Review: "Provenance tracking in the LHCb software"

Patrick O'Leary[1]

[1]Kitware

May 5, 2020

This document is a review of manuscript CiSESI-2019-01-0012 submitted to Computing in Science & Engineering:

Provenance tracking in the LHCb software (Ana Trisovic, Chris R. Jones, Ben Couturier, and Marco Clemencic).

Keywords: cisemag, reproducible research, provenance, review

Content Summary

The authors argue that the best way to foster reproducibility is to integrate it within existing scientific software that is already in use. This technique makes using reproducibility tools seamless and straightforward. They have demonstrated their solution by integrating provenance tracking in the official analysis software used at LCHb high-energy experiment at CERN.

Keywords - reproducible research, provenance.

Contribution

A proposed reproducibility, provenance tracking solution built into the LCHb analysis software. The stored provenance allows understanding how a file was produced and provides sufficient information to entirely reproduce the dataset, eliminating the need for the original input code or even documentation.

The paper is readable but requires some effort to digest. It covers some background material, skips over more fundamental content, and uses imprecise/inappropriate language at times.

References - References are sufficient and appropriate

Overview

Embedding provenance in a software system is a prevalent practice, but doing it for existing and widely used scientific software is uncommon. Thus, the contribution and differentiation, via related work, with previous publications, need to be highlighted.

Detailed Review

Introduction/Abstract/Related Work - I think the reader would benefit from a concise abstract highlighting the thesis statement and the contributions of the work. I like the one from https://arxiv.org/abs/1910.02863. Why am I not reviewing this document? The introduction starts with a list like review of the related work and ends with a thesis statement and the authors' proposed solution.

Section "The LHCb software" - This section provides an overview of the LCHb GAUDI software framework.

Notes:

- Rewrite "high-energy physics experiments like for example ATLAS [15]." Change "like for example" to "such as."
- What is meant by "in a working condition" and later as "in running condition"?
- Rewrite "There is a number of different services within the framework that can used by the Algorithms but some of the main ones are:" "There is"/are, "that can used by", ...
- "python" should be Python.

Section "Implementation of the service" - Describes the organization and methods of the metadata service.

Notes:

- The section title is vague. How about "The provenance tracking service or The metadata service?"
- Redundant, it just restates what was previously stated or at least rewrite "Therefore, the metadata is only be captured once the components and configurations are assigned to the job, at the moment when the output ROOT file is written to the disk." Change "is only be captured."

Section "Using the Provenance tracking service" - The section presents four use-cases, a code snippet for using MetaDataSvc in the Python configuration file, and two ways of examining the info file.

Notes:

- The authors haven't tied Davinci to GAUDI to MetaDataSvc in this section. You have introduced an analysis application but failed to state that it is based on the GAUDI framework.

Use-Cases

1. latest version. How does MetaDataSvc help in reproducibility?
2. minor tweaks to configuration. How does MetaDataSvc help in reproducibility?
3. multiple analysist in one filesystem. How does MetaDataSvc help in reproducibility?
4. version bug. This use-case seems tied to use-case 1. How does MetaDataSvc help in reproducibility?

- MetaDataSvc solves this, how? State how/when the version is captured, written out, where. A similar description for the configuration and whatever else.
- The code snippet "ApplicationMgr().ExtSvc += ['Gaudi::MetaDataSvc']" lacks the necessary context to make it valuable.
- "simply reading and printing the dictionary" do the authors feel that the reader will equate a dictionary used here to the key-value pairs used in a later sentence? If what was written and how was presented more precisely, then key-value pairs and dictionary could be appropriately introduced.
- Here exists the first introduction of the ROOT framework, not a collection of formats or a filesystem-like format.

Section "Conclusion" - Reiterate contributions.

Notes:

- "neat" idea? I think this needs to be presented using more appropriate language.

- Did you demonstrate a number of scenarios or just introduce some uses?
- "that seamlessly mash into researchers' work" I think this needs to be presented using more appropriate language.

Questions

1. How easy was this to implement in a service-based architecture/framework?
2. Do you think this technique could be employed in other architectures as easily?
3. The work was done in 2015 and incorporated into the GAUDI repository in 2017. Are there any end-user results? Is it widely used? Do reesearchers commonly user the info files to address concerns presented in your four use-cases?