Meeting of the LF AI & Data Technical Advisory Council (TAC)

July 27, 2023

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 (1 mins)
- Approval of Minutes from previous meeting (2 mins)
- DocArray update
- OpenLineage Graduation proposal
- Open Discussion

TAC Voting Members - Please note

Please ensure that you do the following to facilitate smooth procedural quorum and voting processes:

- Change your Zoom display name to include your First/Last Name, Company/Project Represented
  - example: Nancy Rausch, SAS
- State your First/Last Name and Company/Project when submitting a motion
  - example: First motion, Nancy Rausch/SAS

TAC Voting Members - Please note

- TAC members must attend consistently to maintain their voting status
- After 2 absences voting members will lose voting privileges
- Voting privileges will only be reinstated after attending 2 meetings in a row

TAC Voting Members

Note: we still need a few designated backups specified on [wiki]

Member Company or Graduated Project	Membership Level or Project Level	Voting Eligibility	Country	TAC Representative	Designated TAC Representative Alternates
4paradigm	Premier	Voting Member	China	Zhongyi Tan	
Baidu	Premier	Voting Member	China	Jun Zhang	Daxiang Dong, Yanjun Ma
Ericsson	Premier	Voting Member	Sweden	Rani Yadav-Ranjan	
Huawei	Premier	Voting Member	China	Howard (Huang Zhipeng)	Charlotte (Xiaoman Hu), Leon (Hui Wang)
Nokia	Premier	Voting Member	Finland	@ Michael Rooke	@ Jonne Soininen
OPPO	Premier	Voting Member	China	Jimmy (Hongmin Xu)	
SAS	Premier	Voting Member	USA	*Nancy Rausch	Liz McIntosh
ZTE	Premier	Voting Member	China	Wei Meng	Liya Yuan
Adversarial Robustness Toolbox Project	Graduated Technical Project	Voting Member	USA	Beat Buesser	Kevin Eykholt
Angel Project	Graduated Technical Project	Voting Member	China	Jun Yao	
Egeria Project	Graduated Technical Project	Voting Member	UK	Mandy Chessell	Nigel Jones, David Radley, Maryna Strelchuk, Ljupcho Palashevski, Chris Grote
Flyte Project	Graduated Technical Project	Voting Member	USA	Ketan Umare	
Horovod Project	Graduated Technical Project	Voting Member	USA	Travis Addair	
Milvus Project	Graduated Technical Project	Voting Member	China	Xiaofan Luan	Jun Gu
ONNX Project	Graduated Technical Project	Voting Member	USA	Alexandre Eichenberger	Andreas Fehlner, Prasanth Pulavarthi, Jim Spohrer
Pyro Project	Graduated Technical Project	Voting Member	USA	Fritz Obermeyer	

Minutes approval

LF AI & DATA

Approval of July 13, 2023 Minutes

Draft minutes from the July 13 TAC call were previously distributed to the TAC members via the mailing list

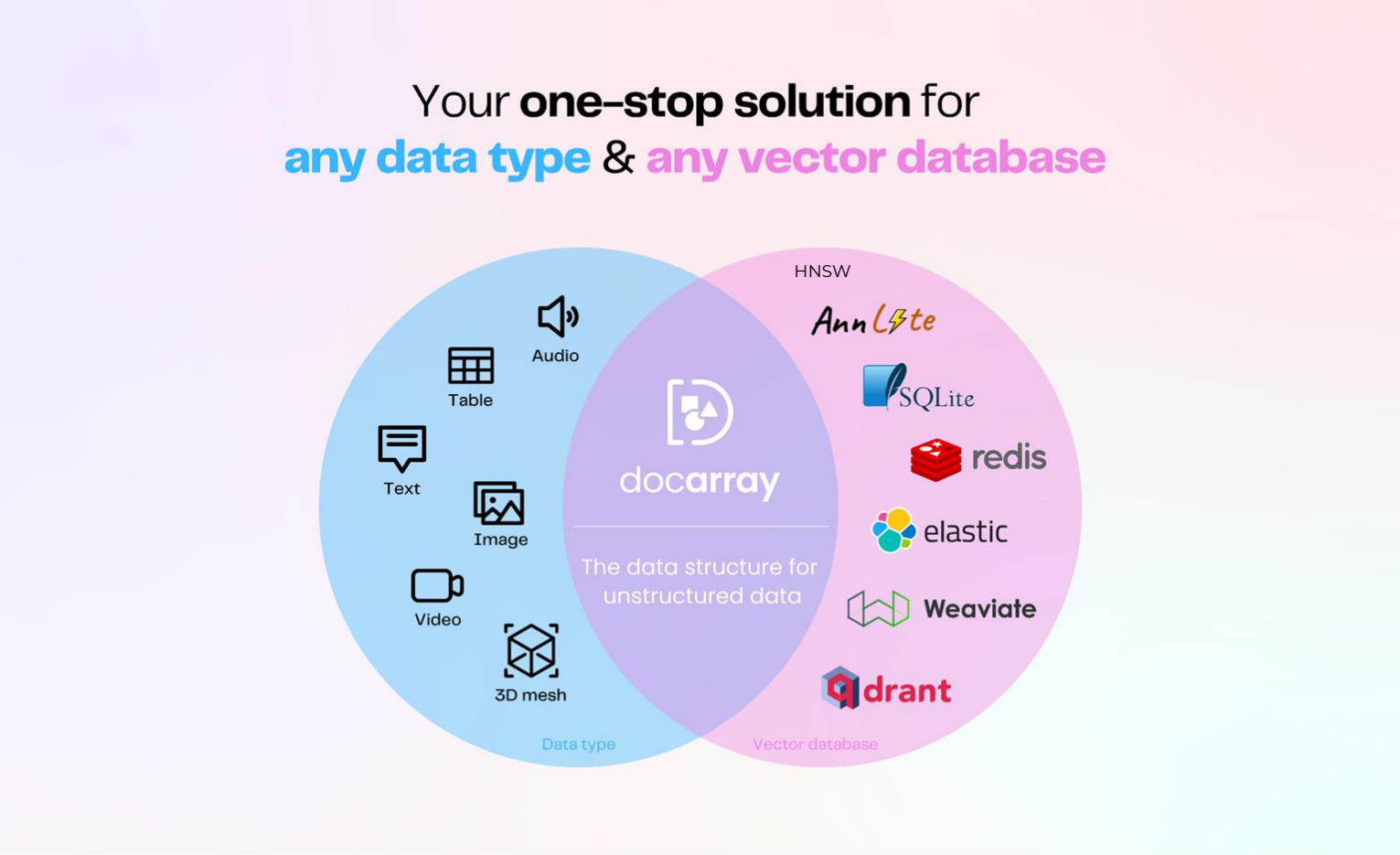
Proposed Resolution:

- That the minutes of the July 13 meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.

DocArray @ LF AI & DATA

Incubation stage

DocArray in a nutshell



LF AI & DATA

DocArray since it joined LF AI & DATA

2022.1

First Release

Built internally by Jina AI

2022.9

Reached 1024 stars

2022.11

Donated to LF AI & DATA

40+ contributors

100k monthly downloads

2023.4

Released v0.30

Complete rewrite to improve the usability.

NOW

Apply for Incubation Stage

2.4k stars (**+100%**)

60+ contributors (**+50%**)

140k monthly downloads (**+40%**)

Impact

Who uses DocArray?

2.400k+	60+	570+	2,000+	140,000+
GitHub stars	Contributors	Used-by	Monthly active docs users	Monthly downloads

Trusted by companies of all sizes and needs

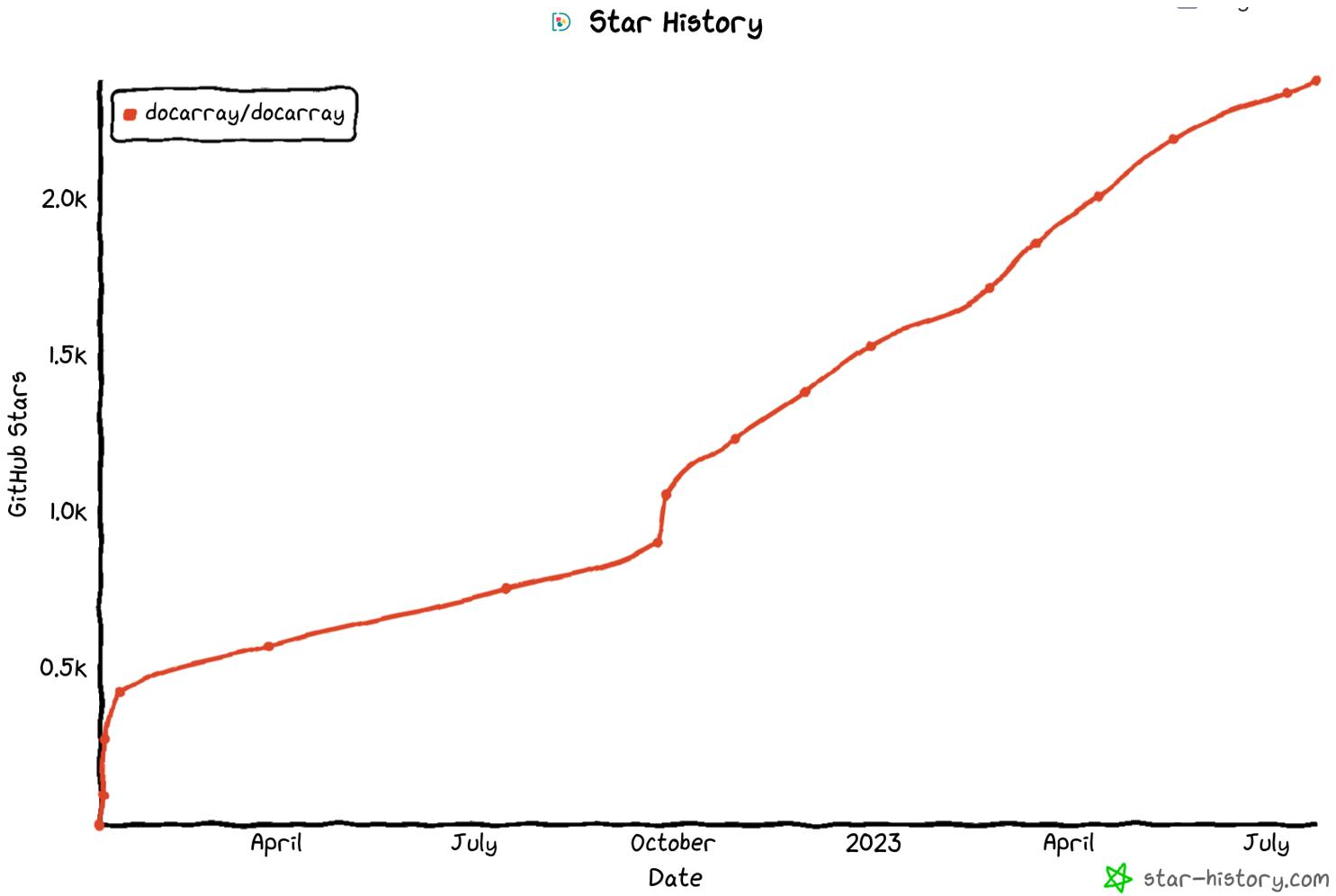
WordLift
HuForce
YAHAHA
masernet
LottieFiles
PBRMAX
ALGOMO
uizard
myntor
salezilla
7dish
FDP Connect ™
OIGA TECHNOLOGIES
YOUR COMPETITIVE ADVANTAGE
BetterStudy.ai

Impact

☆ Overview

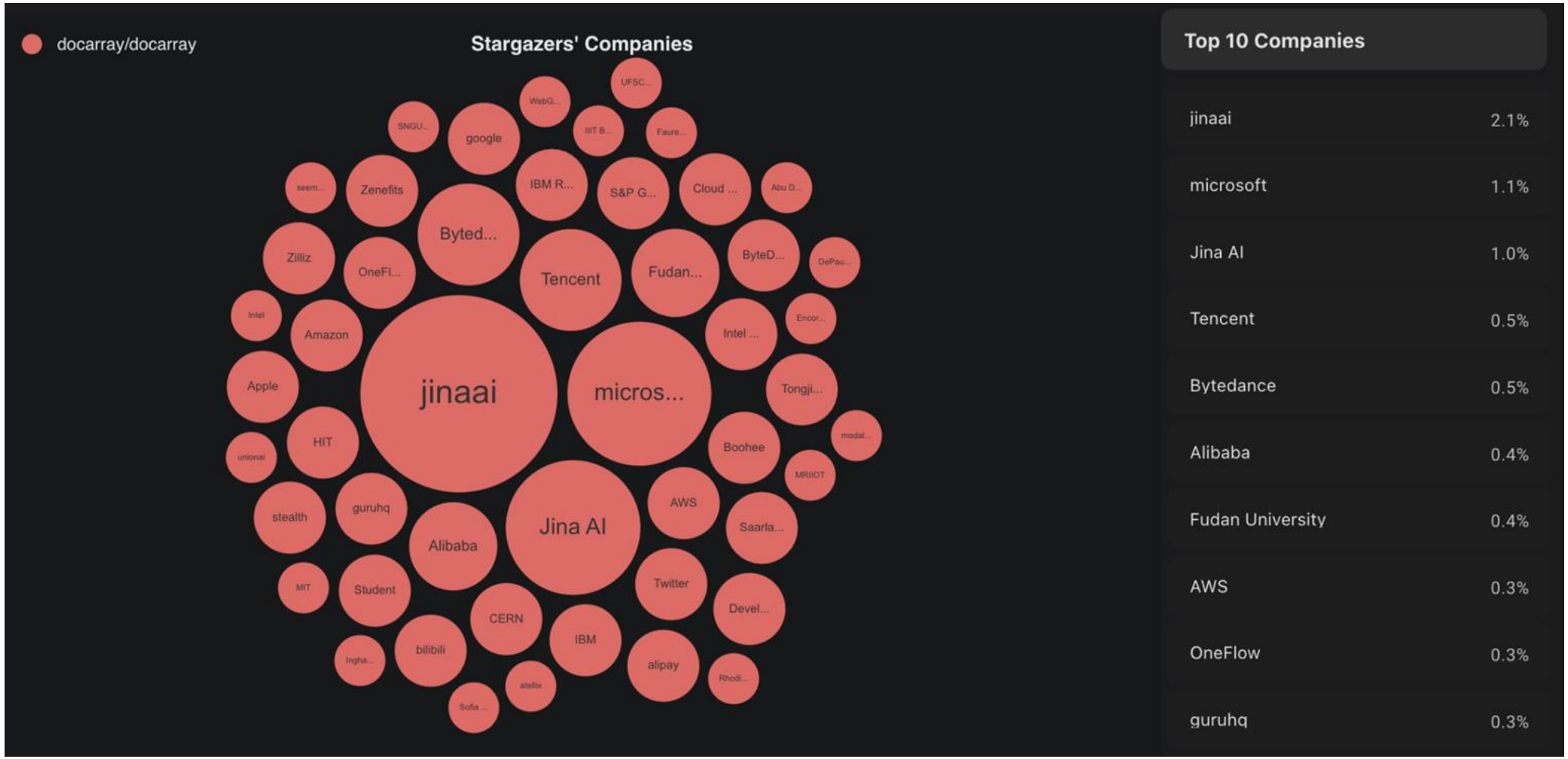
Stars	2280
Commits	6522
Issues	567
Forks	187
PR Creators	69

Data from OSSInsight, 2023.7.25



Impact

☆ Stargazers from Big Techs



LF AI & DATA

Who is contributing to DocArray ?

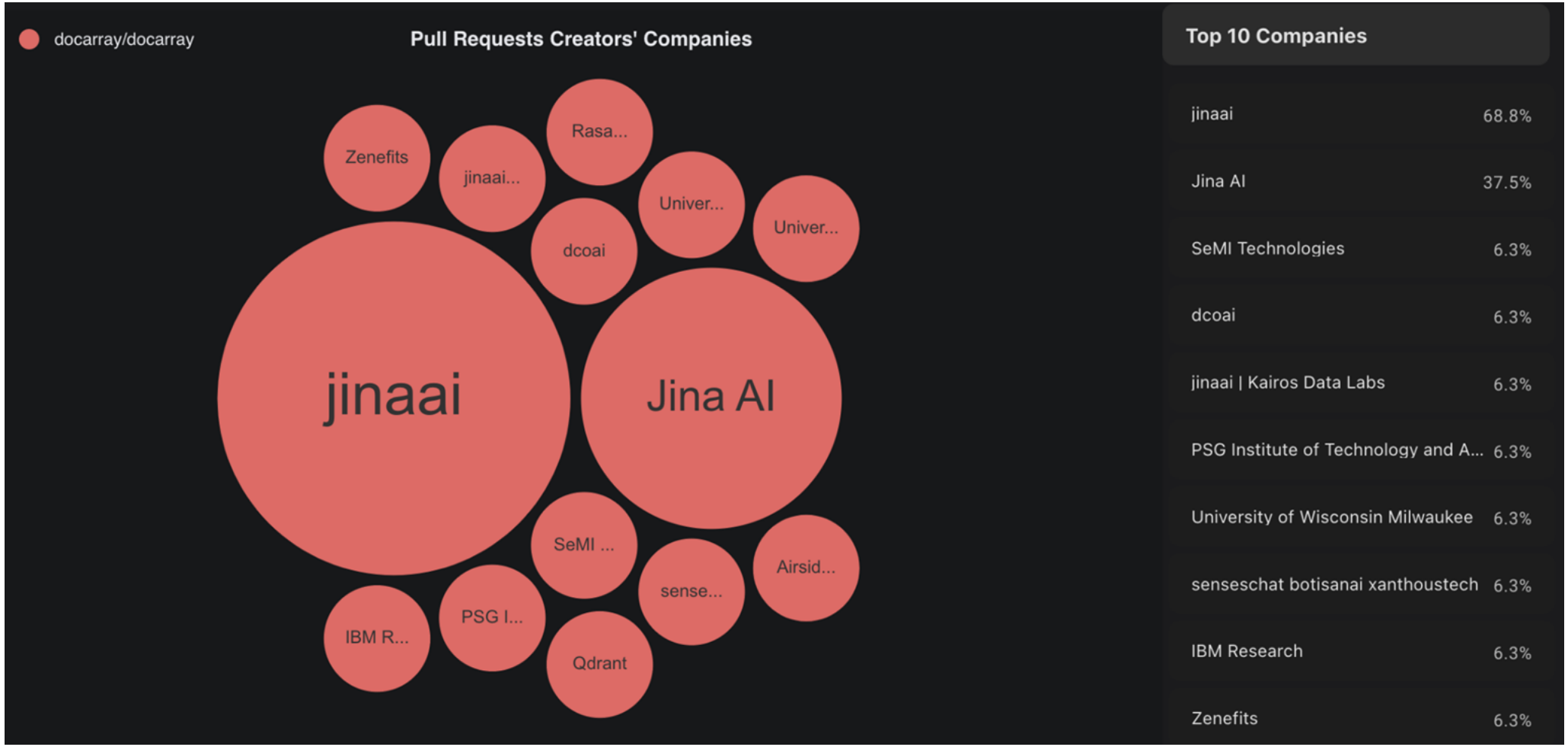
- **3 organizations are actively contributing to DocArray,** i.e. fixing issues, publishing articles, and participating in TSC meetings to discuss the project's future:
  - Jina AI
  - Qdrant
  - Weaviate

- Several more contribute from time to time:
  - Redis
  - IBM

- Individual contributors:
  - Two on-going Google Summer of Code collaborations
  - Dozens of contributors for fixes, documentation, etc ...

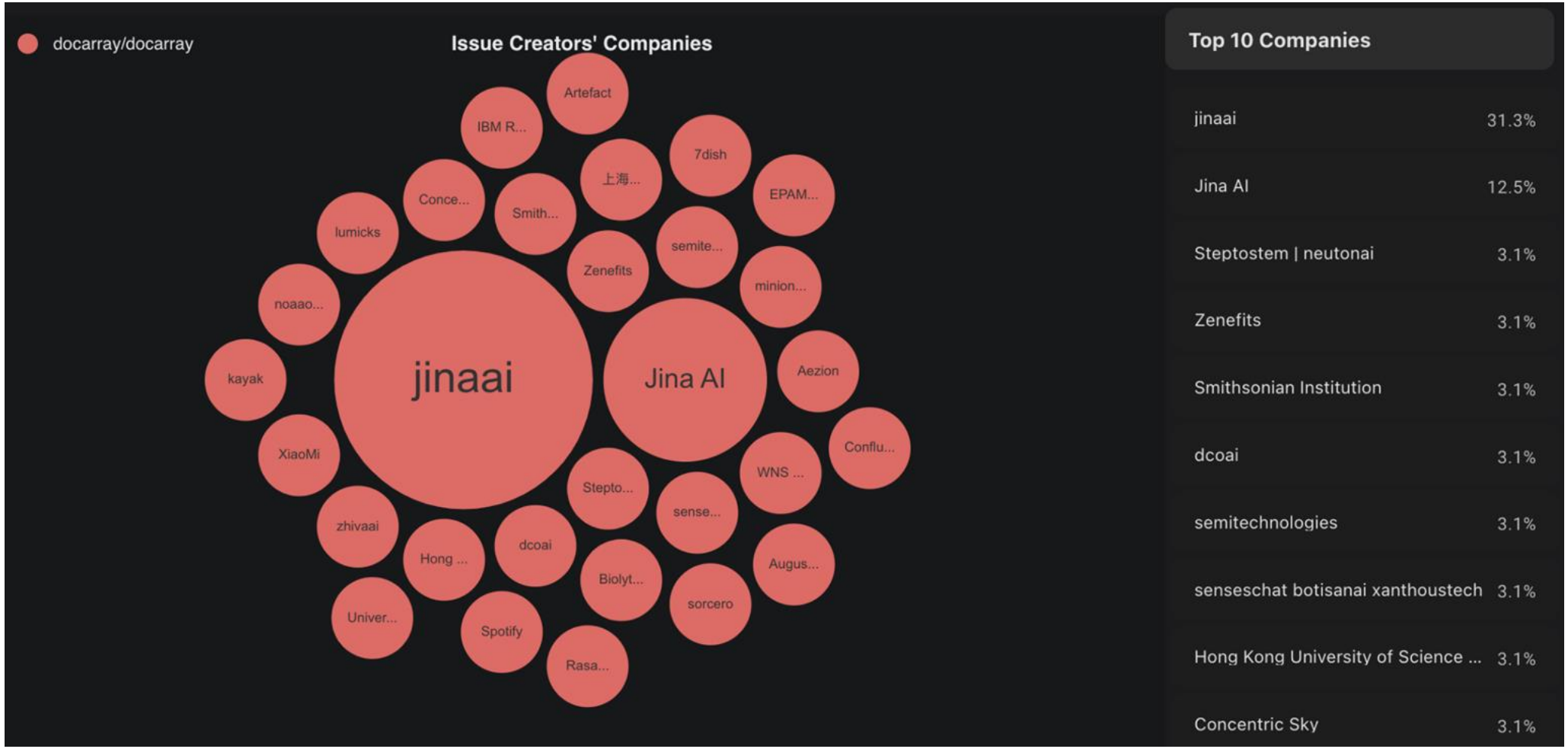
Impact

PRs from various companies



Impact

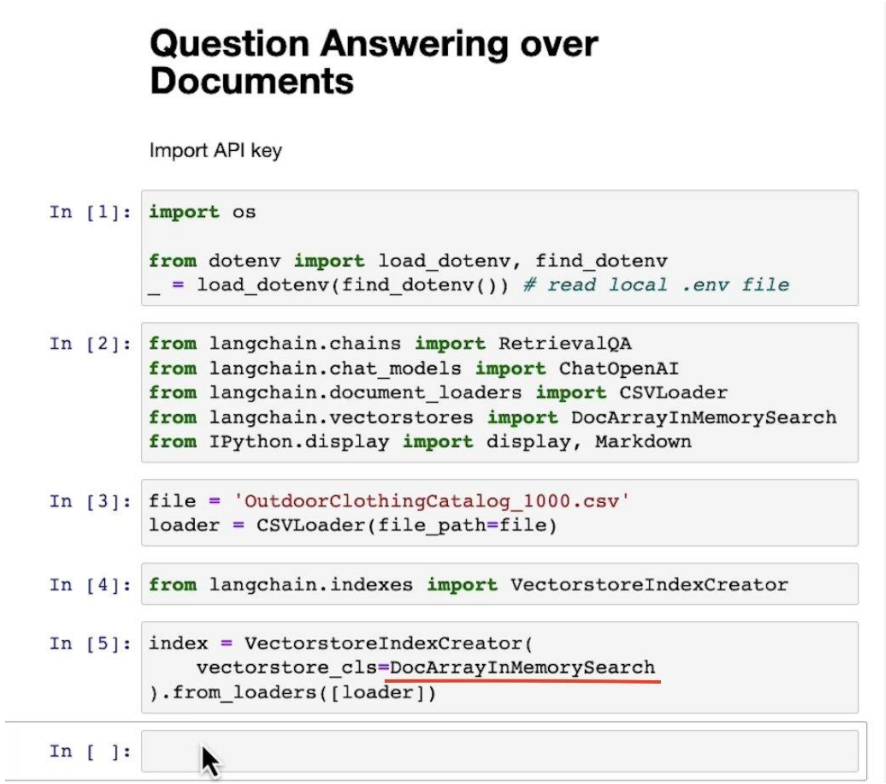
Issues from various companies



LF AI & DATA

DocArray integration in the AI landscape

- Popular packages integrate DocArray
  - LangChain
  - LLama index
  - GptCache
- Adoption in the academy
  - Andrew Ng used DocArray in courses at deeplearning.ai !
  - DocArray is used in CVPR2023 Tutorial: Neural Search in Action



Question Answering over Documents

Import API key

```
In [1]: import os

        from dotenv import load_dotenv, find_dotenv
        _ = load_dotenv(find_dotenv()) # read local .env file

In [2]: from langchain.chains import RetrievalQA
        from langchain.chat_models import ChatOpenAI
        from langchain.document_loaders import CSVLoader
        from langchain.vectorstores import DocArrayInMemorySearch
        from IPython.display import display, Markdown

In [3]: file = 'OutdoorClothingCatalog_1000.csv'
        loader = CSVLoader(file_path=file)

In [4]: from langchain.indexes import VectorstoreIndexCreator

In [5]: index = VectorstoreIndexCreator(
            vectorstore_cls=DocArrayInMemorySearch
        ).from_loaders([loader])

In [ ]:
```

LF AI & DATA

Technical Steering Committee

- TSC voting members:
  - Alaeddine Abdessalem
  - Han Xiao
  - Joan Fontanals Martínez (Chair)
  - Johannes Messner
  - Nan Wang
  - Sami Jaghouar

- TSC Events
  - The TSC meeting is hold every month:
  - Public communication on Discord
  - Public summary on Notion

- Contributing Guidelines
  - https://github.com/docarray/docarray/blob/main/CONTRIBUTING.md
- Governance Information
  - https://github.com/docarray/docarray/blob/main/GOVERNANCE.md

LF AI & DATA

Silver OpenSSF best practice Badge

Link to the badge :
https://bestpractices.coreinfrastructure.org/en/projects/6554?criteria_level=1

docarray

LF AI & DATA

Looking forward to your support in welcoming DocArray to the Incubation Stage!

Q & A

Approval of DocArray to move from Sandbox to Incubation

Proposed Resolution:

- DocArray as an Incubation project of the LF AI & Data Foundation is hereby approved.

OpenLineage LFAI Graduation Application 2023

Julien Le Dem
github: @julienledem
OpenLineage Project Lead

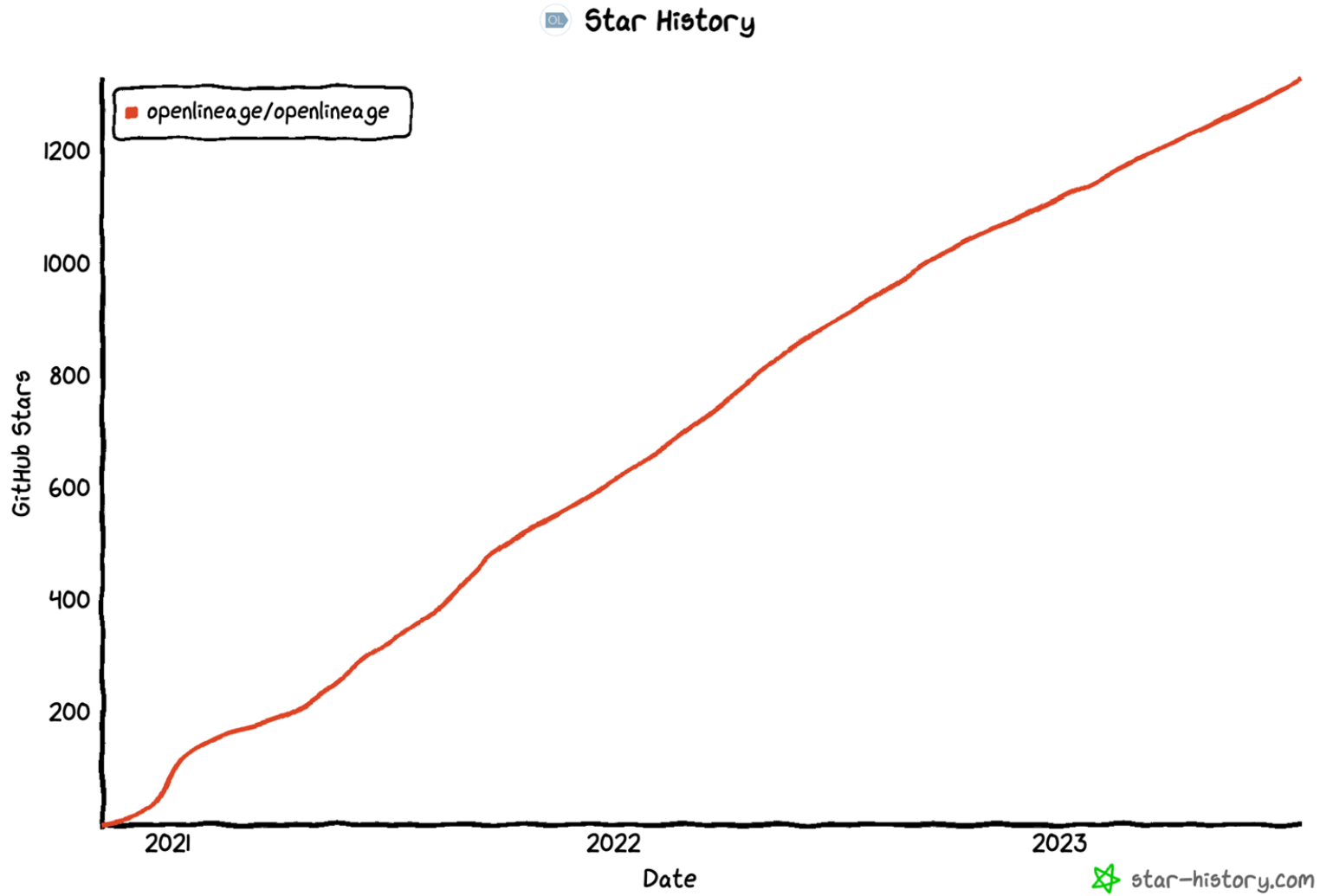
LF AI & DATA

OpenLineage

Background

- **December 2020:** open sourced
- **May 2021:** joined the LFAI
  - 300 GitHub stars
  - Collaboration with Marquez
  - Sandbox status
- **December 2022:** advanced to Incubation status
  - 1.1K GitHub stars
  - Collaborations with: Microsoft, Snowflake, Airflow, Egeria, Northwestern Mutual, Amundsen, GX
- **Today:** ready for graduation
  - 1.3K GitHub stars
  - 1K+ Slack members
  - Collaborations with DataGalaxy, Atlan, Manta, Keboola, Matillion, Microsoft, Snowflake, Airflow, Egeria, Northwestern Mutual, Amundsen, GX

Adoption


Star History
openlineage/openlineage
1200
1000
800
600
400
200
GitHub Stars
2021
2022
2023
Date
star-history.com

Adoption since 2021

	December 2021	December 2022	July 2023	Change
Stars	600	1.1K	1.3K	**+117%**
Forks	44	133	180	**+309%**

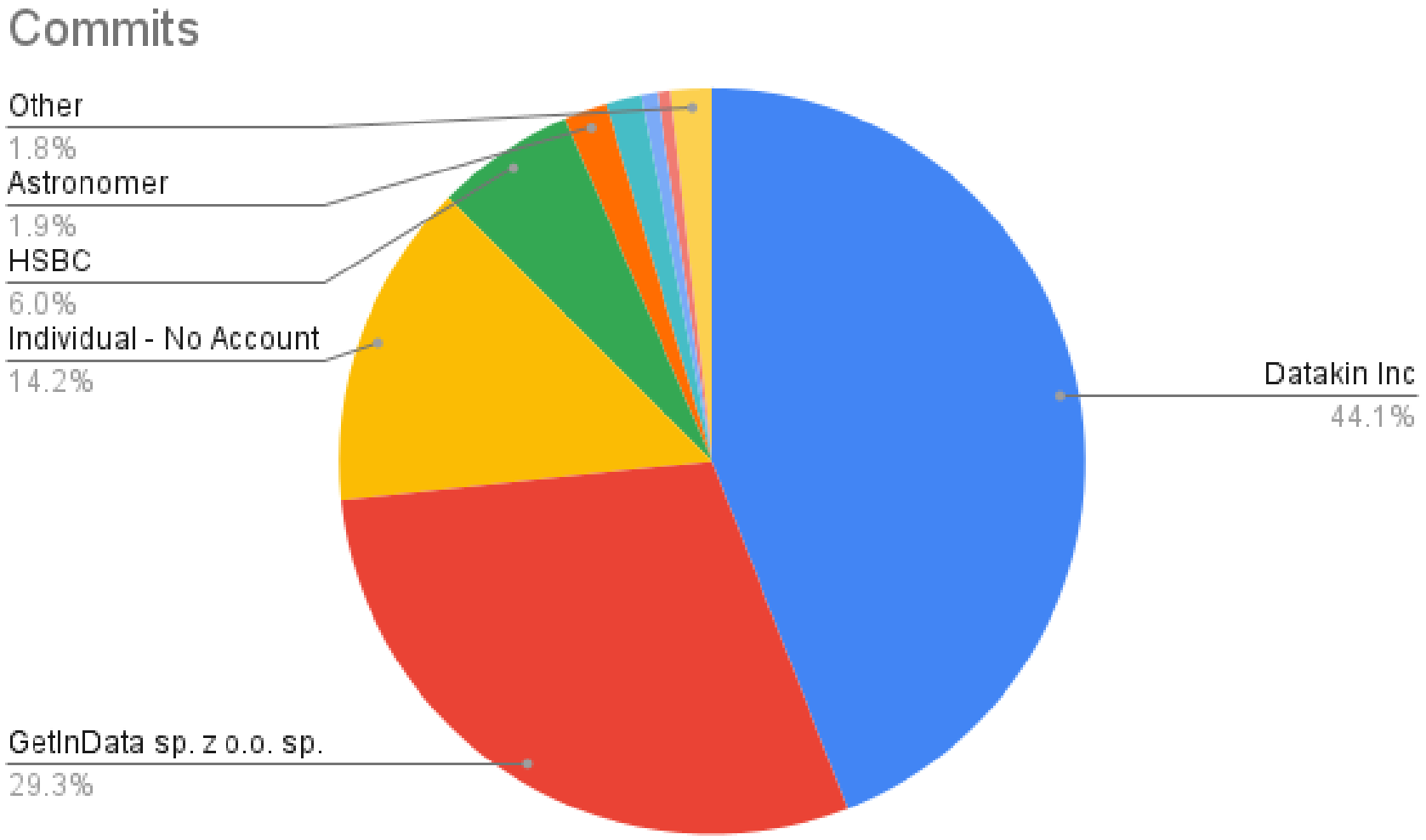
Progress since 2021

- **Contributors**
  - **266** total contributors (DataGalaxy, Bloomberg, Microsoft, Astronomer, HSBC, UBS, Northwestern Mutual, NTT Data Deutschland, others)
- **Best Practices**
  - CII Best Practices Gold Badge
- **Commits**
  - **4.8K**, averaging **~41** per week in the past 12 months
- **Governance and Procedures**
  - TSC, Contributor Guide, Charter, CoC, Code Quality Assurance
- **LFAI Collaborations**
  - Egeria, Marquez, Amundsen

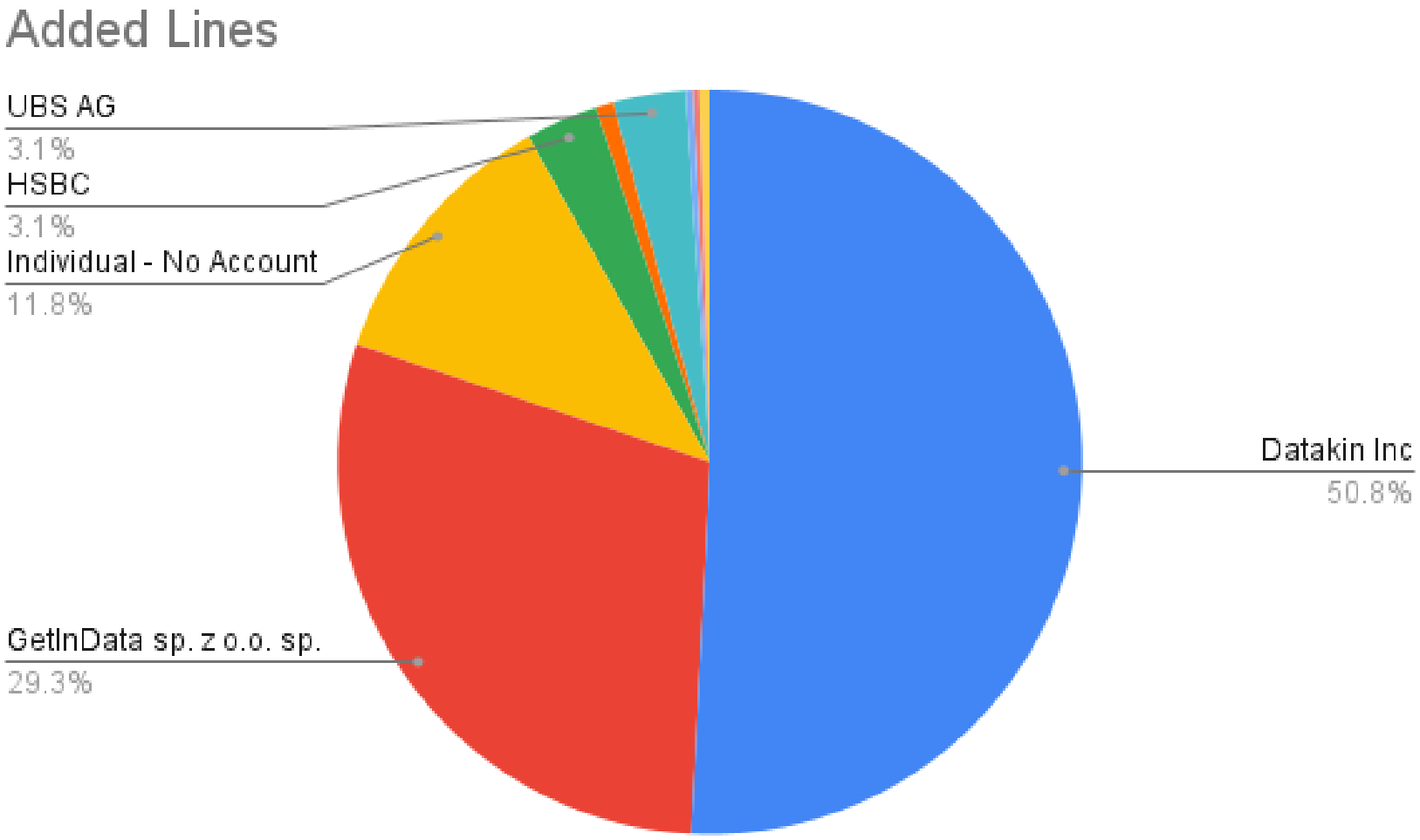
Top Contributing Organizations

Organization	Commits	Added Lines
Datakin Inc	1255	684254
GetInData sp. z o.o. sp. k.	833	395127
Individual - No Account	403	158826
HSBC	172	42307
Astronomer	53	10245
UBS AG	44	42307
NTT DATA Deutschland GmbH	20	4488
WeWork Companies LLC	16	3640

Commits


Commits
Other
1.8%
Astronomer
1.9%
HSBC
6.0%
Individual - No Account
14.2%
Datakin Inc
44.1%
GetInData sp. z o.o. sp.
29.3%

Lines of Code Changed



Top Contributors: Deltas

	December 2021	December 2022	July 2023	Change
% External Commits	15%	40%	43%	**+187%**
% External LOC	13%	57%	51%	**+292%**
# External Commits	15	883	2098	**+13,887%**
# External LOC	17,153	470,946	1,119,240	**+6,425%**

Notable Collaborations

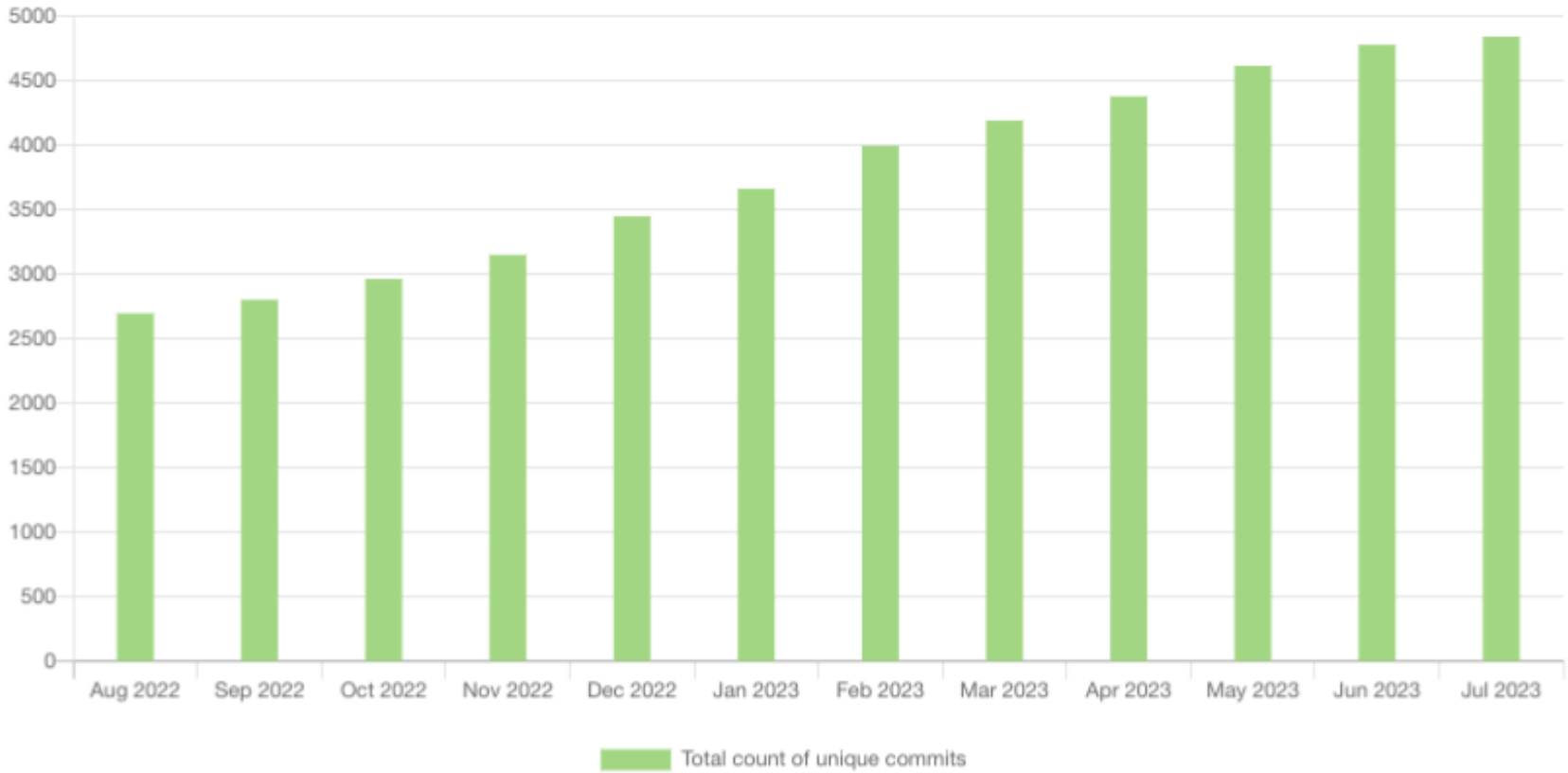
- Apache Flink agent for metadata emission (anonymous)
- DataGalaxy integration (DataGalaxy)
- Keboola metadata producer (Keboola)
- Atlan Airflow event consumer/platform (Atlan)
- Spark integration (Microsoft)
- Snowflake adapter (Snowflake)
- Egeria integration (Egeria)
- Amundsen table lineage extractor (Amundsen)
- Dagster integration (Northwestern Mutual)
- Manta OpenLineage scanner (Manta)
- Great Expectations integration (GX)
- Rust parser interface for SQL integration (supported by Microsoft)
- Interactive developer env, default extractor, & extractors for SQL check operators (Airflow)

Best Practices Badge

Open SSF Best Practices
Core Infrastructure Gold Badge
Earned July 2023

Commits Growth

The growth in terms of the aggregated count of total number of unique commits during the selected time period.



Key Insights

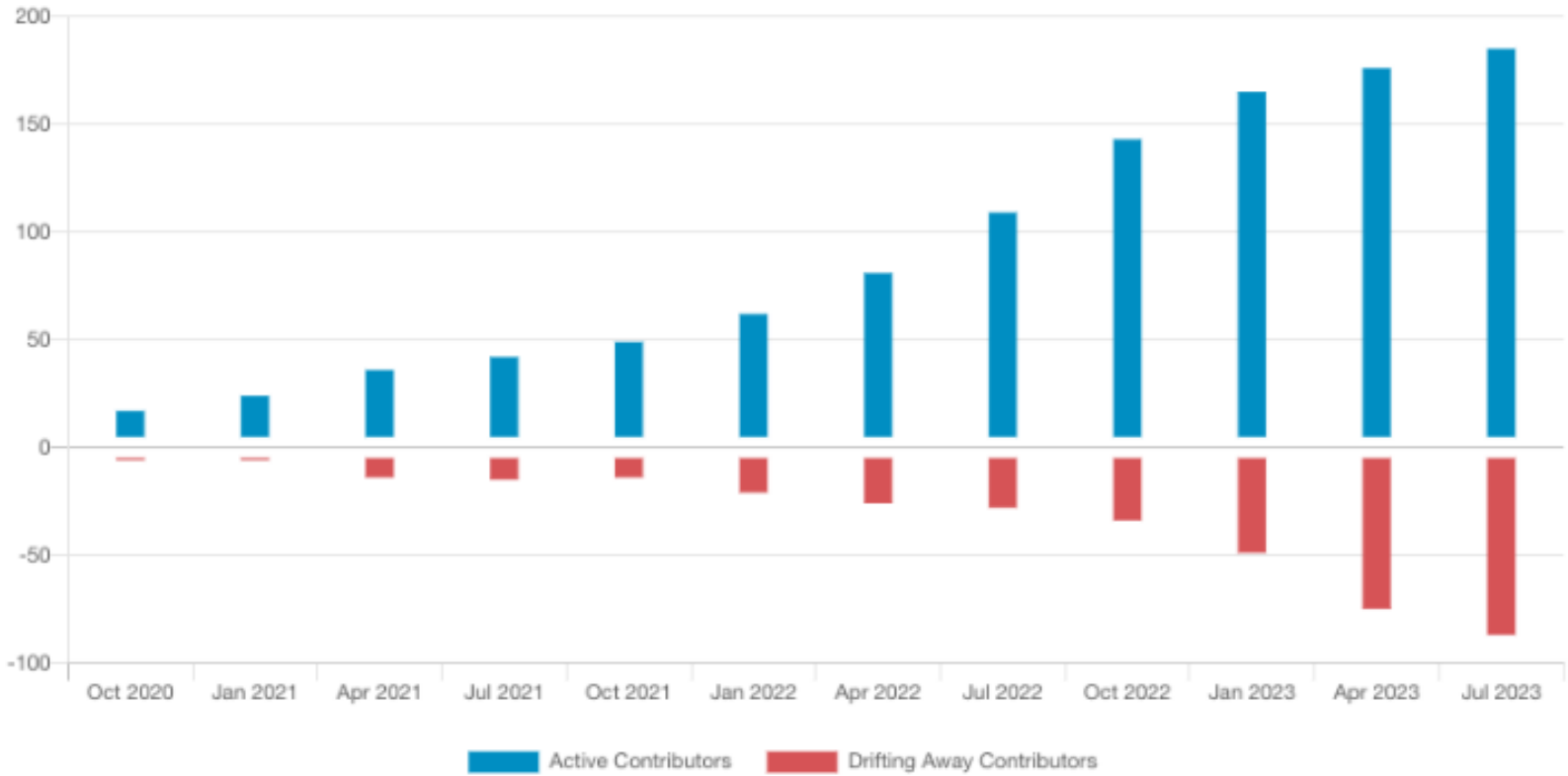
- There has been a **growth** of **79.59%** in the total commits during the last 1 Year.
- An average of **3.79K** commits were pushed by active code contributors during the last 1 Year.

Contributor Growth And Retention

The aggregated count of unique contributors that are active and drifting away for each time interval selected. A contributor performing code activity (Commits/PRs/Changesets) or submitting/resolving bugs within the last 1 year is marked "active" whereas a contributor not performing any code activity in the last 6 months is marked "drifting away".

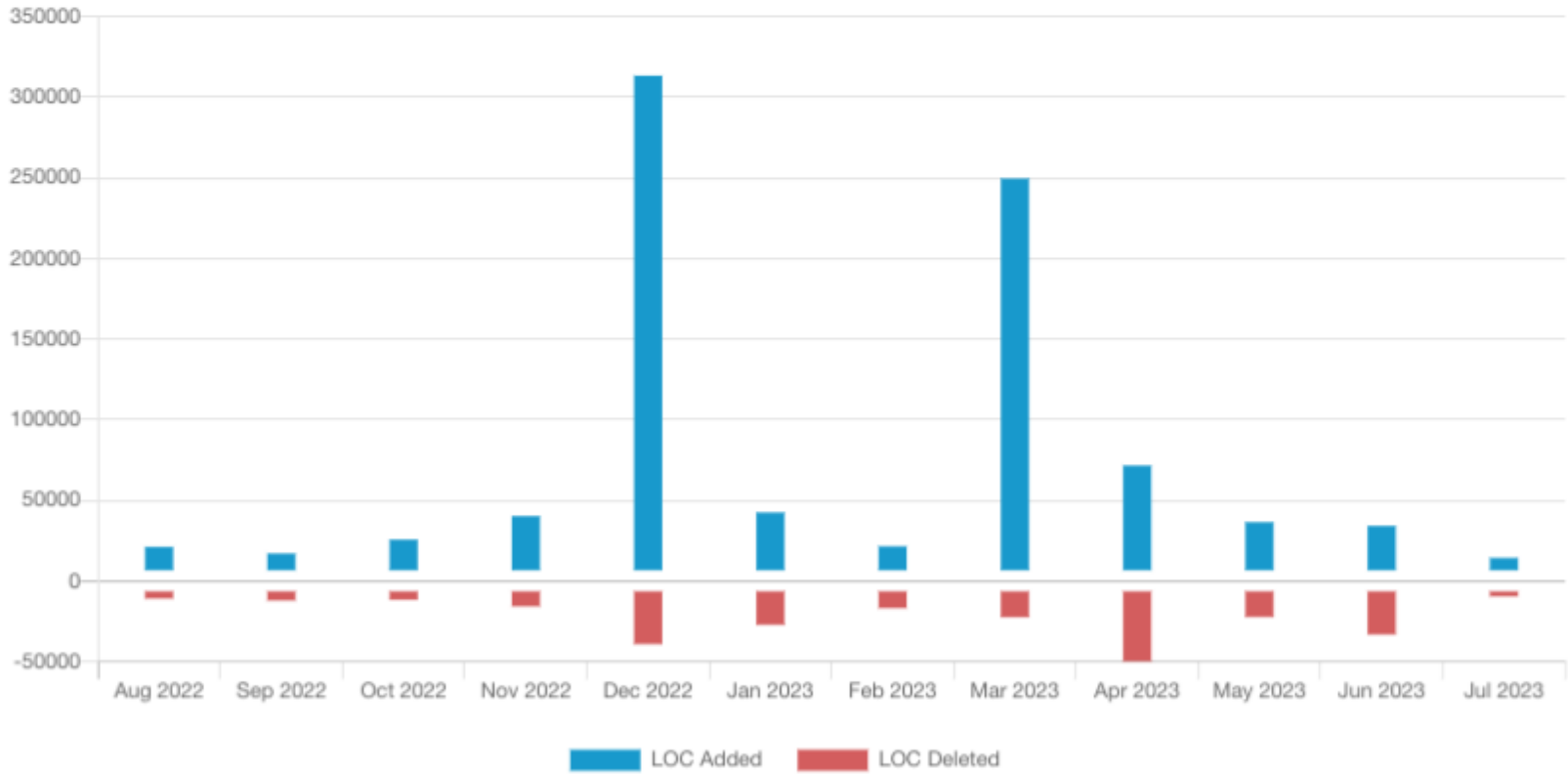
TOTAL CONTRIBUTORS	ACTIVE CONTRIBUTORS	DRIFTING AWAY CONTRIBUTORS	ACTIVE CONTRIBUTORS CHURN RATE
269 ↑1693.33%	180 ↑1400.00%	82 ↑5000.00%+	32.33%

The average count of active contributors was **86** during the last 3 Years.



LOC Added And Deleted

The count of the total lines of code added and deleted aggregated across every unique commit over the given time period.



Key Insights

- The average lines of code changed per commit **increased** by **145.41%**.
- The average LOC churn rate was **76** LOC/commit.
- Across **7** number of total repositories an average of **2.22K** loc were added to a repository on a weekly basis.
- An average of **67.26K** LOC were added during the last 1 Year.

Governance and Procedures

- **Governance document**
  https://github.com/OpenLineage/OpenLineage/blob/main/GOVERNANCE.md:
  - Release authorization process
  - Committer election process
  - Leadership structure
- **Technical Steering Committee document**
  https://github.com/OpenLineage/OpenLineage/blob/main/TECHNICAL_STEERING_COMMITTEE.md
- **Code of Conduct document**
  https://github.com/OpenLineage/OpenLineage/blob/main/CODE_OF_CONDUCT.md
- **Code Quality Assurance document**
  https://github.com/OpenLineage/OpenLineage/blob/main/CODE_QUALITY_AND_SECURITY.md
- **Releasing document containing release process**
  https://github.com/OpenLineage/OpenLineage/blob/main/RELEASING.md
- **New contributor guide**
  https://github.com/OpenLineage/OpenLineage/blob/main/CONTRIBUTING.md

Technical Steering Committee

Chairperson: **Julien Le Dem** (Astronomer, Marquez, Apache Arrow, Apache Parquet)

Voting members:

Mandy Chessell (Egeria, Pragmatic Data Research, Ltd.)	**Maciej Obuchowski** (GetInData, Marquez)	**Ross Turk** (Flox, Marquez)
Daniel Henneberger (DataSQRL)	**Paweł Leszczyński** (GetInData, Marquez)	**Benjamin Lampel** (prev Astronomer)
Drew Banin (dbt Labs)	**Will Johnson** (Microsoft)	**Kengo Seki** (NTT Data)
James Campbell (Superconductive)	**Michael Robinson** (Astronomer, Marquez)	**Minkyu Park** (prev Astronomer)
Ryan Blue (Apache Iceberg)	**Tomasz Nazarewicz** (GetInData, Marquez)	**Michael Collado** (Snowflake, Marquez)
Willy Lulciuc (Astronomer, Marquez)	**Jakub Dardziński** (GetInData, Marquez)	
Zhamak Dehghani (Stealth)	**Howard Yoo** (TechD, Marquez)	

LFAI Collaborations

- **OpenLineage + Egeria**
  - Completed direct integration
  - Egeria's integration daemon supports the OpenLineage API for local processing engines.
  - Egeria contributed the OpenLineage proxy backend in November 2021.
- **OpenLineage + Marquez**
  - Ongoing direct integration
  - Marquez is the reference implementation of the OpenLineage standard, acting as the backend to receive OpenLineage events from all the processing engines OpenLineage supports. Marquez also visualizes the lineage graph in map form and offers analytical data in its UI.
- **OpenLineage + Amundsen**
  - Completed integration
  - Amundsen's `OpenLineageTableLineageExtractor` extracts table lineage information from OpenLineage events.

Graduation Summary

- Joined as **Sandbox** project, May 2021
- Many unique contributors (**266**)
- Numerous involved partners: **Bloomberg, DataGalaxy, Collibra, Manta, Keboola, Matillion, Microsoft, Northwestern Mutual, Snowflake, Airflow, etc.**
- High GitHub star count (**1.3K**)
- Substantial flow of commits (**41** per week avg)
- Collaborations within the LFAI (**Egeria, Marquez, Amundsen**)
- TSC members from **Microsoft, Astronomer, dbt, Superconductive, Snowflake, etc.**
- Advanced to **Incubation** status, December 2022
- CII **Gold** badge

Please vote to approve OpenLineage for Graduation status with the LFAI.

2023 Roadmap

- Add Static Lineage support and release OpenLineage 1.0.0
- Contribute OpenLineage Provider to Airflow (AIP 53)
- Improve the SQL parser
- Add job and run facets to the Flink integration
- Add job/run dependencies facet to Airflow integration
- Add DAG run listener to Airflow integration
- Add ability to annotate datasets outside job runs
- Add facets validator to Python client
- Extend column lineage beyond Spark integration
- Automate updating of docs
- Add dbt Cloud support to Airflow integration
- Improve OpenAPI code generation
- Add support for Apache Atlas
- Add facet to report DAG metadata

Thank you!

Questions?

Approval on OpenLineage Graduation

Proposed Resolution:

- OpenLineage as a Graduation project of the LF AI & Data Foundation is hereby approved.

Upcoming TAC Meetings

LF AI & DATA

Upcoming TAC Meetings

- August 10 – DeepCausality proposal, LF Edge Presentation
- August 24 – New EthicalAI project from Fujitsu, Trusted AI Committee update

Please note we are always open to special topics as well.

If you have a topic idea or agenda item, please send agenda topic requests to tac-general@lists.lfaidata.foundation

Open Discussion

LF AI & DATA

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

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