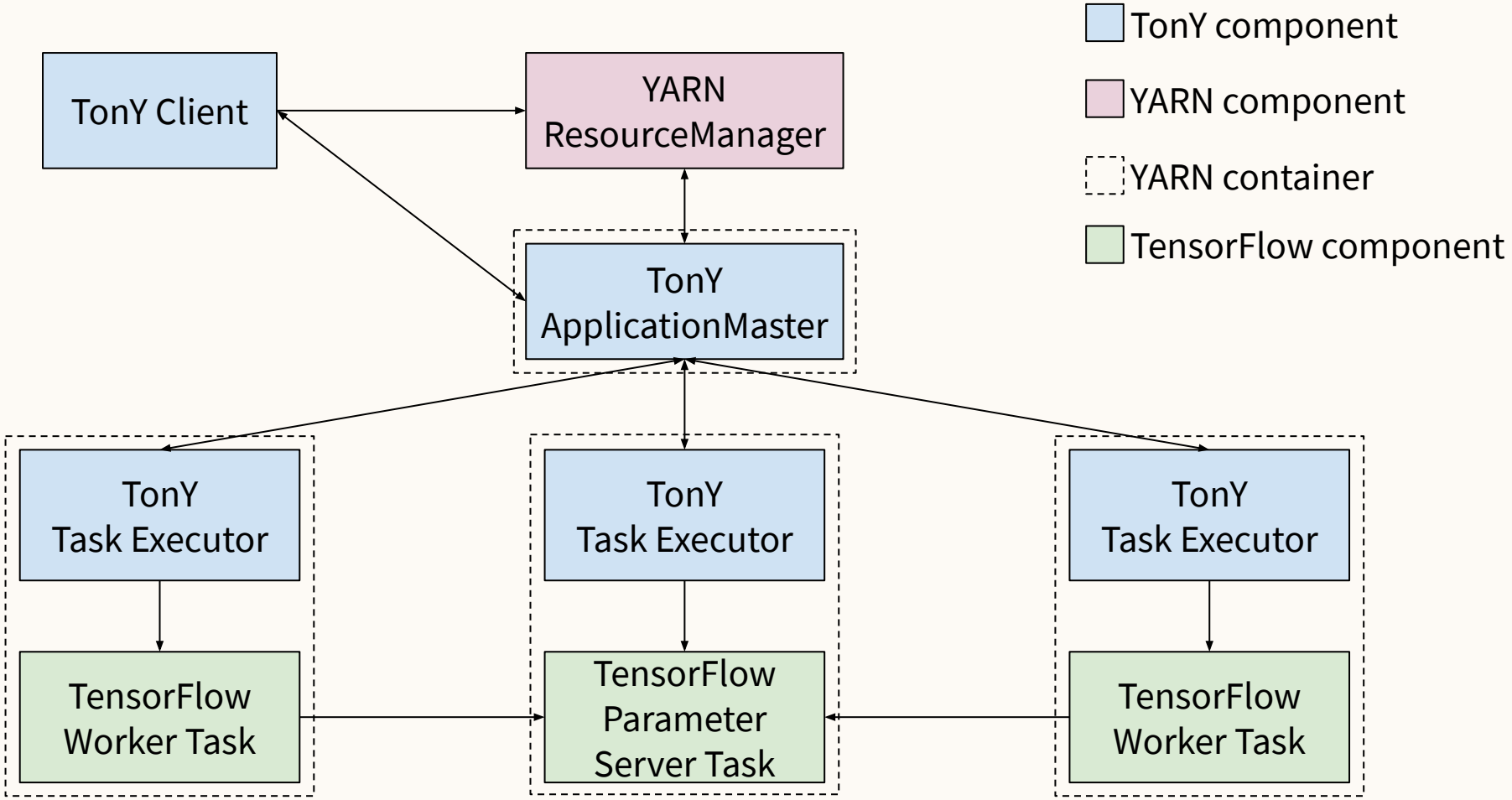
TonY Under the Hood



TonY Under the Hood



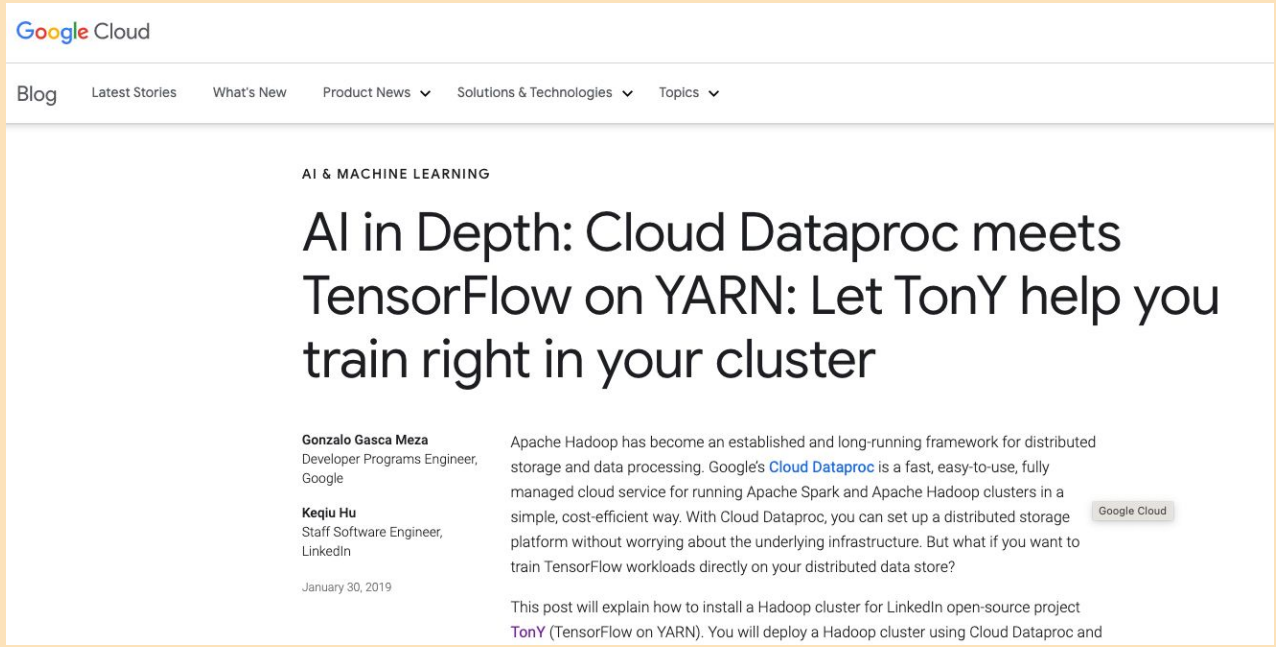
TonY Under the Hood



TonY in Community

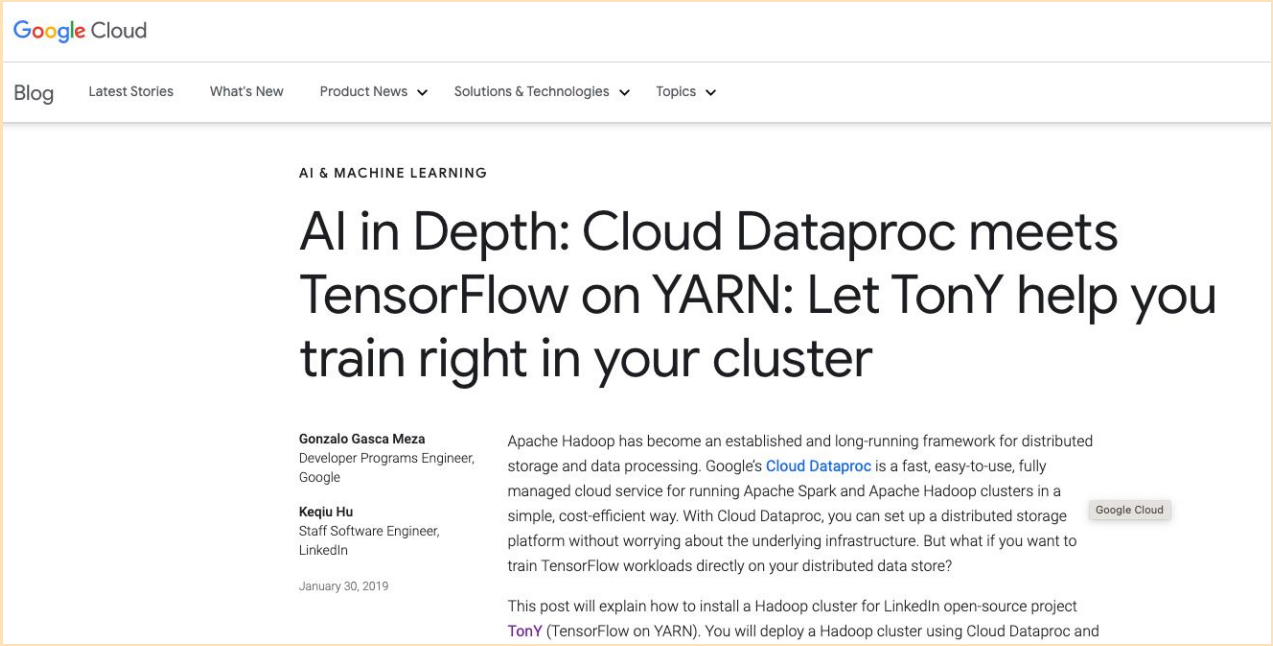
- Open Sourced: https://github.com/linkedin/TonY
- 640 Stars, 150 Forks

- 29 contributors
- Adopted at Google Cloud, iQiyi.com, NetEase, and other companies.
- Horovod integration contribution from iQiyi.



TonY in Community

- TonY @ Strata NY ‘18
- TonY @ AI Conf London ‘19
- Scaling TensorFlow at LinkedIn @ TensorFlow World ‘19
- TonY @ OpML ‘19



Next Step

- Evolve TonY as part of generic AI training platform on Hadoop and Kubernetes.
- History server support.
- Cloud integration.

Thank you

LinkedIn

Summary of open-source solutions

Open-source solution	Pros	Cons
Kubeflow / Kubernetes (Google)	• Large marketplace of libraries and plugins • Active community	• Does not run on Hadoop • May not scale to very large clusters
TensorFlow on Spark (Yahoo!) Spark Deep Learning (Databricks)	• Integrates with Spark	• No GPU resource support until Spark 3.0 (SPARK-20327) • No heterogeneous resource support
TOY: TensorFlow on YARN (Intel) XLearning (Qihoo)	• YARN native, works out-of-the-box	• No GPU resource support
Horovod (Uber)	• Supports synchronous distributed training	• MPI on YARN requires Docker
YARN Native Service	• YARN native	• Distributed TensorFlow requires YARN DNS Registry and ZooKeeper

TAC Vote on Project Incubation Proposal: TonY

Proposed Resolution:

The TAC approves the TonY as an Incubation project of the LF AI & Data Foundation

Next Steps

LF AI & Data staff will work with the TonY project team to onboard the project leading to the announcement of the project joining LF AI & Data

Explore potential integrations between the project and other LF AI & Data projects

Integrate the project with LF AI & Data operations

Annual Project Review - EDL

https://github.com/lfai/proposing-projects/blob/master/proposals/edl.adoc

(contact: Ti Zhou - zhouti@baidu.com)

LF AI & DATA

EDL

Brief Description:

Computing resources on cloud such as Amazon AWS、Baidu Cloud have multi-tenancy. Deep learning model training and inference with elastic resources will be common on cloud. We propose Elastic Deep Learning (EDL) that makes training and inference of deep learning models on cloud easier and more efficient.

TSC Chairperson: Ti Zhou

Contributed by:
Baidu contributed EDL to LF AI in August 2018, and now EDL is the incubating project.

Key Links:

Github: https://github.com/elasticdeeplearning/edl

Homepage: https://elasticdeeplearning.ai/

Artwork: https://artwork.lfaidata.foundation/projects/edl/

Contributions (over last two years)

216.09K
Lines Of Code Changed

158
Commits

11
Contributors

1
No Of Sub Projects

4
Repositories

Top 10 Contributors By Commits — View All

NAME	LINES OF CODE	COMMITS	%
gongweibao	139.01K	60	64.33%
Dong Daxiang	1.61K	52	0.75%
WangXi	56.52K	25	26.16%
wangjiawei04	6.46K	6	2.99%
Ibrahim Haddad	3.85K	5	1.78%
Jiawei Wang	70	3	0.03%
Ti Zhou	681	2	0.32%
jiawei.wang	2.57K	2	1.19%
Tao Luo	8	1	0.00%

Top 10 Organizations By Commits — View All

Key Achievements in the past year

Two Releases

- 0.2.0 fault tolerance
- 0.3.0 knowledge distillation

Enhanced Roadmap

- 0.4.0 live fault tolerance
- 0.5.0 auto scaling

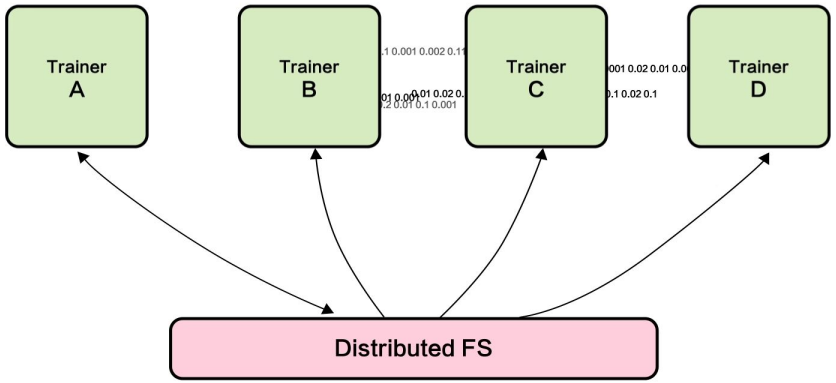
Release 0.2.0 – Fault Tolerance

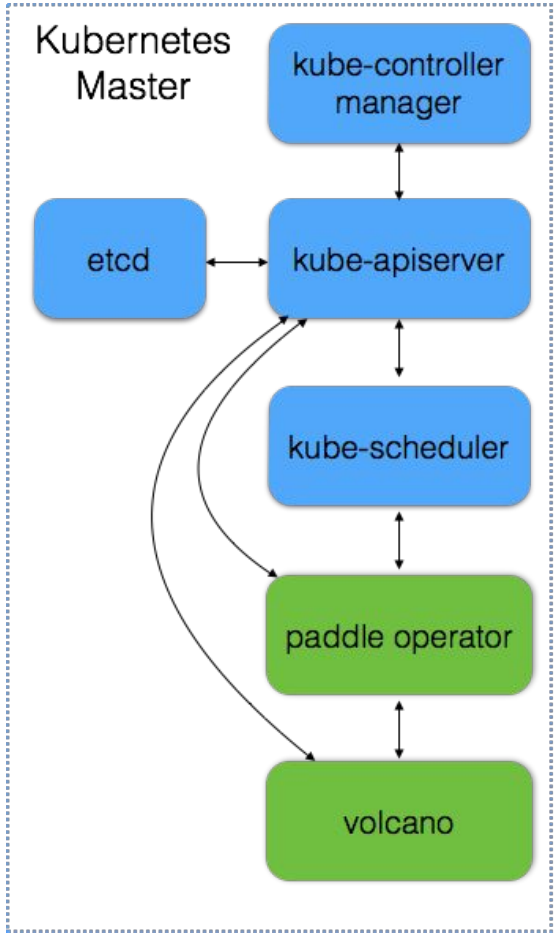
Job recovery with EDL

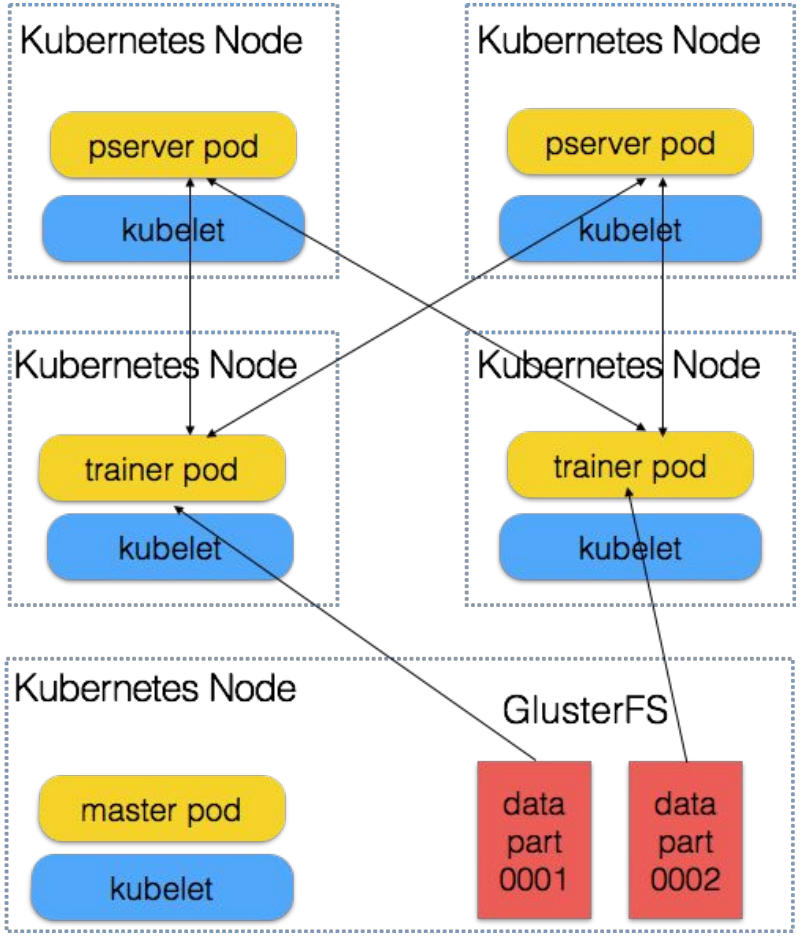
- Snapshots based elastic deep learning
- Training Job can be killed any time
- Job recovery is based on last snapshots

User Interface Design

- Non-intrusive API for PaddlePaddle users
- Minimum user awareness of elastic training







Release 0.3.0 – Knowledge Distillation

Save GPU memory resources

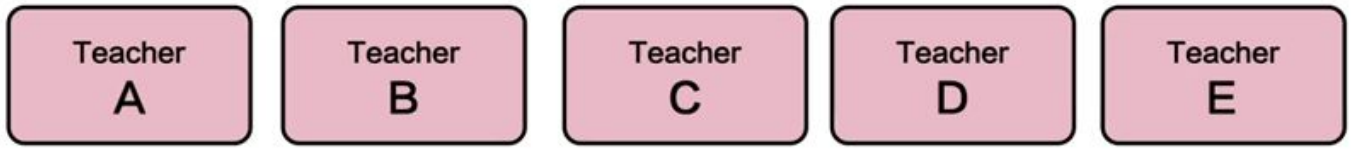
- Heterogeneous resources between teacher and student
- Flexible ratio between teacher and student

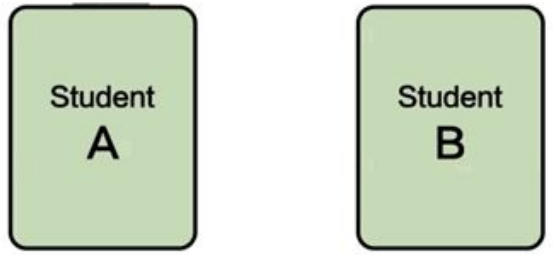
Improve training speed

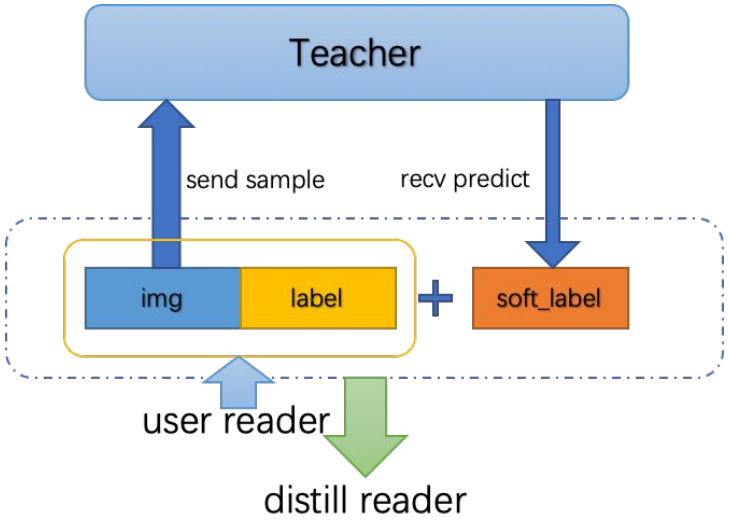
- Larger batch size
- Heterogenous pipelines

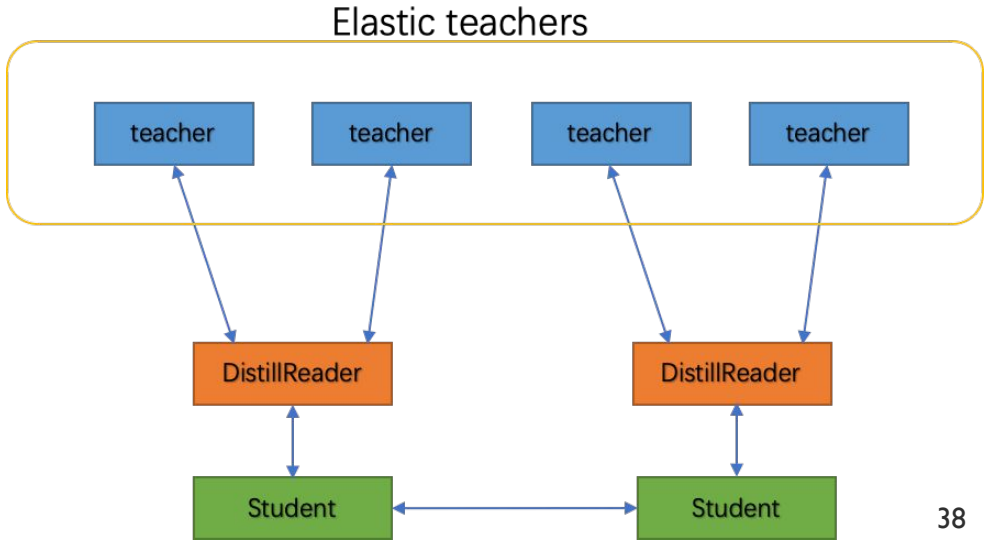
Improve utilization of resources

- Teacher using inference card
- No task failures cased by preemption of online resources
- Make full use of k40 cluster



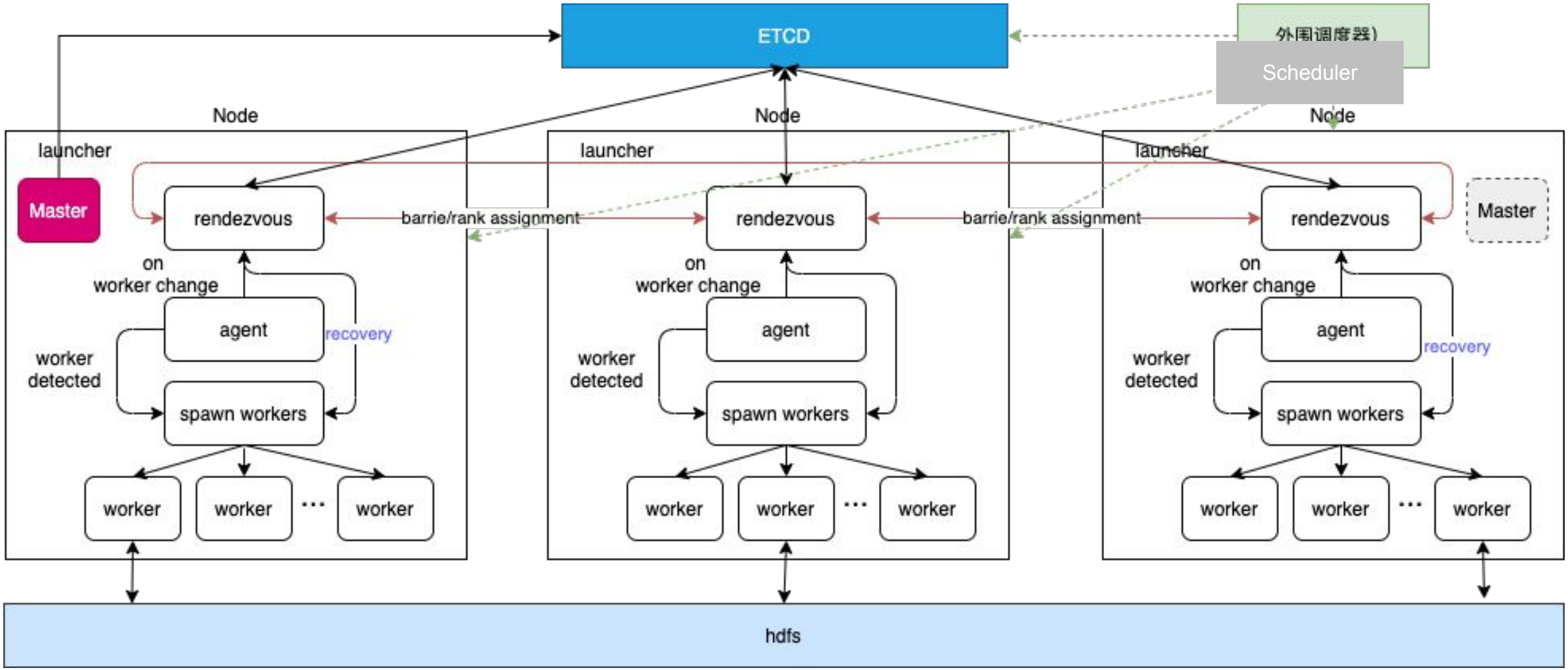






LF AI & DATA

Enhanced RoadMap



Areas the project could use help on

Would like to cooperate with Horovod's elastic training feature, to provide users the state-of-the-art elastic training on both collective mode and parameter server mode.

Feedback on working with LF AI & Data

Benefit from LF AI & Data Slack, Zoom, Mailing Lists for WG meetings, etc.

Value opportunities to work with related projects (ONNX/Horovod/etc).

LF AI & Data - Ongoing Annual Project Reviews

LF AI & DATA

Annual Review schedule

Date	Project
April 6, 2021	Egeria
April 6, 2021	OpenDS4all
May 20, 2021	ONNX
July 15, 2021	Acumos
July 29, 2021	Angel
July 29, 2021	Adlik
Aug 26, 2021	EDL
Aug 26, 2021	Sparklyr
Sept 9, 2021	Marquez
Sept 9, 2021	Milvus
Sept 23, 2021	NNStreamer
Sept 23, 2021	ForestFlow
Oct 7, 2021	Ludwig
Oct 7, 2021	Amundsen
Oct 21, 2021	AI Fairness 360
Oct 21, 2021	AI Explainability 360
Oct 21, 2021	Adversarial Robustness Toolbox
Nov 4, 2021	Horovod
Nov 4, 2021	FEAST
Nov 18, 2021	SOAJS
Nov 18, 2021	Delta
Dec 2, 2021	DataPractices.org
Dec 2, 2021	JanusGraph
Dec 16, 2021	Pyro
Jan 6, 2021	Datashim
Jan 6, 2022	Flyte
Jan 20, 2022	RosaeNLG
Jan 20, 2022	SubstraFramework
	MLX
	VulcanKompute

Schedule: https://wiki.lfaidata.foundation/pages/editpage.action?pageId=43286684

LF AI & Data - General Updates

LF AI & DATA

LF AI & DATA Landscape

309 projects
2.2M+ GitHub Stars
450M+ LoC
40K+ developers

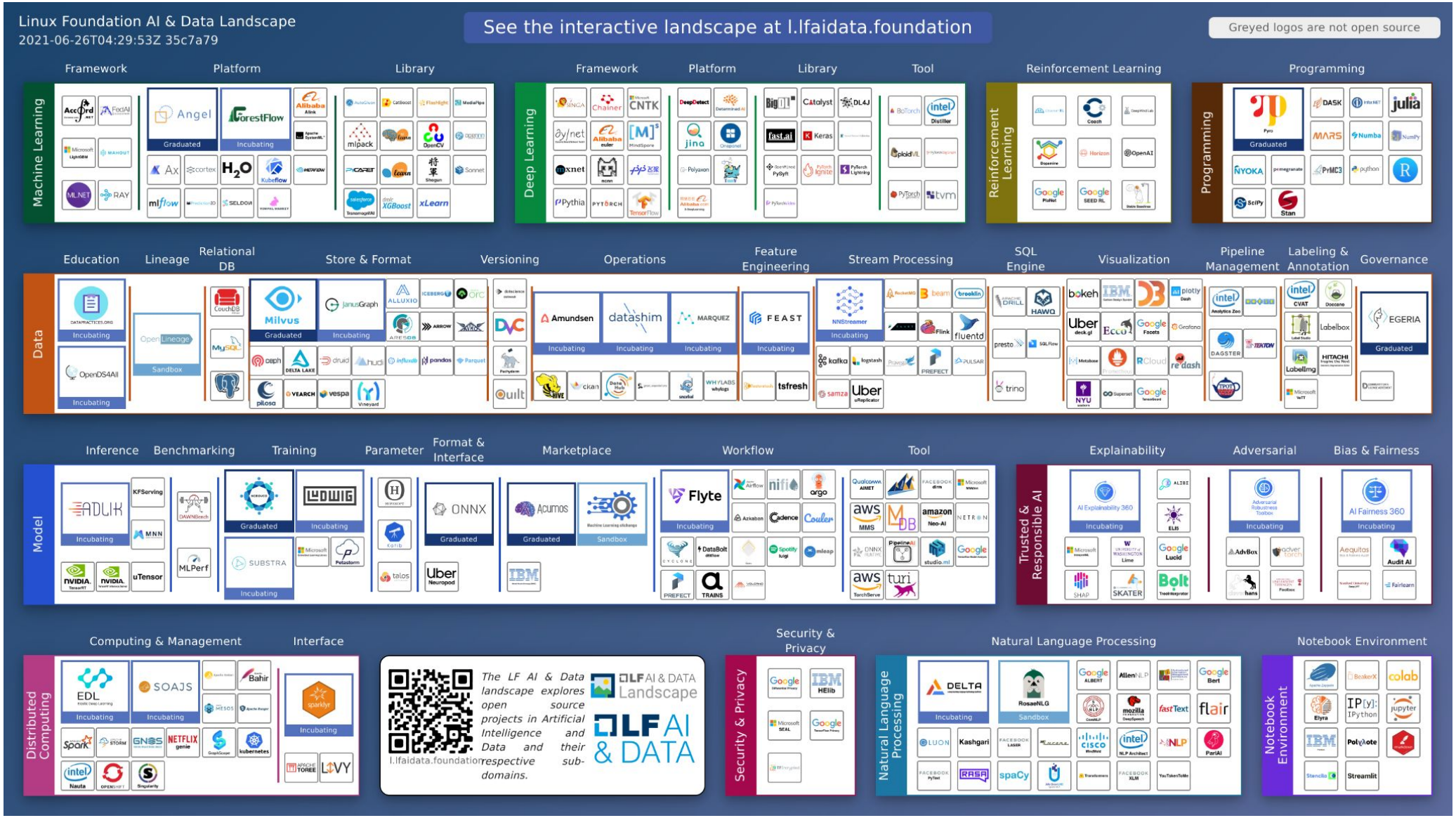
140+ founding orgs

10+ universities

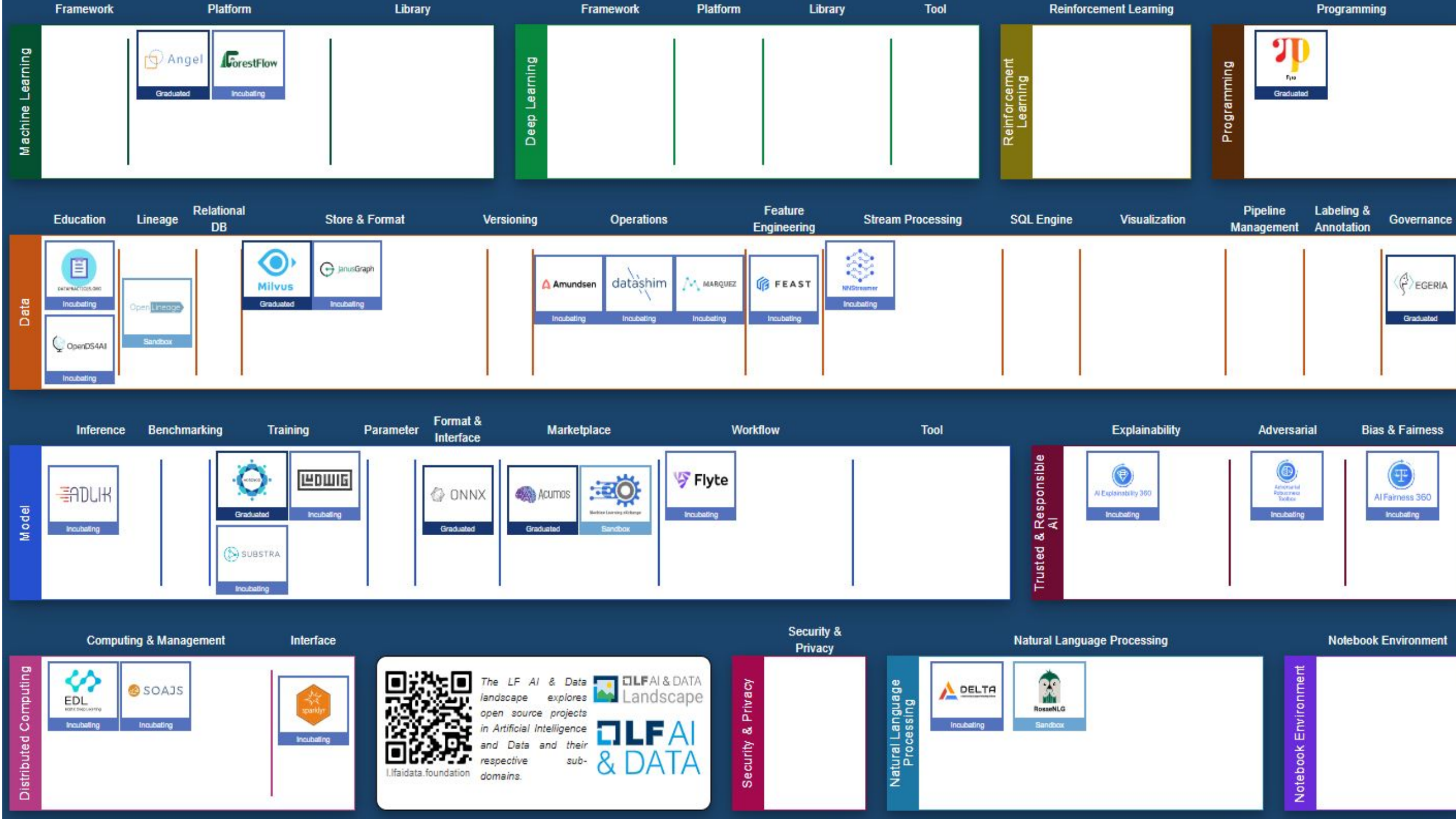
13 licenses

>450M LoC

>1M new LoC weekly

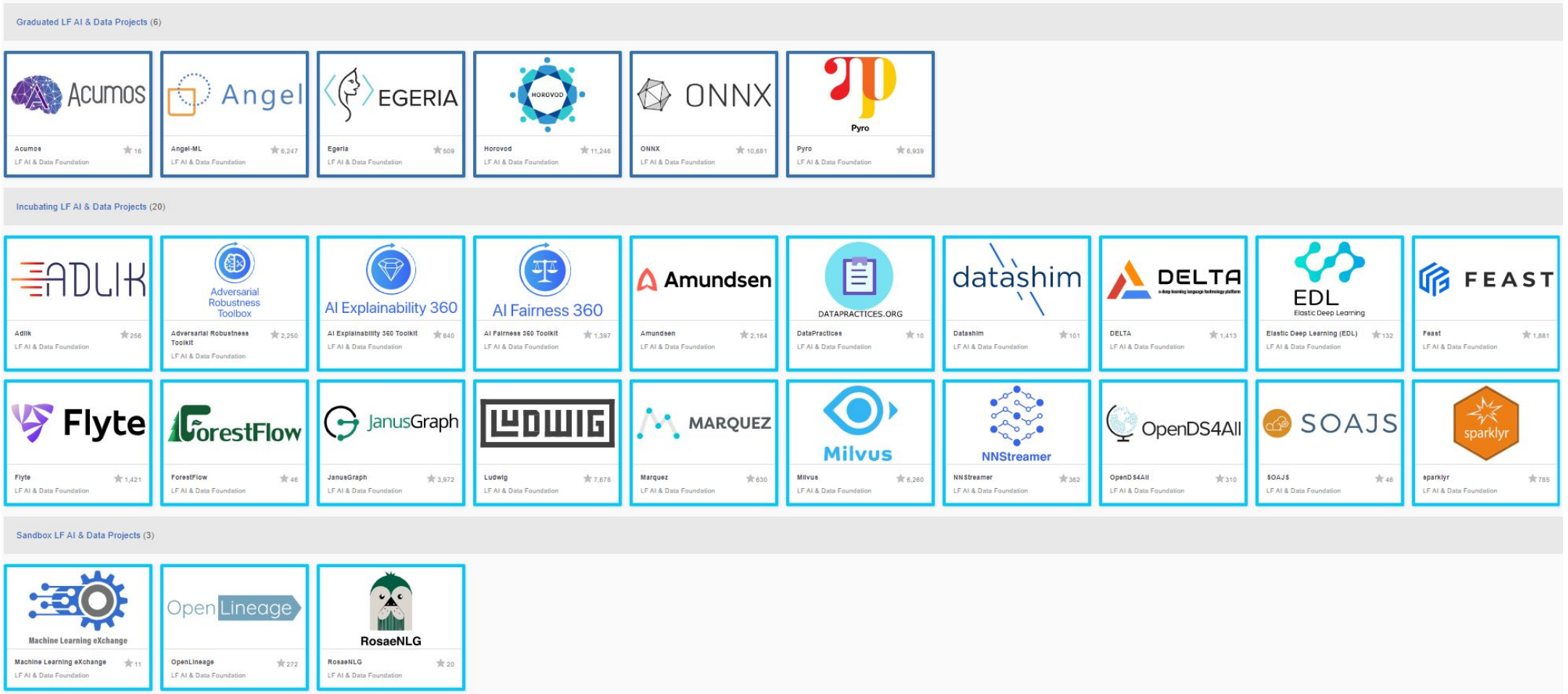


Help us improve the landscape:
https://github.com/lfai/lfai-landscape/


Machine Learning
Framework
Platform
Library
Angel
Graduated
ForestFlow
Incubating
Deep Learning
Framework
Platform
Library
Tool
Reinforcement Learning
Reinforcement Learning
Programming
Programming
Fyle
Graduated
Data
Education
Incubating
OpenDS4All
Incubating
Lineage
OpenLineage
Sandbox
Relational DB
Milvus
Graduated
Store & Format
JanusGraph
Incubating
Versioning
Operations
Amundsen
Incubating
datashim
Incubating
MARQUEZ
Incubating
Feature Engineering
FEAST
Incubating
Stream Processing
NNStreamer
Incubating
SQL Engine
Visualization
Pipeline Management
Labeling & Annotation
Governance
EGERIA
Graduated
Model
Inference
ADLIK
Incubating
Benchmarking
Training
Graduated
LUDWIG
Incubating
SUBSTRA
Incubating
Parameter
Format & Interface
ONNX
Graduated
Marketplace
Acumos
Graduated
Sandbox
Workflow
Flyte
Incubating
Tool
Trusted & Responsible AI
Explainability
AI Explainability 360
Incubating
Adversarial
Incubating
Bias & Fairness
AI Fairness 360
Incubating
Distributed Computing
Computing & Management
EDL
Incubating
SOAJS
Incubating
Interface
Incubating
l.lfaidata.foundation
The LF AI & Data landscape explores open source projects in Artificial Intelligence and Data and their respective sub-domains.
LF AI & DATA Landscape
LF AI & DATA
Security & Privacy
Security & Privacy
Natural Language Processing
Natural Language Processing
DELTA
Incubating
RosaeNLG
Sandbox
Notebook Environment
Notebook Environment

Projects

https://landscape.lfai.foundation/card-mode?project=company



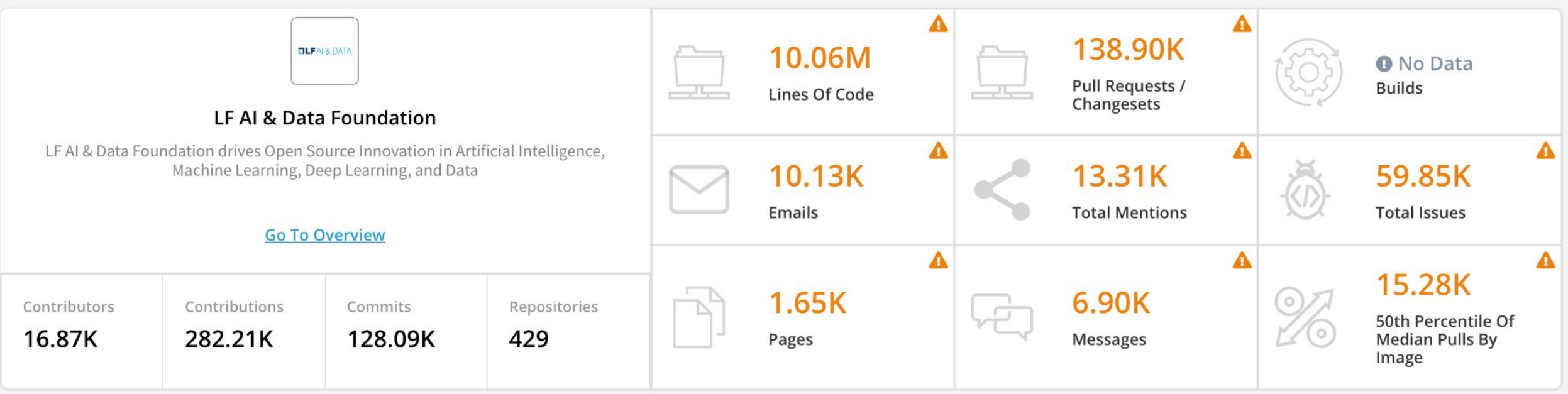
* Missing Substra Framework (pending logo redesign)

New projects in 2021

→ Track incoming proposals via: https://github.com/lfai/proposing-projects

1. **Datashim:** Open source enablement and acceleration of data access for Kubernetes/Openshift workloads in a transparent and declarative way
2. **Flyte:** Production-grade, declarative, structured and highly scalable cloud-native workflow orchestration platform
3. **RosaeNLG:** Open source project, template-based Natural Language Generation (NLG) automating the production of relatively repetitive texts based on structured input data and textual templates, run by a NLG engine
4. **Substra Framework:** Low-layer framework, offering secure, traceable, distributed orchestration of machine learning tasks among partners.
5. **ML eXchange:** Data and AI Assets Catalog and Execution Engine
6. **Kompute: Blazing fast, mobile-enabled, asynchronous, and optimized for advanced GPU processing usecases.**
7. **Open Lineage:** Open standard for metadata and lineage collection designed to instrument jobs as they are running

Active and growing developer community



Looking to host a project with LF AI & Data

- Hosted project stages and life cycle:
  https://lfaidata.foundation/project-stages-and-lifecycle/
- Offered services for hosted projects:
  https://lfaidata.foundation/services-for-projects/
- Contact:
  Jim Spohrer (TAC Chair) and Ibrahim Haddad (ED, LF AI & Data)

Promoting Upcoming Project Releases

We promote project releases via a blog post and on LF AI & Data Twitter and/or LinkedIn social channels

For links to details on upcoming releases for LF AI & Data hosted projects visit the Technical Project Releases wiki

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

Note on quorum

As LF AI & Data is growing, we now have 18 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.

Upcoming TAC Meetings

LF AI & DATA

Upcoming TAC Meetings (Tentative)

- July 29: TBD - Annual project review + TBD
- Aug 12: TBD - Annual project review + TBD

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

TAC Meeting Details

- To subscribe to the TAC Group Calendar, visit the wiki: https://wiki.lfaidata.foundation/x/cQB2
- 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

LF AI & DATA

Legal Notice

- 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.

LF AI & DATA