The Workflow Management System for Data Processing towards Photon Sources

Hao-Kai SUN
on behalf of IHEP-CC & HEPS-CC

Abstract

The new-generation light sources, such as the High Energy Photon Source (HEPS) under construction, are one of the advanced experimental platforms that facilitate breakthroughs in fundamental scientific research.

These large scientific installations are characterized by numerous experimental beam lines (more than 90 at HEPS), rich research areas, and complex experimental analysis methods, leading to many data processing challenges: high-throughput multi-modal data, flexible and diverse scientific methodology, and highly differentiated experimental analytical processes.

This project will use the idea of "workflow" to independently design and implement a set of graphical general-purpose management systems to solve the following key problems:
(1) how to quickly share and apply data processing methods to experiments by beamline scientists, experiment users, and methodology developers during analysis;
(2) how researchers can flexibly customize and monitor complex and diverse data processing processes;
(3) how the whole process of experimental analysis can be applied in batches to similar experiments and the results can be reproduced.

Introduction

In the framework of data processing and analysis software, algorithms or computational units, typically implemented by scientists, constitute the smallest units. Workflows delineate the invocation relationships and execution logic among algorithms and may serve as algorithms themselves invoked by other workflows.
As illustrated in [**Fig.1**], Daisy's compute engine manages the initialization, resource allocation, and execution of algorithmic entities, while the data repository oversees the creation and propagation of data objects among algorithms. We offer a user-friendly, flexible, and intelligent graphical interface for managing, orchestrating, and monitoring the entire data analysis workflow, facilitating complex workflow orchestration such as algorithmic parallelism, multiple input-output configurations, conditional execution, and nested workflows. This enables the rapid design of end-to-end experimental data analysis workflows.

Fig.3(a). List of Autogenerated Nodes for Integrated Algorithms.

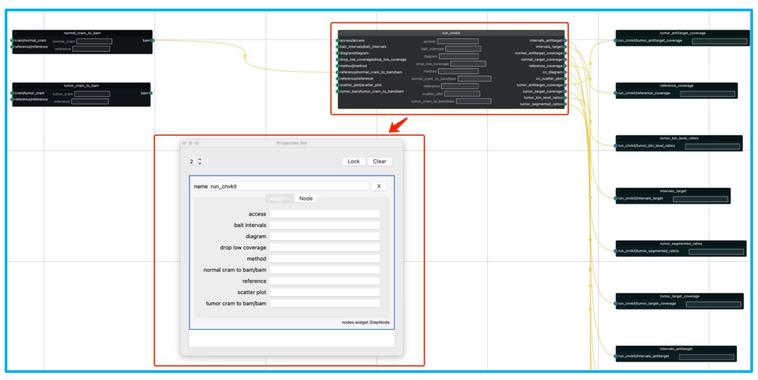

Fig.3(b). Parameter Configuration Interface for Compute(Algorithm) Nodes.

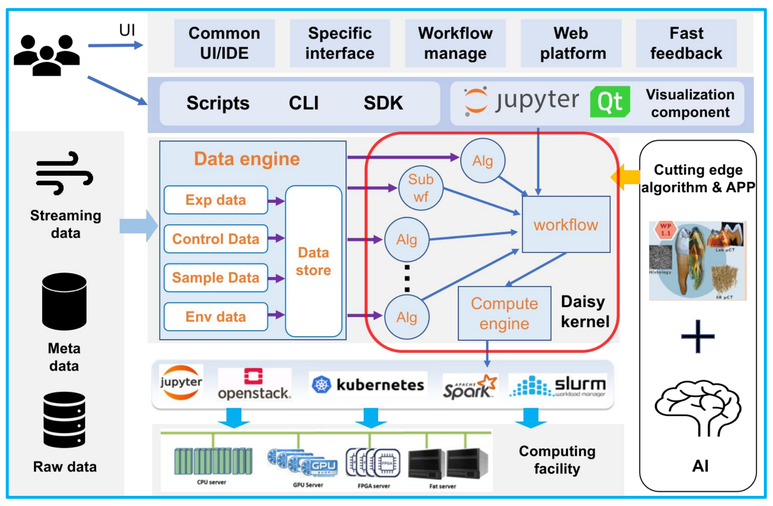

Fig.1 The Daisy Software Framework and where workflow system locates.

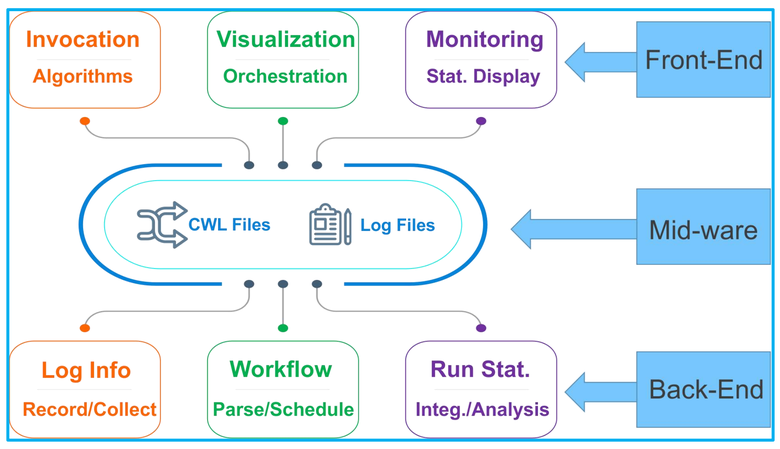

Fig.2 The Architecture of Daisy workflow management system.

Technical Roadmap

The system architecture [**Fig.2**] is divided into three levels: the front-end includes a visual orchestration and monitoring interface; the mid-ware layer comprises a general workflow description language and runtime logs; and the back-end involves workflow parsing, scheduling, dispatching, logging, and monitoring systems.

(a) The front-end draws inspiration from existing software in various fields, as well as the IT industries, independently developing a set of features to support:
- Automatic generation of computational nodes of methodological software for existing algorithms.
- Drag-and-drop arrangement and configuration.
- Workflow saving, loading, and modification.
- A graphical user interface with dynamic monitoring and automatic statistical analysis.

(b) The mid-ware layer operates as an independent intermediary, decoupling the graphical front-end from the back-end. Common Workflow Language (CWL) is adopted as the general workflow description language.
(c) The back-end is responsible for specific execution logic processing, comprising:
- CWL file parser.
- Communicate with the compute engine.
- Analysing log info and streamlined and provided to the core layer of Daisy and various subsystems after minimal wrapping.

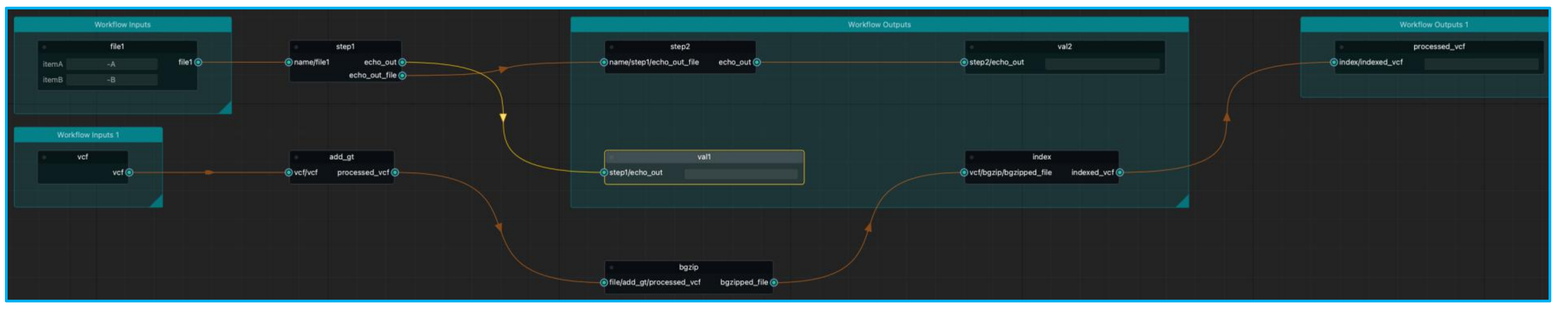

Fig.4 Complex Workflow Orchestration and Management with Nesting and Branching Support; Additionally, in Dark styling theme.

Project Progress

The current focus of the project development lies predominantly on the front-end and mid-ware layer. Leveraging Python packages and Qt technologies, the main accomplishments include:

(a) The graphical interface for algorithm node [**Fig.3(a)**] generation within Daisy, which facilitates flexible configuration of inputs, outputs, and parameters [**Fig.3(b)**].
(b) Visual workflow orchestration allowing mouse-driven interconnection of nodes, definition of execution sequences, and support for complex workflows such as branching and nesting [**Fig.4**].
(c) Importing[**Fig.5(a)**] and exporting[**Fig.5(b)**] of workflow description files in the CWL format, enabling experimental users to save, load, modify, and share their entire data processing workflows.

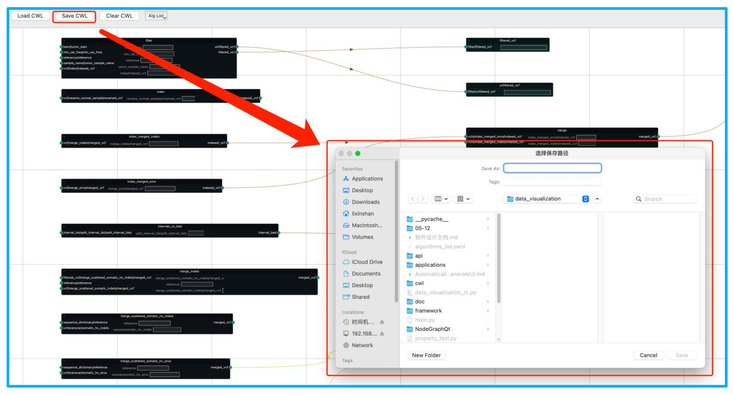

Fig.5(a). The arranged workflow can be exported as a CWL file.

Future Planning

According to the software development plan on HEPS, this project will steadily progress, focusing on the following aspects:

(a) Enhancing the details of the visual orchestration interface in the front-end, iteratively through discussions with beamline staff and users.
(b) Implementing the backend portion in Daisy-Core, initially developing the monitoring module rapidly with callback functions to facilitate testing with other layers; subsequently transitioning to a log + message pattern.
(c) Implementing interface designs to integrate with Daisy's computation engine, including parsing and converting CWL into Daisy algorithmic workflows.
(d) Following a modular design approach, the front end's visual orchestration and monitoring system will be decoupled from Daisy, facilitating its extension to other large-scale scientific facilities and experimental platforms.

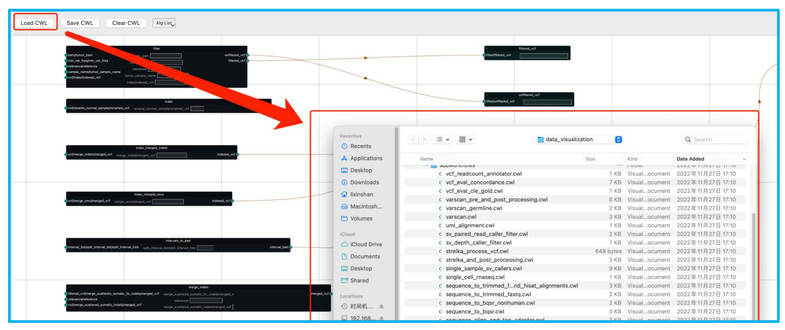

Fig.5(b). Existing or Exported CWL Files Importable as Visual Workflows

Contact:

Hao-Kai SUN
Computing Center, IHEP, CAS
Email: sunhk@ihep.ac.cn
Phone: +86-18602612453

References:

1. Qi F. et al. The Design of Science Data Platform for High Energy Photon Source[J]. Frontiers of Data and Computing, 2020, 2(2):
2. Dong, Y. et al. Exascale image processing for next-generation beamlines in advanced light sources. Nat Rev Phys 4, 427–428 (2022).
3. Arnold, O. et al. Mantid-Data Analysis and Visualization Package for Neutron Scattering and mu-SR Experiments. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 764 (2014): 156-166
4. Hu, Y. et al. Daisy: Data analysis integrated software system for X-ray experiments. EPJ Web of Conferences 251, 04020 (2021)
5. W. De Nolf, H. Payno, O. Svensson, & G. Koumoutsos. (2022). ewoks (0.0.5a). Zenodo. https://doi.org/10.5281/zenodo.6075054
6. Ahmed, A.E. et al. Design considerations for workflow management systems use in production genomics research and the clinic. Sci Rep 11, 21680 (2021). https://doi.org/10.1038/s41598-021-99288-8
7. R. Mitchell et al., "Exploration of Workflow Management Systems Emerging Features from Users Perspectives," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 4537-4544, doi: 10.1109/BigData47090.2019.9005494.