

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

September 23, 2021

 LF AI & DATA

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
- › Approval of Minutes from previous meetings
- › Integration Update: Triton, Milvus, and Feast
- › NNStreamer Project Update
- › 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 Soinenen	jonne.soininen@nokia.com
OPPO	Jimin Jia*	jjajimin@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 August 26th, 2021 Minutes

Draft minutes from the August 26th TAC call were previously distributed to the TAC members via the mailing list

Proposed Resolution:

- › That the minutes of the August 26th meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.

Approval of September 9th, 2021 Minutes

Draft minutes from the September 9th TAC call were previously distributed to the TAC members via the mailing list

Proposed Resolution:

- › That the minutes of the September 9th meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.

Update on Integration: Triton, Milvus, Feast

September 23, 2021

Jun Gu

 THE **LINUX** FOUNDATION

 **LF** AI & DATA

NVIDIA Merlin addresses Recommender System challenges



NVIDIA Merlin

HugeCTR

Triton

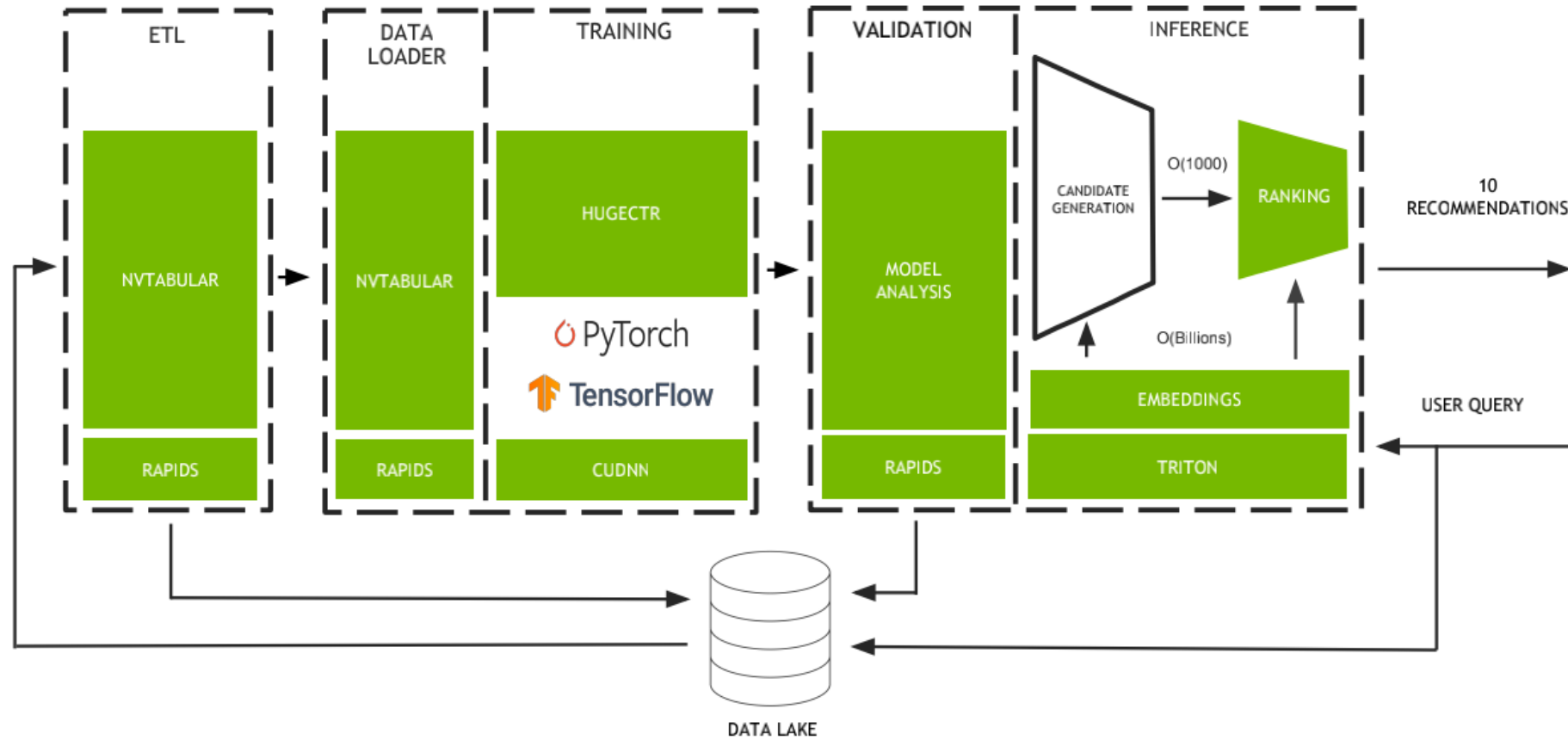
Challenge

Solution







	ETL	Data Loading	Training	Inference
Challenge	Pipelines are slow and complex	Using common item-by-item loading can be slow	Embedding tables of large deep learning recommender systems can exceed memory	High throughput to rank more items is difficult while maintaining low latency
Solution	GPU-accelerated and easy-to-use ETL pipelines prepares datasets in minutes	Asynchronous and GPU-accelerated dataloader for PyTorch and TensorFlow/Keras	Easy data and model parallel training allow to scale TB size embeddings	High throughput, low-latency production deployment

NVIDIA Merlin is an open-source library to deploy recommender systems end-2-end

NVIDIA Merlin Accelerates Every Stage in Recommender Pipeline



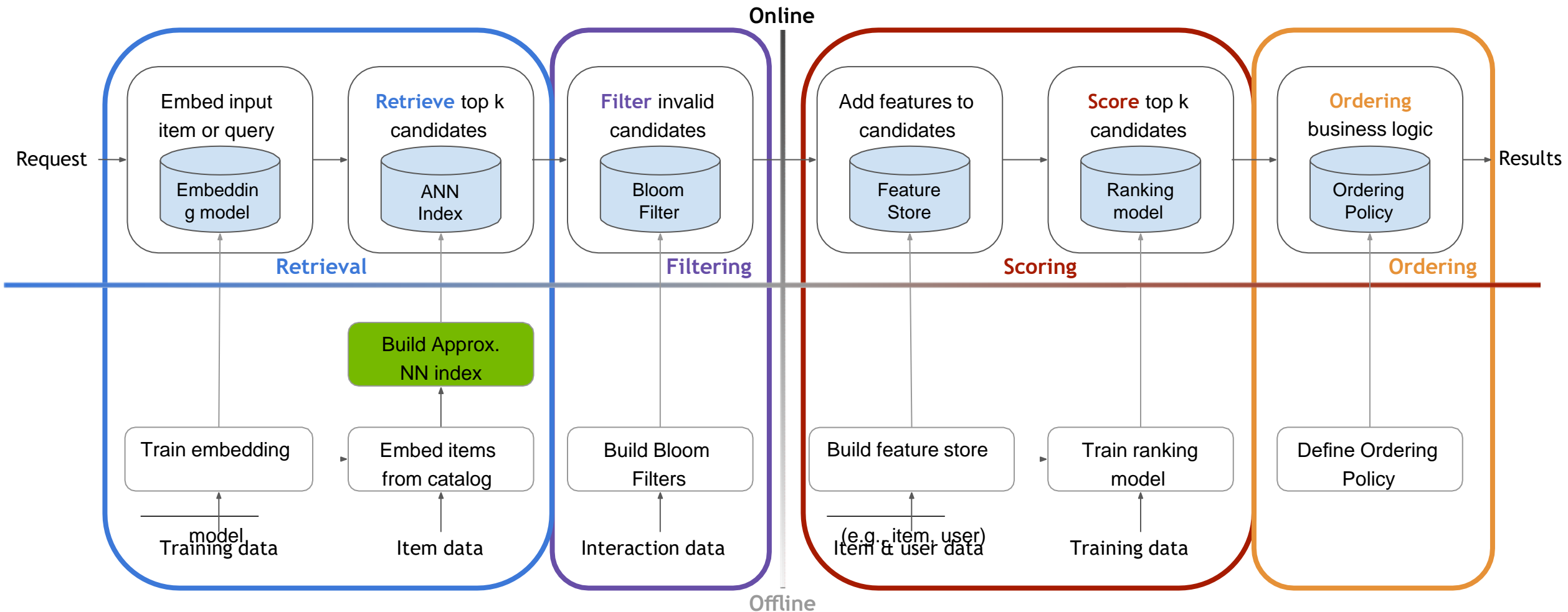
Recommenders On GPU Customer

Company	Use Case	Workflows	Outcome
	Accelerating real-time dynamic pricing for in-store grocery items to avoid waste	Merlin NVTabular, Dask and TF to scale data preprocessing Multi-GPU train a wide and deep model	10X ETL Operations Speedup With Merlin NVTabular
	Increased accuracy in advertising recommendations for 1B MAU	Integration of HugeCTR framework into advertising recommendation training framework	Training time reduced from 20 hrs to 3 7x speed-up in comparison to the previous TF solution using the same GPU infrastructure
	Recommenders curate real-time content feeds, encourage inspirational sharing	Using NVIDIA T4 GPUs, Triton and TensorRT	DL inference cost efficiency by 50% and decrease serving latency by over 60%
	Provide a delightful customer experience through highly relevant search results and recommendations	SotA CV uses Triton on V100 to process 300B pins(images) every night to extract embedded features	3x higher throughput with Triton dynamic batching
	Personalization and discovery features for three sided marketplace	RAPIDS.AI and Merlin NVTabular for TF Dataloader	Training time reduced from 1 hr (CPU) to 5 min (GPU) Reduced cost by 95% on NVIDIA A100 GPUs
	Improve playlist recommendations	TensorFlow Model Analysis (TFMA) - Improved with RAPIDS and NVTabular	Model validation time reduced from 3.5 hours (CPU) To 10 secs (GPU)

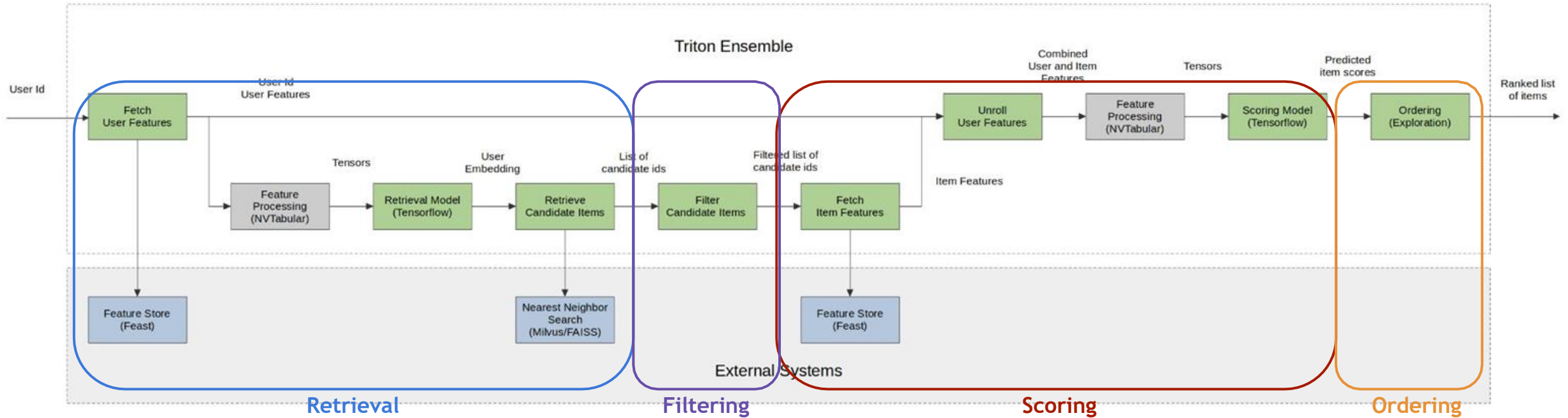
Four Stage Rec Sys Inference POC w/ Triton, Milvus and Feast

- Goal
 - The prototype leverages Triton model ensemble to perform four-stage inference for recommender model integrating Feast for feature retrieval and Milvus for candidate generation
- Functionality
 - The prototype takes a user id from the incoming request at inference and return an ordered list of recommendations
- Stages
 - Retrieval - nearest neighbor search w/ Milvus (Faiss-GPU)
 - Filtering - remove items from current session
 - Scoring - predict likelihood of item interaction
 - Ordering - softmax exploration
- External components
 - [Milvus](#) - open-source vector database for similarity search. Uses FAISS as as one of ANN algorithms (FAISS, Annoy, HNSWlib)
 - [Feast](#) - open-source feature store for model training and online inference

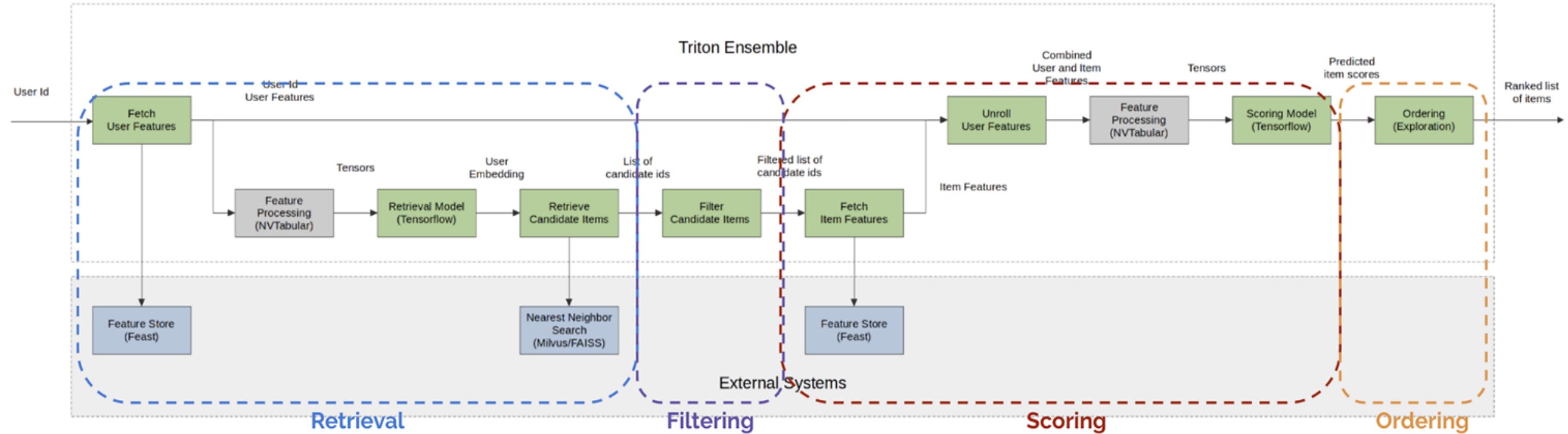
Four-stage Recommender Systems



POC diagram with Triton, Milvus and Feast



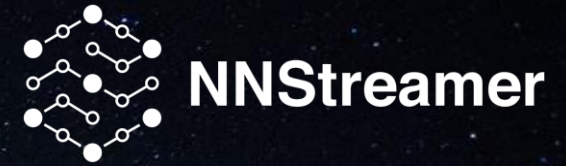
POC diagram with Triton, Milvus and Feast



Annual Review for NNStreamer

Myungjoo Ham (myungjoo.ham@gmail.com)
9/23/2021

NNStreamer



Brief Description:

NNStreamer is a set of Gstreamer plugins that allow Gstreamer developers to adopt neural network models easily and efficiently and neural network developers to manage neural network pipelines and their filters easily and efficiently.

Contributed by:

Samsung in March 2020 as an Incubation Project

Key Links:

Github: <https://github.com/nnstreamer/nnstreamer>

Artwork:


<https://github.com/lfai/artwork/tree/master/projects/nnstreamer>

Mailing lists:

- > [nnstreamer-announce](#)
- > [nnstreamer-technical-discuss](#)
- > [nnstreamer-tsc](#)

Incubation/Graduation Project review criteria

To be accepted into the Graduation stage, a project must meet the Incubation stage requirements plus:

- Need contributors from more organizations.
- Have reached [380 stars on GitHub](#).
- Core Infrastructure Initiative Best Practices: "Passing"  <https://bestpractices.coreinfrastructure.org/en/projects/4401>
- *A substantial ongoing flow of commits and merged contributions for the past 12 months**.
- Receive the affirmative vote of two-thirds of the TAC and the affirmative vote of the Governing Board.
- Have completed at least one collaboration with another LF AI & Data hosted project
- Have a technical lead appointed for representation of the project on the LF AI & Data Technical Advisory Council.

Organizations contributing

SAMSUNG

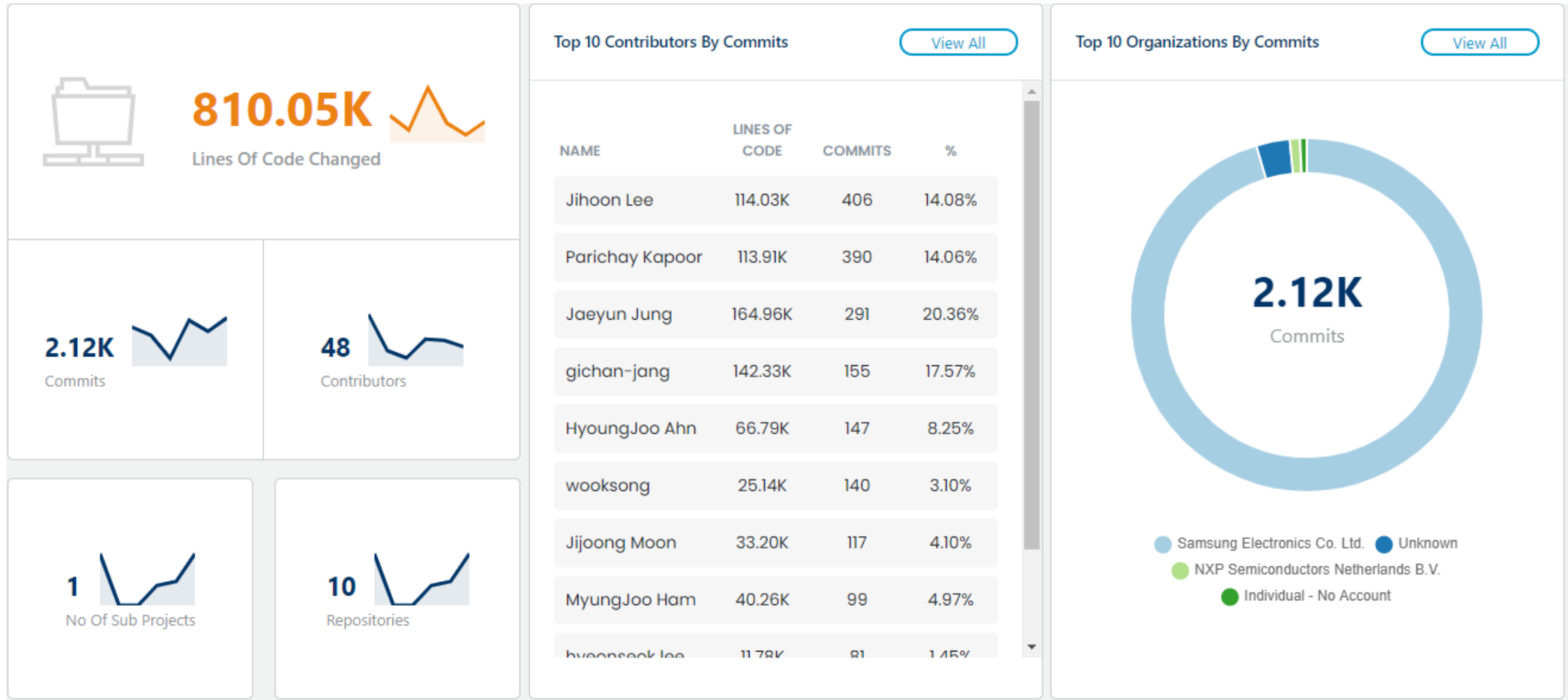


Hobbyist/Students

# Contributors	# Commits
> 20	> 2000
1	19
1	1
12 (A. Arthurs, J. Lee, C. Hall, W. Lee, D. Lee, C. Jeon, B. Kim, D. Kim, N. Jang, H. Park)	> 30

Counting nstreamer.git only (excluding other git repos) 2021-09-14

Contributions



2021-09-16, last 1 year

Key Achievements in the past year

Commercialization / Deployment

- Samsung
 - 2020: Galaxy Watch 3
 - 2021: Bixby Service in Galaxy Series
 - 2022 (WIP): TV, AR apps, Robotics, Home appliances
- Companies contacted for their products: NXP, Dell (Pravega), Collabora, Huawei, ODKMedia, FAInders,
- Tizen (5.5, 6.0, 6.5), Android JCenter, Yocto (meta-neural-network), Ubuntu (PPA), MacOS (Homebrew)

Features & Adaptation

- More stream data types: flexible-tensor, sparse-tensor, flexbuf
- More stream path manipulators: IF-branch, Join, Rate, Crop,
- Adaptation: TF-lite delegation, Tensor-RT (Nvidia), TVM, Lua scripts
- Gst-to-Pbtxt(mediapipe) converter. (for WYSIWYG pipeline editor in the future)


New Concept: Edge-AI (“Among-Device AI”)

- MQTT Pub/Sub, Query Server/Client (Prototype)

New Subproject: NNTrainer (On-Device Neural Network Training)

Papers: ICSE 2021 SEIP: “NNStreamer: ...” (nnstreamer) & “LightSys: ...” (nnstreamer’s subproject, “TAOS-CI”)

Areas the project could use help on

- Linux distro deployment
 - Deploy to Debian, OpenSuse, Fedora, ...
- Adding more automated test suites
 - E.g., Coverity, ARM-cloud for testing ARM binaries, ...
- Connection with Matter  matter
 - 2.0+ (2022) features are related w/ “Home IoT Connectivity”.
 - We’d like to contribute “Inter-device AI stream protocol”
 - Possibly, w/ plugins or libraries for MediaPipe, DeepStream

TAC Open Discussion

Upcoming TAC Meetings

Upcoming TAC Meetings (Tentative)

- › Oct 7, 2021: Ludwig Annual project review, Amundsen Annual project review; OC update
- › Oct 21, 2021: AI Fairness 360, AI Explainability 360, Adversarial Robustness Toolbox Annual project reviews

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

LF AI & Data - Ongoing Annual Project Reviews

Date	Project	Presenter
April 6, 2021	Egeria	Mandy Chessell (slack) - TAC recording / deck
April 6, 2021	OpenDS4all	Andre de Waal (slack) - TAC recording / deck
May 20, 2021	ONNX	Jim Spohrer (slack) - TAC recording / deck
July 15, 2021	EDL	Ti Zhou (slack) deck
July 29, 2021	Angel	Bruce Tao (slack) (confirmed) deck
July 29, 2021	Adlik	Meng Wei (slack) (confirmed) deck
Aug 12, 2021 (potentially Aug 12)	Sparklyr	Sigrid Keydana Yitao Li (slack) (confirmed) deck
Aug 12, 2021	Milvus	Jun Gu (slack) (confirmed)
Aug 26, 2021	Kendro new project into incubation	Yetunde Dada <yetunde_dada@mckinsey.com>
Sept 9, 2021	Marquez	Julien le Dem (slack) (confirmed)
Sept 9, 2021	Acumos	Amit Kumar (slack) (tentative)
Sept 23, 2021	NNStreamer	MyungJoo Ham (slack) (confirmed)
Sept 23, 2021	ForestFlow	Ahmad Alkilani (slack) (confirmed)
Oct 7, 2021	Ludwig	Piero Molino (slack) (confirmed)
Oct 7, 2021	Amundsen	Mark Grover (slack) (confirmed)
Oct 21, 2021	AI Fairness 360	Animesh Singh (to be asked)
Oct 21, 2021	AI Explainability 360	Animesh Singh (to be asked)
Oct 21, 2021	Adversarial Robustness Toolbox	Animesh Singh (to be asked)
Nov 4, 2021	Horovod	Travis Addair (to be asked)
Nov 4, 2021	FEAST	Willem Pienaar (to be asked)
Nov 18, 2021	SOAJS	Antoine Hage (to be asked)
Nov 18, 2021	Delta	Kun Han (to be asked)
Dec 2, 2021	DataPractices.org	Patrick McGarry (to be asked)
Dec 2, 2021	JanusGraph	Jason Plurad (to be asked)
Dec 16, 2021	Pyro	Fritz Obermeyer (to be asked)
Jan 6, 2021	Datashim	Yiannis Gkoufas (to be asked)
Jan 6, 2022	Flyte	Ketan Umare (to be asked)
Jan 20, 2022	RosaeNLG	Ludan Stoeckle (to be asked)
Jan 20, 2022	SubstraFramework	Camille Marini (to be asked)
	Machine Learning Exchange	Animesh Singh (to be asked)
	VulcanKompute	Alejandro Saucedo (to be asked)
	OpenLineage	Julien le Dem (to be asked)
	MARS	Chris Qin (to be asked)

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

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/j/430697670>

Open Discussion

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.