Extensible Scientific Workflow Engine and Ecosystem for Genome Analysis Workflows

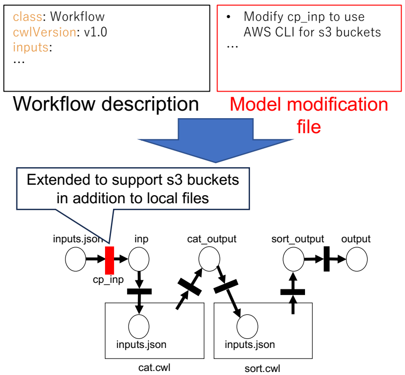
Due to the diversity of platforms, such as HPC and cloud platforms, and the variety of required features for the workflow engines, it is hard to support all of them in a workflow engine. We propose ep3, a workflow engine that can extend its supported platforms and features by modifying its internal state transition model rather than modifying its implementation directly. We have also developed an ecosystem for genome analysis workflows to utilize workflow execution records.

Research Goal

- ✓ Make it easier for system administrators to provide platform-specific optimizations and features without modifying workflow descriptions or workflow engines directly
- ✓ Utilize existing workflow-related technologies, e.g., Common Workflow Language (CWL), Docker, and Singularity
- ✓ Execute genome analysis workflows that are published in workflow registries, e.g., Common Workflow Library, in appropriate computing resources
- ✓ Make it easier to analyze workflow execution records to investigate characteristics of workflows

ep3: A workflow engine for CWL that aims to have a pluggable architecture

- ✓ By modifying its internal state transition model, we can support new cloud platforms and introduce new features with existing command line tools, e.g., AWS CLI and qsub
- ✓ We have also integrated ep3 with other systems such as:
  - ✓ Virtual Cloud Provider (VCP)
  - ✓ CWL-metrics for collecting workflow execution records
  - ✓ Apptainer for HPC containers (ongoing)



```
class: Workflow
cwlVersion: v1.0
inputs:
…
```

Workflow description

- Modify cp_inp to use AWS CLI for s3 buckets

…

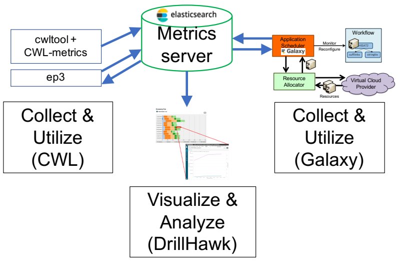
Model modification file

Extended to support s3 buckets in addition to local files

Modified state transition model

Ecosystem for Genome Analysis Workflows

- ✓ We have also developed an ecosystem to utilize execution records that consists of:
  - ✓ Metrics server that automatically collects execution records of genome analysis workflows
  - ✓ Resource scheduler for ep3 and Galaxy. It allocates and releases suitable computing resources using VCP
  - ✓ DrillHawk, which is a visualizer of workflow execution records



cwltool + CWL-metrics
ep3
elasticsearch
Metrics server
Application Scheduler
Galaxy
Monitor
Reconfigure
Workflow
HISAT2
cufflinks
stringtie
Resource Allocator
Virtual Cloud Provider
Resources
Collect & Utilize (CWL)
Visualize & Analyze (DrillHawk)
Collect & Utilize (Galaxy)