IRIS-HEP Fellowship Proposal

Ajay R. Rawat

University of Washington, Seattle

Duration : June 2021 to August 2021
WBS : Analysis Systems (AS)
Project : Integrating REANA Backend into ROB for evaluating workflows in the cloud

Funding Period: 3 months of FTE during Summer 2021

The ***R**eproducible **O**pen **B**enchmarks for Data Analysis Platform (ROB)*[1] is designed for analyzing, benchmarking, and ranking data analysis workflows in a controlled competition-style format. ROB's prime use case is to benchmark and compare different data analysis workflows on the same data set, for example: benchmarking various top taggers. On the other hand, the ***Re**producible research data **ana**lysis platform (REANA)*[2] is a cloud platform for running containerized data analysis pipelines.

ROB is designed to be largely independent of the way workflows are specified and the backend that is used to execute them. Workflow steps in ROB originally were executed within Docker containers that are provided by the user. Currently, ROB runs on the user/coordinator's local machine along with all of the workflows using flowServ[3] as the workflow engine. Aaron Wang[4] added the option for users to submit Jupyter notebooks as workflow steps that will be executed using papermill (still on the user machine). He also included test cases to illustrate how to use ROB to evaluate Jupyter Notebooks. The main goal for this project is to extend the list of backends that can be used to execute the workflows. This is particularly important for computation-intensive workflows that require high-performance computing clusters. The primary goal of this project is to allow the ROB to use REANA's workflow engine to evaluate the workflow/notebooks and obtain the benchmarks. REANA has a Python API that can be used to upload the files, run the workflow and download the results. This API can be integrated within ROB to allow users to run their workflow in the REANA's cloud platform. The major advantage of using REANA is the ability to execute on a powerful compute cluster, e.g. on a High-Performance Computing environment (HPC), and the ability to run complex workflows using the Yadage[5] workflow system. ROB would also be able to run parameterized Jupyter Notebooks as workflow steps on the REANA cloud platform with the help of papermill. If time permits, another goal for this project is to use other cloud platforms like Google Cloud Platform, Azure, etc. to run workflows. This could be

achieved by translating an instance of a ROB workflow into a sequence of commands that run Docker containers in the cloud and collect the final results.

Under the supervision of Professor Shih-Chieh Hsu at the University of Washington and the technical guidance provided by Heiko Müller at New York University, Ajay Rawat will develop the backend to allow ROB to execute workflows on the REANA cloud platform. He will build upon Aaron's work to execute Jupyter Notebooks on REANA using papermill. Ajay's work will be available on the public repositories of ROB.

Schedule:

- **Month 1 (REANA Serial):**
  - Week 1:
    - Familiarize with the flowServ (ROB back-end) and the REANA API service
  - Week 2 to 4:
    - Integrate the REANA Python API into ROB
    - Test the code with the example workflows available
    - Test using a complicated workflow like TreeNiN[6]
- **Month 2 (Jupyter Notebooks):**
  - Week 1 and 2:
    - Create a Schema for Notebook evaluation that specifies the format of submitted notebooks
    - Update flowServ to use papermill to evaluate notebooks on REANA cloud
  - Week 3 and 4:
    - Test with MNIST and other example notebooks
    - Run a complete Benchmark test on ROB
- **Month 3 (Yadage + other cloud services):**
  - Week 1 to 3:
    - Integrate Yadage workflow in flowServ
    - Run final tests for the projects
    - Experiment with Google Cloud/Azure/AWS for running workflows
  - Week 4:
    - Finish the Documentation for the project and specification for the code
    - Upload the final code along with the final documentation

References:

- [1] ROB: Müller, Heiko and Macaluso, Sebastian: ROB: Reproducible Open Benchmarks for Data Analysis Platform, https://indico.cern.ch/event/809820/contributions/3632640/
- [2] REANA: reanahub.io
- [3] flowServ: Müller, Heiko: https://github.com/scailfin/flowserv-core
- [4] Aaron Wang: IRIS-HEP fellow
- [5] Yadage: Heinrich, Lukas https://yadage.readthedocs.io/en/latest/
- [6] TreeNiN: Macaluso, Sebastian https://github.com/diana-hep/TreeNiN