Physics Examples for Reproducible Analysis

Student: Daniel Prelipcean
Technical University of Munich, Germany

Supervisor: Dr. Tibor Šimko
CERN, Switzerland

CERN Summer Student Programme

September 20, 2019

Abstract:

REANA is a reusable and reproducible data analysis platform allowing researchers to structure their analysis pipelines and run them on remote containerised compute clouds, for preservation and reproducibility purposes. The present work aimed at creating new REANA pilot examples and enhancing existing ones in order to test and improve the user experience of the platform. These examples showcase both: i) the variety of use cases that REANA is adapted to; ii) the technology stack behind it. The focus is given on the explicit work carried out and the knowledge attained during the Summer Programme.

Index terms: reproducible science, computational workflows, data analysis

1 Introduction

The REANA (REusable ANAlysis) platform [1] is a reproducible and reusable data analysis platform allowing researchers to specify computational workflow steps in several declarative languages (such as CWL [2] or Yadage [3]) to run their data analysis pipelines on containerised compute clouds, using different compute backends (such as Kubernetes and HTCondor). The computational workflows associated to research data analyses can be represented in the form of a Directed Acyclic Graphs (DAG) that topologically order the analysis steps and define their dependencies. The DAG workflows may consist of running several tens of thousands of computational tasks. The platform was developed with data analysis reproducibility and reusability in mind. [4]

2 Reusable Analysis Examples

To support the community to adopt reproducible science practices in their analyses, several REANA examples have been created, covering a wide range of use cases. Their purpose is to demonstrate the usage of REANA in the context of high energy particle physics data analysis carried out at the European Organization for Nuclear Research (CERN), using software specific to each collaboration/experiment. The workflow to create these examples is illustrated in Fig. 1.

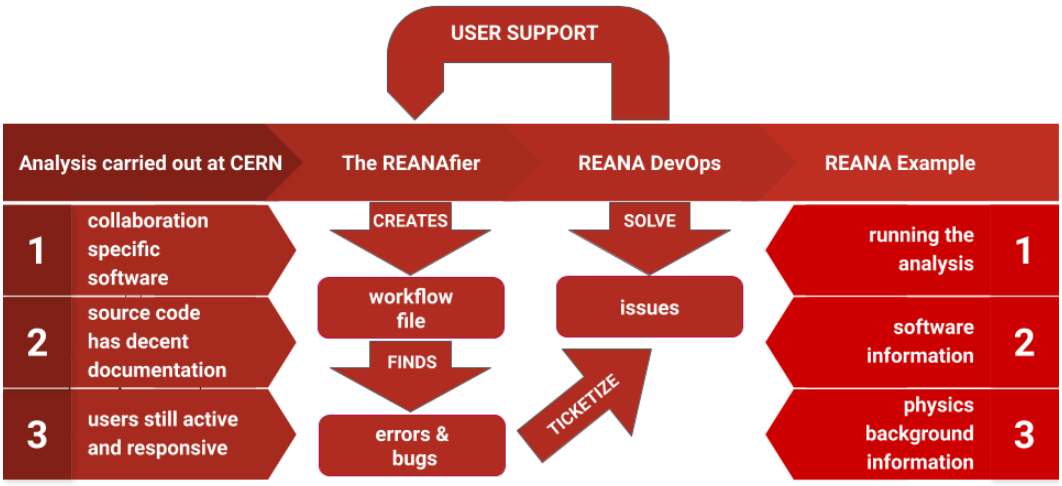

Figure 1: Typical workflow to create REANA examples from existing analysis carried out at CERN. As a first result of this Summer Project, the user experience has been vastly improved by adding new functionalities to REANA.

2.1 ALICE Transverse Momentum P_t Analysis Example

The CERN Open Data Platform

CERN has embarked the initiative to openly publish data from LHC experiments as open data to the public. [5] "The CERN Open Data portal is the access point to a growing range of data produced through the research performed at CERN. It disseminates the preserved output from various research activities, including collision and simulated datasets, with accompanying software and documentation, which is needed to understand and analyse the data being shared.

The portal adheres to established global standards in Data Preservation and Open Science: the products are shared under open licenses; they are issued with a digital object identifier (DOI) to make them citable objects in the scientific discourse." [6] Moreover, they adhere to the FAIR Guiding Principles for scientific data management and stewardship. [7]

However, there is no explicit guarantee that any analysis code published online is actually working several years after it was originially published, if there is no continuous integration tool that checks it (GitLab integration has been another Summer Project). Therefore, we wanted to test whether the ALICE open data example analysis can be actually reproduced using the information provided on the portal at the time of its publishing.

A new REANA example has been set up using a material published by the ALICE collaboration, namely: a simple transverse momentum P_t analysis script using any of the published datasets. The original analysis has been successfully reproduced using a containerised AliPhysics environment.

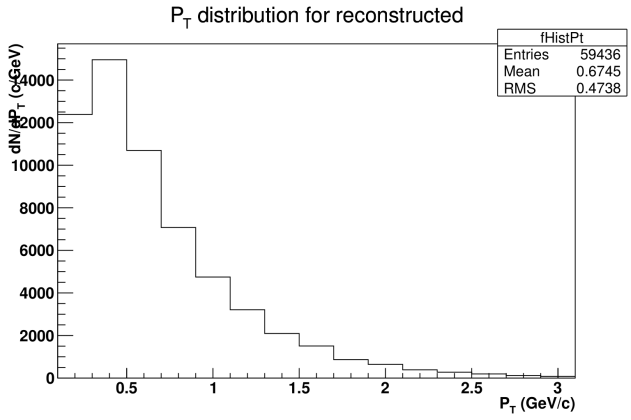

Figure 2: Figure obtained from CERN Open Data platform, without any changes in the published code, using the Pb-Pb data sample at the collision energy of 2.76 TeV per nucleon pair from run number 139038.

2.2 LHCb Rare Charm Decay Search Example

Struggles using Large Local Files and ROOT warnings

Data analysis in experimental partile physics is notorious for the humongous amount of data that it handles. For most REANA examples, datasets are pulled from different (mostly online) sources, and are not uploaded directly using the reana-client upload file option. Some use cases, however, require manual uploads of files by the researchers.

This LHCb example studying the rare decay $D^+_{(s)} \to \pi^+\mu^+\mu^-$ uses a 10 GB ROOT file. Debugging or changing parameters for different tests would then require uploading this large file at each run, which is redundant after the first upload in the REANA workspace. Hence, future work includes:

i allowing upload of files larger than the memory size by buffering technique.

ii reusing the same file(s): REANA architecture could be restructured to use a shared workflow workspace, making debugging or iterative changes easier.

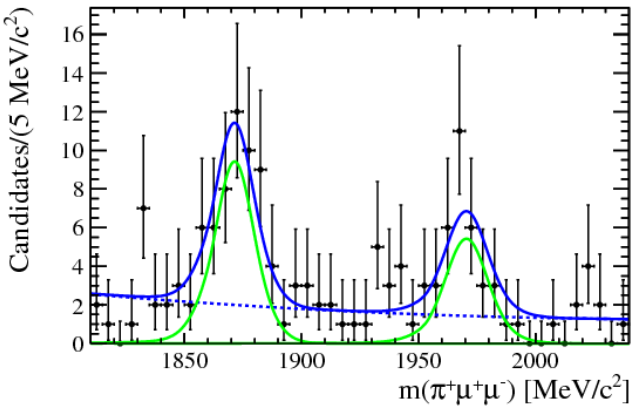

Figure 3: Centrality plot, obtained by running the analysis locally

Moreover, the ROOT analysis scripts would cause several warnings, although not leading to fatal errors. Nevertheless, such messages overflow the logging output and had to be cleaned. The ROOT forum is a prime example of user support at CERN.

2.3 ATLAS RECAST Example

Dealing with Docker User Permissions

One of the main technological advancements that allows now reproducibility of science are the docker images/containers, which are isolated from one another and bundle their own software, libraries and configuration files. The REANA docker images are openly available online.

This REANA reproducible analysis example demonstrates a RECAST [8] analysis using ATLAS Analysis Software Group stack.

The user (e.g. researcher doing data analysis) only has to create a simple Dockerfile that mentions the required libraries for the execution of their code. One of their main features is that docker containers are more lightweight than virtual machines.

For this REANA example for the ATLAS collaboration, some files are in private workspaces not accessible by all users inside the docker container, hence the need to slightly change the original docker containers.

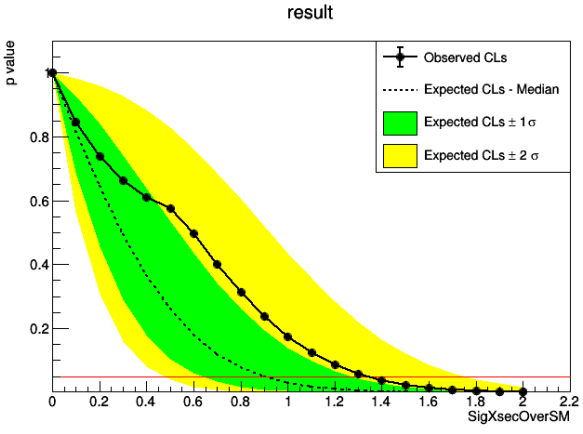

Figure 4: Statistical analysis done via the RECAST method.

2.4 CMS Higgs to Four Leptons Decay

Entrypoint scripts
The discovery of the Higgs boson [9] created a lot of enthusiasm, and parts of this achievement have been reproduced inside REANA with this example.

It studies the Higgs-to-four-lepton decay channel $H \to ZZ^* \to 4$ leptopns that led to the Higgs boson experimental discovery in 2012. The example uses CMS open data released in 2011 and 2012.

For convenience purposes, the CMSSW docker image automatically sets the environment variables and aliases for the CMS user. When running the image through REANA though, the CMS entrypoint script is overwritten and it has to be re-executed.

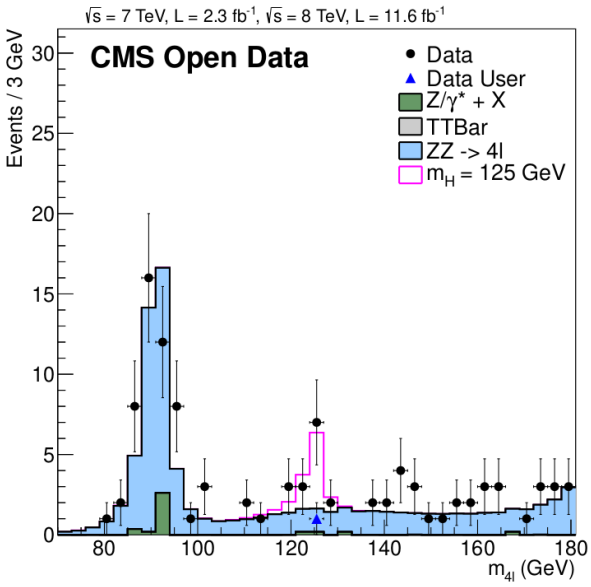

Figure 5: Possible Higgs decay to four leptons.

3 Workflow Factory for CMS AOD Reconstruction

Some projects may have standardized/parametrized workflows, which only require some parameter changes for a different run, with small or no code changes. Such a project is the CMS reconstruction for file formats: from RAW to Analysis Object Data (AOD) files. For this purpose, we have designed a workflow factory demo that automatically creates the workflows based only on the selection of the dataset.

The workflow can be logically divided into two steps:

i) Polling information from the associated CERN Open Data page of the dataset
ii) Change the template files using the polled information

Three such workflows (for different years and datasets) have been manually created and tested, to be used as references for the workflow factory.

```
$ cms-reco --load-config recid=39
# alias for cernopendata-client get-record --recid 39 | tee config.json
$ cms-reco --create-workflow
#   Directory cms-reco-SingleElectron-2011 created.
$ cd cms-reco-SingleElectron-2011
$ reana-client run
```

Figure 6: Commands to create and execute a workflow using mostly default parameters for the dataset with record id 39, namely the SingleElectron 2011 dataset.

This significantly improves the user experience: before, one would have to manually extract all relevant information and create a reana.yaml file from scratch; but now, all the user has to do is install the package and run a simple (customizable) command. This paves the way for a more general workflow factory that could be used for arbitrary workflows.

4 Workflow Systems and Partial Workflow Execution

The main and simplest workflow system is serial (built in-house), which REANA uses by default. However, more interesting examples benefit from more complicated workflow systems, which may have worthy features, e.g. parallelism.

One regular task for REANA is to constantly update to the latest version of the respective workflow system, which may require API changes and other functionalities. Nevertheless, one feature that deemed important is the partial workflow execution, which implies that the user may specify how many and which steps of the workflow are actually executed. This proves itself important in debugging.

As the CWL and YADGE have implemented this tool, it felt natural to include this feature for the serial workflow system as well.

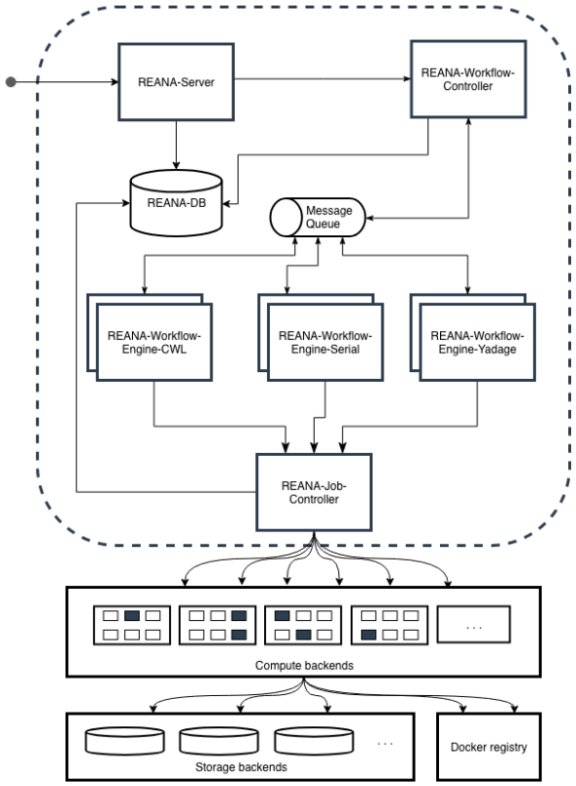

Figure 7: REANA cluster architecture and its components. Note the support for several workflow systems (CWL, Serial, Yadage). [1]

5 Conclusion

The aim of the REANA reproducible analysis platform is to provide modern information technology tools and solutions via a platform to scientists, such that they can structure their data analys in a reproducible, preservable, and shareable manner. This poses both technological and sociological challenges. On the technology side, one has to deal with large container sizes and massively parallel workflows. On the sociological side, one has to persuade researchers that adopting reproducibility best practices is not cumbersome, but it is actually helping their daily research.

We have taken several typical examples from both data production and data analysis phases of the research conducted at CERN in order to study the feasibility and to demonstrate the applicability of the REANA approach. We hope that the advancement of technology through readily-applicable examples will help to drive the cultural change towards fair and open science.

Acknowledgements

Over the past 13 weeks, I have received help and guidance from so many amazing people. Firstly, I would like to thank my supervisor, Dr. Tibor Šimko, for his guidance and introducing me to this wonderful open science movement.

I am proud to have been part of the IT-CDA-DR department at CERN, working with the Open and Reproducible Research team: Diego Rodriguez Rodriguez, Rokas Mačiulaitis, Jan Okraska and Leticia Wanderley. Their help was the key to any progress in my project. Additionally, all the other members of the department were extremely friendly, helpful and informative.

Secondly, thank you to Jean-Yves Le Meur for selecting me and giving me so many projects to choose from for the best of my abilities.

Additionally, thank you to the CERN Summer Student team, especially Adriana Bejaoui. It has been the most incredible experience, and I am so grateful to have taken part in CERN.

References

[1] T. Šimko, L. Heinrich, H. Hirvonsalo, D. Kousidis, D. Rodrıguez, *"REANA: A system for reusable research data analyses"*, Computing in High Energy Physics 2018, Sofia, Bulgaria, 9–13 July 2018. DOI: https://doi.org/10.1051/epjconf/201921406034

[2] P. Amstutz, M. R. Crusoe, N. Tijanić (editors), B. Chap- man, 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"*. https://arxiv.org/abs/1706.01878

[4] X. Chen, S. Dallmeier-Tiessen, R. Dasler, S. Feger, P. Fokianos, J. .B. Gonzalez, H. Hirvonsalo, D. Kousidis, A. Lavasa, S. Mele, D. Rodrıguez, T. Šimko, T. Smith, A. Trisovic, A. Trzcinska, I. Tsanaktsidis, M. Zimmermann, K. Cranmer, L. Heinrich, G. Watts, M. Hildreth, L. Lloret Iglesias, K. Lassila-Perini, S. Neubert, *"Open is not enough"*, Nature Physics **15** 113–118 (2019). https://www.nature.com/articles/s41567-018-0342-2

[5] CERN announcement about publishing LHC data on the COD platform Retrieved on September 17^{th}, 2019, from: home.cern/news/news/accelerators/cern-makes-public-first-data-lhc-experiments

[6] CERN Open Data Platform. Retrieved on September 12^{th}, 2019, from: opendata.cern.ch/docs/about

[7] M. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. Boiten, L. da Silva Santos, P. Bourne, J. Bouwman, A. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, R. Finkers, A. Gonzalez-Beltran, A. Gray, P. Groth, C. Goble, J. Grethe, J. Heringa, P. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. Lusher, M. Martone, A. Mons, A. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B.Mons, *"The FAIR Guiding Principles for scientific data management and stewardship"*, Scientific Data **3** 160018 DOI: https://doi.org/10.1038/sdata.2016.18

[8] K. Cranmer, I. Yavin, *"RECAST: Extending the Impact of Existing Analyses"*, Journal of High Energy Physics, 12 Oct 2010. https://arxiv.org/abs/1010.2506

[9] The CMS collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", Phys. Lett. B 716 (2012) **30**. https://arxiv.org/abs/1207.7235