Machine Learning Hub for Tapis

Simplifying MLOps for Research

Dhanny Indrakusuma
Nathan Freeman
Joe Stubbs

TACC TEXAS
The University of Texas at Austin

Overview

Machine Learning Hub (ML Hub) is a dynamic platform for the Tapis Framework, designed to streamline the machine learning workflow for researchers utilizing TACC's HPC systems.

Through our close collaboration with the ICICLE AI Institute, we identified the various challenges individuals faced when attempting to integrate machine learning workflows into the Tapis framework.

ML Hub offers user-friendly access for researchers, leveraging Hugging Face Hub API to provide open-source pre-trained models, and Tapis API for user authentication. Using ML Hub, users can browse and download models, execute inferences, as well as train and fine-tune models on TACC's HPC cluster.

Building upon NSF investments into the Tapis project, ML Hub aims to democratize machine learning, simplifying complex machine learning tasks by leveraging TACC's HPC resources and promoting inclusive research and innovation.

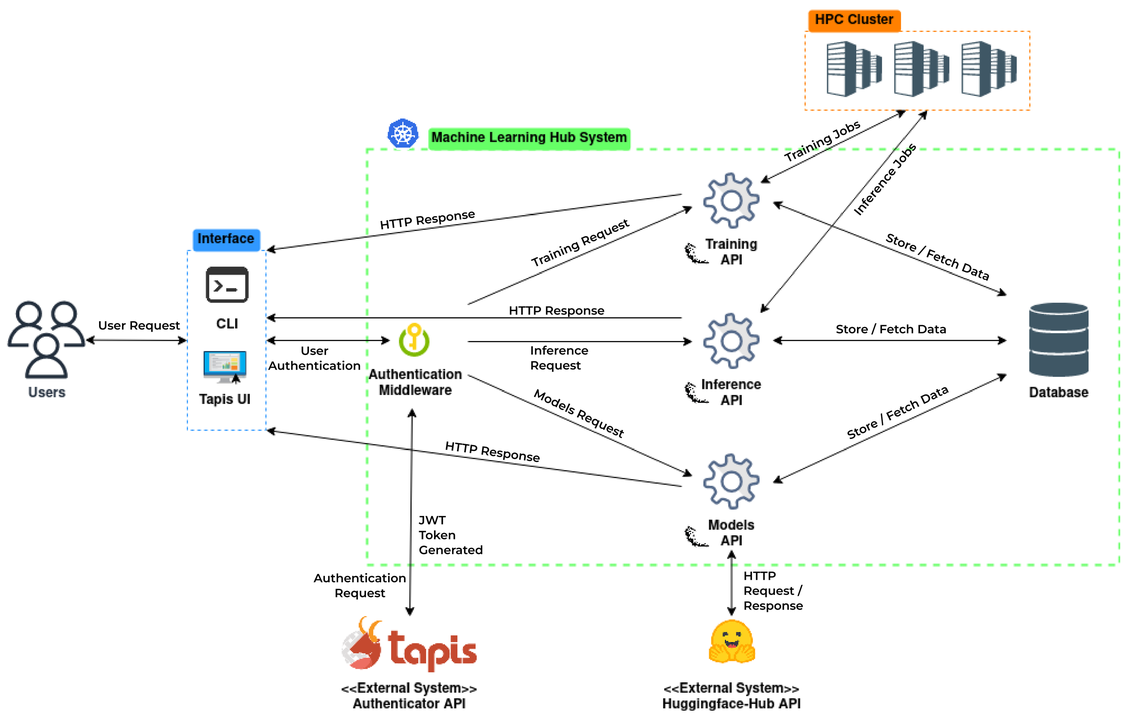
ML Hub Architecture

ML Hub is designed as a set of user-friendly microservices, with each service providing an independent REST API over HTTP.

The portal is codified using OpenAPI v3 definitions and implemented in Python's Flask and FastAPI. By integrating Hugging Face Hub API into ML Hub, we aim to provide open-source pre-trained machine learning models, allowing researchers and developers to harness state-of-the-art AI capabilities.

Currently, the Models Overview and Models Download functions in ML Hub provide a single point of access for non- technical users to explore and download available machine learning models. All service requests are authenticated using a JSON Web Token (JWT) generated from the Tapis API.

Future efforts involve implementing the Inference Client and Training Engine, as well as ensuring seamless integration into the Tapis UI.



Key Functionalities

Models Overview: The portal showcases top Hugging Face models, allowing users to filter by author and access detailed information of a specific model.

Models Download: Users can obtain specific models, with options to either download a binary file of the model or a zip file containing the model's repository, cached in a version-aware manner.

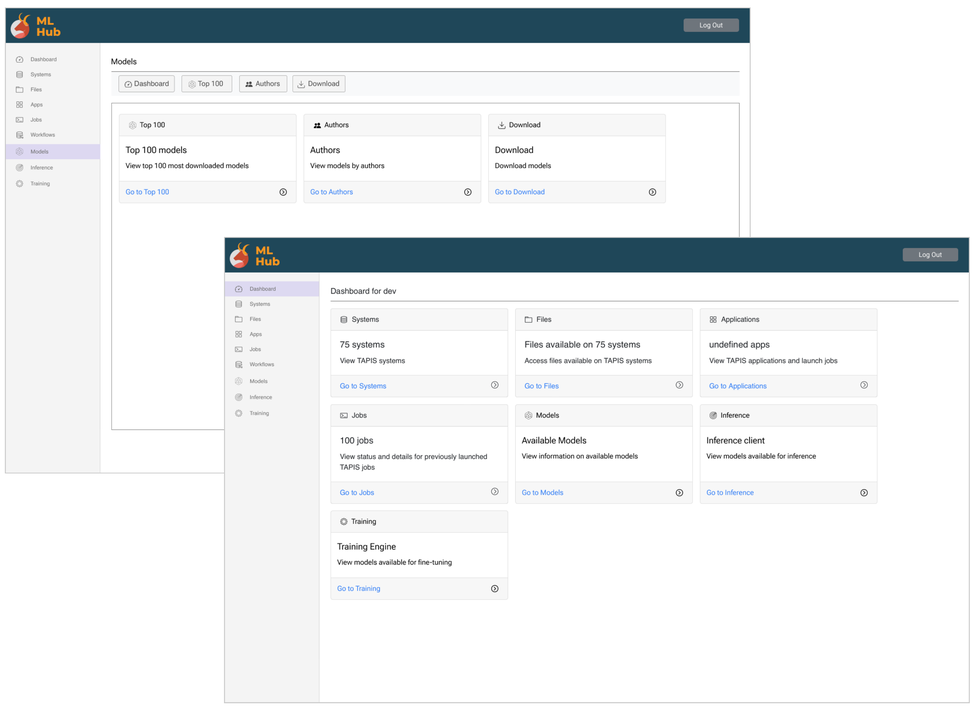
Inference Client: Facilitating server initiation for machine learning model inference on TACC's HPC cluster, enabling rapid prototyping.

Training Engine: Enabling users to train, fine-tune models and showcase them on TACC's HPC cluster, eliminating technical complexities and bottlenecks.

Tapis UI Integration

Machine Learning Hub's web interface will be built on top of the Tapis UI, a serverless gateway implemented in React and Typescript.

The user-friendly interface allows users to access files, applications, and jobs within their Tapis account. Futhermore, users can browse and download available machine learning models, run inferences and submit model training jobs to HPC cluster hosted by TACC.



Learn More

For more information and demo, scan the QR code:

or visit:

https://dhannywi.github.io/ml-hub

Acknowledgement: Machine Learning Hub is part of the Tapis project, supported by the National Science Foundation Division of Advanced CyberInfrastructure, award numbers #1931439 and #1931575.