Search for computational workflow synergies in reproducible research data analyses in particle physics and life sciences

Tibor Šimko
CERN
Geneva, Switzerland
tibor.simko@cern.ch

Kyle Cranmer
New York University
New York, USA
kyle.cranmer@nyu.edu

Michael R. Crusoe
Common Workflow Language
Vilnius, Lithuania
mrc@commonwl.org

Lukas Heinrich
New York University
New York, USA
lukas.heinrich@cern.ch

Anton Khodak
Wellcome Sanger Institute
Hinxton, UK
anton.khodak@sanger.ac.uk

Dinos Kousidis
CERN
Geneva, Switzerland
dinos.kousidis@cern.ch

Diego Rodríguez
CERN
Geneva, Switzerland
diego.rodriguez@cern.ch

***Abstract*—We describe the REANA reusable and reproducible research data analysis platform that originated in the domain of particle physics. We integrated support for running Common Workflow Language (CWL) workflows that originated in the domain of life sciences. This integration allowed us to study the applicability of CWL to particle physics analyses and look for synergies in computational practices in the two communities.**

***Index Terms*—reproducible science, data preservation, data analysis, computational workflows**

I. REANA

REANA [1] is a reusable and reproducible research data analysis platform that offers tools to particle physicists to structure their input data, analysis code, compute environments and computational workflow steps so that the analysis can be instantiated and run on remote containerised compute clouds.

The researchers use a command-line client to interact with the REANA platform. The platform consists of a set of micro-services (written in Python) that communicate over REST API (using OpenAPI). The platform uses container technologies (Docker) and runs user jobs on supported compute platforms (Kubernetes, OpenStack) using several distributed storage backends (Ceph, EOS).

The basic architecture of the platform is illustrated in Figure 1.

II. Workflow-as-a-Service

The REANA platform aims at supporting diverse computational workflow practices used in the scientific community. There exists many different workflow systems that the researchers actively use, from an informal set of shell scripts running the analysis jobs to using complex structured workflow systems describing the computational steps in a formalised manner.

Simple user needs may often be expressed by means of the sequence workflow execution pattern where each step of the workflow runs linearly after the previous steps finished. The support for these sequential workflows is provided by the `Serial` workflow engine component of the REANA platform.

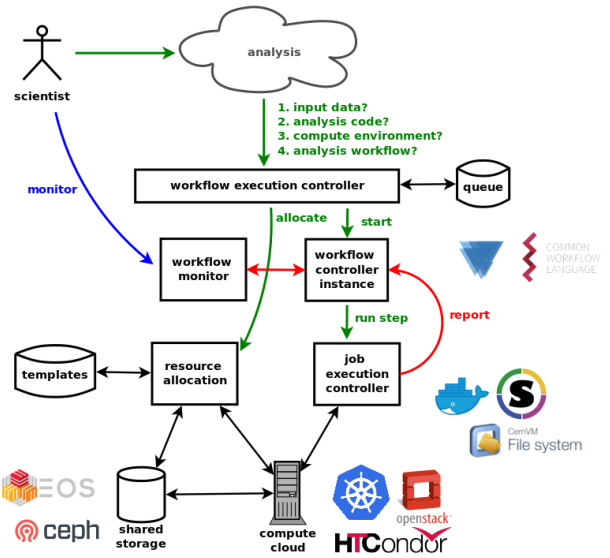

Fig. 1. A block diagram of the REANA reusable analysis platform showing the main components of the system and the supported backend tools.

More complex user needs call for expressing computational steps of the workflow as a Directed Acyclic Graph (DAG) where parts of the workflow can branch out, run in parallel, and merge back again later. The `Yadage` workflow engine [3] is one such example that was born in the domain of particle physics.

The REANA platform supports multiple workflow systems by means of pluggable workflow engine components, such as `Serial` and `Yadage`. This provides an efficient and scalable

CERN-IT-2018-004
29/10/2018

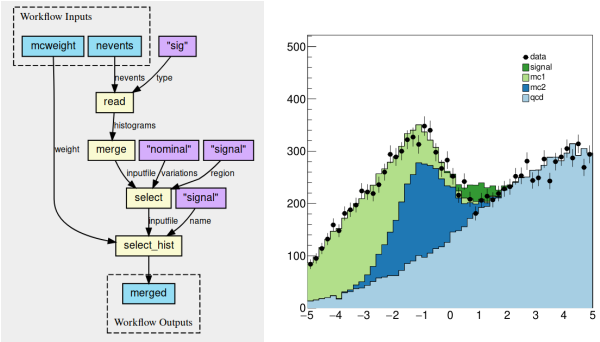

Fig. 2. A part of typical Beyond Standard Model particle physics search analysis expressed in the Common Workflow Language standard. An example plot simulating a hypothetical BSM signal (`signal`) over Standard Model background predictions (`mc1`, `mc2`, `qcd`) showing an excess of signal over data (`data`) at the pre-fit configuration of the model.

"workflow-as-a-service" solution to instantiate and manage diverse scientific workflows for diverse user groups of the platform.

III. Common Workflow Language

The Common Workflow Language (CWL) [2] standard for describing analysis workflows originated in the domain of bioinformatics and life sciences and became used in other data-intensive scientific disciplines such as astronomy. The CWL workflow system comes with several implementations and composition and visualisation tools. Our aim was to study the feasibility of expressing typical particle physics workflows in the CWL standard to explore the synergies with the CWL ecosystem.

The support for running CWL workflows was carried out by means of integrating the CWL reference implementation `cwltool`. A new `CWL` workflow engine component of the REANA platform was designed and implemented. The running of the CWL conformance test suite shows 100% compliance with the Common Workflow Language standard. The new component enables researchers to use REANA to run CWL workflows on scalable Kubernetes clouds.

A typical particle physics analysis workflow was ported from Yadage to CWL (see Figure 2). Most of the computational constructs were directly translatable to CWL idioms. Some of the advanced constructs such as "multi-level cascading map-reduce" would necessitate further extension to the CWL "scatter-gather" concept (see Figures 3, 4).

IV. Conclusions

We have described the integration of the Common Workflow Language standard with the REANA reusable analysis platform. The newly developed `CWL` workflow engine component enables researchers to run CWL workflows on Kubernetes-orchestrated containerised compute clouds.

We have studied the feasibility of using Common Workflow Language idioms in typical particle physics data analysis workflows taking the typical Beyond Standard Model search analyses as an example. The majority of the necessary particle

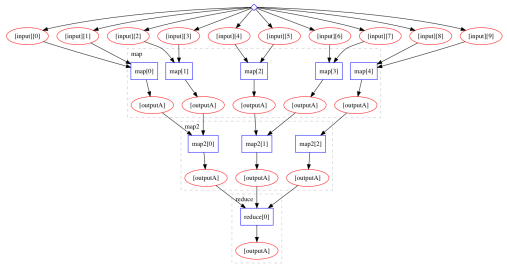

Fig. 3. An example of the multi-level cascading map-reduce idiom. The first-level "map" operates on batches of up to three nodes and generates intermediate outputs for the second-level "map" operating on batches of up to two nodes.

```
stages:
  - name: map
    dependencies: [init]
    scheduler:
      scheduler_type: multistep-stage
      parameters:
        input: {stages: init, output: input, unwrap: true}
      batchsize: 3
      scatter:
        method: zip
        parameters: [input]
  - name: map2
    dependencies: [map]
    scheduler:
      scheduler_type: multistep-stage
      parameters:
        input: {stages: map, output: outputA, unwrap: true}
      batchsize: 2
      scatter: ...
  - name: reduce
    dependencies: [map2]
    scheduler:
      scheduler_type: singlestep-stage
      parameters:
        input: {stages: 'map2', output: outputA}
```

Fig. 4. The Yadage workflow definition corresponding to the multi-level cascading map-reduce computational graph presented in Figure 3. This Yadage concept would necessitate an extension of the CWL scatter-gather idiom.

physics workflow concepts were directly supported by CWL; some advanced concepts such as the dynamically cascading map-reduce computations would necessitate further extension to the CWL scatter-gather idiom.

The present work paves the way towards reproducible science synergies in particle physics and life sciences. The developed tools are generic and can be applied to further scientific disciplines to widen the applicability of the platform in the scientific computational research data analysis ecosystem at large.

Acknowledgment

The CWL integration with the REANA platform was partially supported by the United States' National Science Foundation grant PHY-1247316 "Data and Software Preservation for Open Science" (DASPOS).

References

[1] REANA reusable analysis platform, `http://www.reana.io/`
[2] P. Amstutz, M. R. Crusoe, N. Tijanić (editors), B. Chapman, J. Chilton, M. Heuer, A. Kartashov, D. Leehr, H. Ménager, M. Nedeljkovich, M. Scales, S. Soiland-Reyes, L. Stojanovic, "Common Workflow Language, v1.0" Specification, Common Workflow Language working group, `https://doi.org/10.6084/m9.figshare.3115156.v2`
[3] K. Cranmer, L. Heinrich, "Yadage and Packtivity — analysis preservation using parametrized workflows", `arXiv:1706.01878`