Building Efficient AI Pipelines with Self-Learning Data Foundation for AI

Annmary Justine
Hewlett Packard Labs
Hewlett Packard Enterprise
Fort Collins, USA
annmary.roy@hpe.com

Aalap Tripathy
Hewlett Packard Labs
Hewlett Packard Enterprise
Austin, USA
aalap.tripathy@hpe.com

Revathy Venkataramanan
Hewlett Packard Labs
Hewlett Packard Enterprise
Columbia, USA
revathy.venkataramanan@hpe.com

Sergey Serebryakov
Hewlett Packard Labs
Hewlett Packard Enterprise
San Jose, USA
sergey.serebryakov@hpe.com

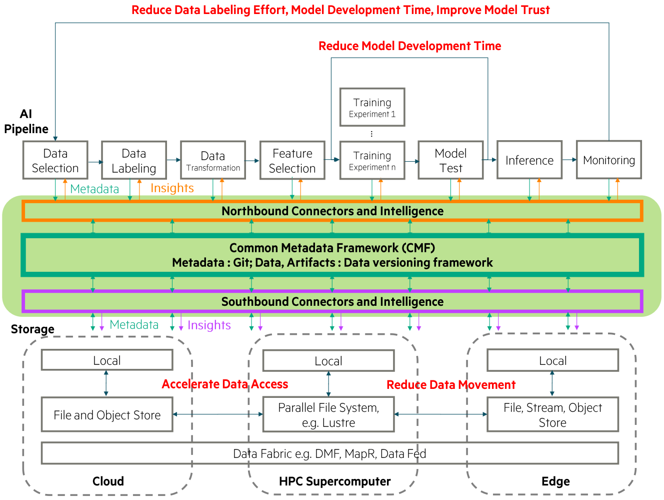
Martin Foltin
Hewlett Packard Labs
Hewlett Packard Enterprise
Fort Collins, USA
martin.foltin@hpe.com

Cong Xu
Hewlett Packard Labs
Hewlett Packard Enterprise
San Jose, USA
cong.xu@hpe.com

Suparna Bhattacharya
Hewlett Packard Labs
Hewlett Packard Enterprise
Bangalore, India
suparna.bhattacharya@hpe.com

Paolo Faraboschi
Hewlett Packard Labs
Hewlett Packard Enterprise
San Jose, USA
paolo.faraboschi@hpe.com

Development of Artificial Intelligence (AI) models is a multistage process that includes data collection, selection, labeling and augmentation; feature selection; model training, testing and refinement. These stages are organized as directed acyclic graphs called ML (machine learning) or AI pipelines and are parametrized by hyperparameters that have direct impact on end-to-end performance. AI model development is therefore a complex optimization problem in high-dimensional space of data characteristics and hyperparameters. This space is especially complex in "AI for Science" pipelines that often include multiple models applied in sequence, or models built incrementally as new data is collected. Tools available in the industry focus on hyperparameter optimizations for model training stages, however, they lack in visibility and optimization across all stages, with most notable deficiencies in data selection, pre-processing, and model retraining stages. To fill this gap, we have been developing a self-learning Data Foundation for AI that records lineage, hyperparameters and metrics of AI pipeline runs and learns from this metadata to optimize subsequent runs. We introduced the Data Foundation at 2022 SMC Conference [1] and describe here new capabilities.



Reduce Data Labeling Effort, Model Development Time, Improve Model Trust
Reduce Model Development Time
AI Pipeline
Data Selection
Data Labeling
Data Transformation
Feature Selection
Training Experiment 1
⋮
Training Experiment n
Model Test
Inference
Monitoring
Metadata
Insights
Northbound Connectors and Intelligence
Common Metadata Framework (CMF)
Metadata : Git; Data, Artifacts : Data versioning framework
Southbound Connectors and Intelligence
Storage
Metadata
Insights
Local
Local
Local
Accelerate Data Access
Reduce Data Movement
File and Object Store
Parallel File System, e.g. Lustre
File, Stream, Object Store
Data Fabric e.g. DMF, MapR, Data Fed
Cloud
HPC Supercomputer
Edge

Figure 1: Self-Learning Data Foundation for AI (highlighted green) in AI software stack

First, we present the Federated Common Metadata Framework, a core component that enables metadata management in distributed AI pipelines. Second, we show examples of Data Foundation intelligence that i) recommends AI model and initial hyperparameter seeds for a given task to reduce AI model training time; and ii) assists with energy and carbon footprint tracking and optimization for AI pipelines. We conclude with the discussion of future work.

A. *Federated Common Metadata Framework*

Complex AI pipelines often run in distributed environments across different sites spanning HPC computing facility (e.g., for model training, retraining, coupling to HPC simulations, etc.), edge (e.g., AI inference, monitoring, active learning, etc.) and cloud (e.g., cloud bursting). Multiple teams may collaborate on model development, each responsible for different stages or covering different subspaces of hyperparameter optimization space. We developed Common Metadata Framework (CMF) to enable management of artifacts (intermediate data, models) and metadata alongside pipeline code with Git-like simplicity [1]. CMF has been extended here to support distributed pipelines (see Figure 2). It enables tracking and storing metadata and data locally at each site, and distribution of metadata subsets that are merged with metadata from other sites to provide lineage and provenance tracking. It decouples the data management from the metadata management, enabling to share only the required data when needed, reducing data movement and ensuring data privacy. This is enabled by i) hierarchical organization of pipeline metadata facilitating insertion to the appropriate branch in the distributed pipeline lineage tree, ii) a mechanism to index artifacts allowing to merge pipeline lineage trees and metadata from independently executed steps based on input/output artifacts, iii) independent management of metadata and data allowing to keep the data local while sharing metadata, and, iv) peer-to-peer model to facilitate merging of metadata and lineage at any site without need for central coordinator. In the paper we will discuss how Federated CMF enables pipeline reproducibility and incremental improvement of results after fine-tuning AI models at the edge. We use a high energy physics particle trajectory reconstruction pipeline example that

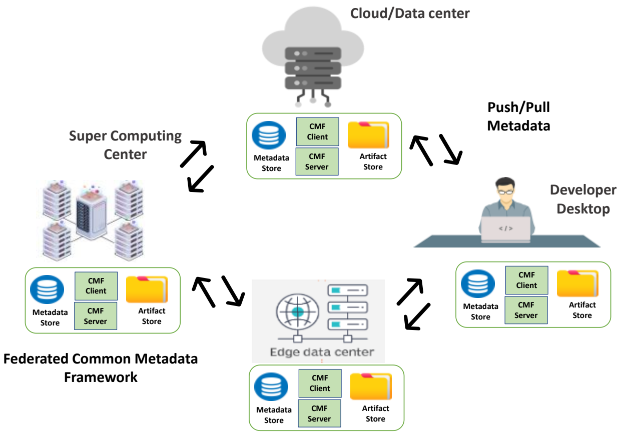


Figure 2: Federated Common Metadata Framework

involves progressive filtering of points on a common trajectory by multiple AI models.

B. AI model and hyperparameter recommendation

As we mentioned previously, the process of developing AI models involves the search of pipeline stages' hyperparameters (e.g., selecting model architectures, strategies for data selection and pre-processing, optimization algorithms). This requires significant human effort, compute resources, and energy. In previous work we demonstrated how Data Foundation for AI learns from previous runs of a pipeline to optimize the pipeline and reduce this effort [1]. In this work we extend the Data Foundation intelligence to utilize a-priori knowledge and metadata captured from executions of hundreds of different pipelines to recommend a small set of models and hyperparameters best fitting for a given task and dataset. These recommendations can be used as a seed (good known configurations) for AutoML methods such as neural architecture search (NAS) and Bayesian hyperparameter search to accelerate model development. Central to our architecture (see Figure 3) is an evaluator that transforms the universe of ML pipeline metadata into a knowledge graph. Unique tasks represented by nodes are connected by edges if and only if they meet a similarity threshold employing carefully designed similarity metric based on task categories, modalities, and dataset characteristics. We will demonstrate the benefits by considering examples in two domains: AIOps (anomaly detection for time-series data) and Telco (churn prediction). For the latter, we achieved 1.5 to 12x speedup in model development for different use cases as compared with baseline (Bayesian search without seed configurations).

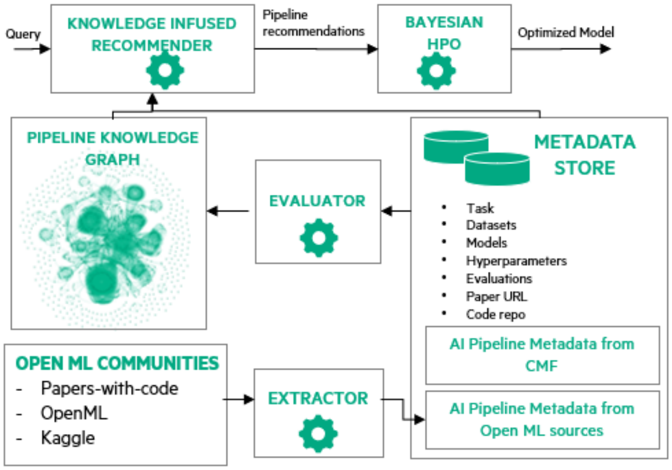


Figure 3: Knowledge-infused AI model and hyperparameter recommender system

C. AI pipeline energy & carbon footprint analysis and optimization

Accurate reporting of carbon footprint in distributed AI pipelines enables researchers to implement targeted energy-efficient optimizations. Carbon cost should be computed at the time of the experiment rather post fitting, as it is found that estimating carbon footprint post executions results in inaccurate results [2] Federated CMF tackles this problem by measuring the system metrics automatically at the time of execution and taking into consideration PUE of the data center and renewable energy in the grid, to provide the cost of execution automatically at the end of the experiment cycle. Further, since Federated CMF can track various stages in a pipeline across different geographies and teams, it can track the cumulative carbon footprint of a pipeline from the pre-processing stage to the inference and retraining stages. This end-to-end observability enables evaluation of various trade-offs like energy efficiency at training vs energy efficiency at inference and cost of retraining. The long-term vision into various trade-offs enables Federated CMF to accelerate research in optimizing pipelines for energy efficiency.

D. Future Work

We will discuss future research directions towards reducing model development effort and increasing model trust by intelligent data selection and metalearning from historical experience.

REFERENCES

[1] A. Justine, et. al., *Self-Learning Data Foundation for Scientific AI*, SMC 2022, CCIS volume 1690, Springer 2023, https://link.springer.com/chapter/10.1007/978-3-031-23606-8_2

[2] D. Patterson, et. al., *Carbon Emissions and Large Neural Network Training*. ArXiv. https://doi.org/10.48550/arXiv.2104.10350