LF AI & Data Foundation Technical Advisory Council (TAC) Meeting Minutes
May 4, 2023 (6:00am - 7:00am PST) via Zoom

TAC Meetings are bi-weekly and open to everyone in the LF AI & Data community.

Voting Notes
Under the original charter rules we need 50% attendance from the voting members (represented by premier members + graduate projects) to meet quorum. To ensure quorum is met more consistently, the TAC implemented the following attendance and voting eligibility rule effective Dec 1, 2022: Voting members of the TAC who miss 2 consecutive meetings will not be allowed to vote and must then attend 2 consecutive meetings before their voting privileges are reinstated.
With the new attendance and voting rule adopted by the TAC, the number of eligible voting members varies per week and is tracked here. Please see meeting minutes on the TAC Wiki for details on attendance and voting eligibility on a per meeting basis.

Please see the TAC Wiki for more information about voting.
Please see current voting members, including alternates, on the TAC Wiki.

Voting details for today's meeting: 10 voting members were eligible to vote and 7 voting members were in attendance. Quorum was met for this meeting.

TAC Voting Member	Voting Representative	3/23/2023
Quorum	**Attendance/No. Eligible Votes**	**0.7**
4paradigm	Zhongyi Tan	Jerry Tan
Baidu	Jun Zhang	
	Alternate: Daxiang Dong	
	Alternate: Yanjun Ma	
Ericsson	Rani Yadav-Ranjan	
Huawei	Howard (Huang Zhipeng)	Huang Zhipeng
	Alternate: Charlotte (Xiaoman Hu)	
	Alternate: Leon (Hui Wang)	
Nokia	Michael Rooke	Michael Rooke
	Alternate: Jonne Soininen	

OPPO	Jimmy (Hongmin Xu)	
SAS	*Nancy Rausch	Nancy Rausch
	Alternate: JP Trawinski	
ZTE	Wei Meng	Wei Meng
	Alternate: Liya Yuan	
Adversarial Robustness Toolbox Project	Beat Buesser	Beat Buesser
	Alternate: Kevin Eykholt	
Angel Project	Jun Yao	
Egeria Project	Mandy Chessell	
	Alternate: Nigel Jones	
	Alternate: David Radley	
	Alternate: Maryna Strelchuk	
	Alternate: Ljupcho Palashevski	
	Alternate: Chris Grote	
Flyte Project	Ketan Umare	
Horovod Project	Travis Addair	
Milvus Project	Xiaofan Luan	
	Alternate: Jun Gu	
ONNX Project	Alexandre Eichenberger	Alexandre Eichenberger
	Alternate: Jim Spohrer	
	Alternate: Prasanth Pulavarthi	
	Alternate: Andreas Fehlner	Andreas Fehlner
Pyro Project	Fritz Obermeyer	

Note for Voting Members

Please ensure you attend the bi-weekly TAC meetings to maintain voting eligibility. If you have not already provided an alternate representative, please email Nancy Rausch (Nancy.Rausch@sas.com), TAC Chair AND operations@lfaidata.foundation to designate an alternate representative. It is critical to meet quorum during the meetings especially when there are voting items on the agenda.

LF Attendees
Ibrahim Haddad, Executive Director, LF AI & Data
Lucy Hyde, Program Manager, Linux Foundation
Reden Martinez, Project Coordinator, Linux Foundation

Invited Guests/Presenters
Xu Chen, DMetaSoul
Yadong Zhu, DMetaSoul
Guangui Yang, DMetaSoul

Call to Order
Nancy Rausch (NR) called the meeting to order at 6:01am Pacific and Lucy Hyde (LH) recorded the minutes.

NR reviewed the antitrust policy notice.

Agenda
NR reviewed the agenda for the meeting. There were no further changes or additional topics added.

- › Roll Call (2 mins)
- › Approval of Minutes from previous meeting (2 mins)
- › Amundsen Project Update (20 minutes)
- › LakeSoul (20 Minutes)
- › Foundation Update (Ibrahim Haddad – 20 minutes)
- › Open Discussion

Approval of Minutes
NR presented a resolution to approve the minutes of the April 6, 2023 TAC meeting.

Proposed Resolution:

That the minutes of the April 6, 2023 meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.

Jerry Tan made the first motion to approve the minutes, and Michael Rooke seconded the motion.

APPROVED - By vote of the TAC, the minutes of the April 6, 2023 meeting of the Technical Advisory Council were approved.

Nancy Rausch asked members to please review the [Project Lifecycle Document](). Ibrahim Haddad (IH) stated that feedback was included from multiple members, and that the proposal to approve the document is on the TAC agenda in two weeks.

Annual Review of Amundsen

Kristen Armes (KA) presented an annual review for Amundsen. KA provided an overview of Amundsen's data discovery and search capabilities, and the additions to the homepage. KA stated that the databuilder ingestion framework ingests from multiple sources into Neo4j (a graph DB), and further pushes the data into Amundsen and Elastic.

KA highlighted contributing organizations to Amundsen, as well as contribution/contributor metrics from LFX Insights. KA highlighted key achievements: continued maintenance of Amundsen even with reduced headcount; 38+ companies are officially using Amundsen; a cohesive frontend strategy to set a unified direction for future contributors; the improved search service with fuzzy search, stemming, usage based rankings, multivalue filters; revamped Neo4j metadata publisher to increase speeds by 97%, and added support for newer Neo4j versions; Slack community reached over 3k members, and 3.9k stars on GitHub. KA stated that the project could use assistance with the onboarding process for users/contributors, front end migrations and other challenges, technical support to users, etc.

LakeSoul

Xu Chen (XC) presented an introduction to LakeSoul, a cloud-naive realtime lakehouse framework. XC gave an overview of reasons for donation to LF AI to include a vendor-neutral, not for profit environment; increase in users by outreach and an increase of contributors from developer users; collaboration with other projects in the Foundation; open governance and open source licensing; neutral management of project by the Foundation; and trust in contributors and adopters in management of the project.

XC stated that DMetaSoul is a startup company based in Beijing, focused on building better data and intelligence infrastructure and provided a background of the modern data stack, to include common characteristics such as cloud native infrastructure to promote scalability and elasticity, transition from ETL to ELT, storage format such as open data format and concurrent updates, and business intelligence and analytics.

XC stated that LakeSoul's goals focus on a cloud-first approach without dedicated storage; centralized metadata management; streaming data ingestion, increase efficiency in data analytics

and AI in the data lake. XC stated that LakeSoul does not focus on creating a new compute engine, file format, or optimizations. XC outlined each layer: the Meta Service (schema, ACID control, data distribution), the storage layer (cloud storage drivers and hot data caches), and the query engine (engine connector, and query optimizer). XC covered data modeling for physical data (files are stored physically in parquet format, where the table could optionally have a primary key constraint and multi-level range partitions) and metadata (commit where a file sequence with add/delete operations, commits sequence with commit types that constitutes a snapshot, and monotonic increasing number that identifies a snapshot and timestamp for versioning). XC stated that LakeSoul uses centralized metadata management through PostgreSQL as it is generally available through most cloud vendors, transaction support, granular write conflicts, advanced features such as trigger functions, and includes a Java wrapper and Spark/Flink's catalog interface implementations.

XC stated how LakeSoul utilizes a two-phase commit protocol (prepare and commit phase), fine-grained write conflict resolution, auto schema evolution, and snapshot read, rollback, and cleanup. XC discussed the native IO layer, which encapsulates read/write logics, provides easier integration, and is cloud native. XC stated that the native IO layer is implemented using an asynchronous reader/writer in Rust, with a C interface and Java/Python wrapper, and Spark DataSource v2 and Flink DynamicTableFactory implementations. XC discussed LakeSoul's streaming data ingestion, which supports synchronization of multiple tables from relational databases and multiple topics from message queues, and incremental streaming for data transformations. XC stated that LakeSoul also supports streaming joins, reducing maintenance overhead of a large stateful stream job, compute overhead of full joins, and achieves a higher throughput with lower latency.

XC provided application examples, such as an end-to-end real-time LakeHouse, with an incremental streaming pipeline without extra scheduling, unlimited storage where historical data can be accessed and updated, the ability to run BI/AL on the LakeHouse, and piping the data to external downstream systems. XC stated that LakeSoul can build Machine Learning Datasets in real time, for classification, forecasting, or recommendations. XC stated that LakeSoul can feed data directly into machine learning frameworks (MLLib, PyTorch, and Flink ML).

XC gave an overview of the current state, in that LakeSoul was open sourced in December 2021 under Apache 2.0 license. XC stated that LakeSoul currently has eleven contributors, four from other organizations, and early adoptions from aviation/banking/a research lab. XC stated that there was possible collaboration with OpenLineage and Marquez, Sparklyr and Feast, and building tabular training datasets for PyTorch, Angel ML, and FATE. XC elaborated that in the future, LakeHouse plans to improve data warehousing, echosystems, and performance.

Proposed Resolution:

That the LakeSoul project is approved by the Technical Advisory Council (TAC) as a Sandbox project of the LF AI & Data Foundation.

Jerry Tan made the first motion to approve, and Wei Meng seconded the motion.

APPROVED - By vote of the TAC, the DeepRec project was approved as an Incubation project.

Foundation Update
Ibrahim Haddad (IH), Executive Director of the LF AI & Data Foundation provided updates on the Foundation. Please see slide deck for details.

Open Discussion
There were no topics added.

Upcoming TAC Agendas
   › May 16 – Resilience Sandbox project, Update on the Project Lifecycle Document
   › June 1 – Project review <to be scheduled>

Please note the TAC is always open to agenda suggestions and guest presentations. If you have a topic you would like to request, please email tac-general@lists.lfaidata.foundation for review and coordination via the TAC Chair accordingly.

Closing
With no further business, the meeting was adjourned by NR at 6:53am Pacific.

Chat:
08:59:10 From Ibrahim Haddad to Everyone:
   Hi everyone 👋
08:59:22 From Jerry Tan-4Paradigm to Everyone:
   Hi, Ibrahim
08:59:34 From Ahmed Abdelmonsef (IBM) to Everyone:
   Hello everybody, GM, GA and GE 🙂
08:59:38 From Ibrahim Haddad to Everyone:
   Jerry, thanks for the reminder on WeChat ,
09:01:18 From LF AI & Data Zoom General to Everyone:
   Reacted to "Hello everybody, GM,..." with 👍
09:07:40 From LF AI & Data Zoom General to Everyone:

Here is the lifecycle document for review
https://docs.google.com/document/d/1VFoMcXlbOKa4cqTLkaWX-8_7-uh11hZT3b6p0wvjUHM/edit#heading=h.gjdgxs

09:15:54 From Ibrahim Haddad to Everyone:

Great progress on the growing the dev community 🎆

09:17:51 From Ibrahim Haddad to Everyone:

Are you able to identify any strong contributor who can be possibly promoted to a maintaier

09:59:46 From Junping Du to Everyone:

Congrats!

09:59:49 From Lucy Hyde | Program Manager to Everyone:

Congratulations, what an incredible project!

09:59:55 From Yadong Zhu to Everyone:

Many thanks

09:59:56 From Ahmed Abdelmonsef (IBM) to Everyone:

Congrats

10:00:07 From susan malaika (IBM) - non-voting to Everyone:

Congratulation!!!! LakeSoul

10:00:17 From Ibrahim Haddad to Everyone:

Congrats Xu and LakeSoul team. Great presentation and a very cool project.

10:00:20 From Ibrahim Haddad to Everyone:

Thanks everyone for attending

10:00:52 From Ali Hashmi to Everyone:

Thank you