On-Demand Distributed Computing Workflow for Physics Analysis at the CMS Experiment

Diyaselis Delgado[1], Kati Lassila-Perini[2], Clemens Lange[3]

[1] University of Puerto Rico Mayaguez, [2] Helsinki Institute of Physics, [3] European Organization for Nuclear Research

Abstract

The CMS experiment is a strong contributor to the CERN Open Data Portal. CMS Open Data project releases data collected from proton-proton collisions at the LHC to the public. It publishes research level data together with the environment, software and instructions. These data can be used by the scientific community and the general public for physics analysis. This talk will describe the development and tests of simplified examples for the use of open data, analysis preservation and a new reproducible research data analysis platform.

Introduction

CERN Analysis Preservation (CAP)

The CAP Framework[1] is a central platform for the four LHC collaborations, developed to address the need for the long-term preservation of the data analysis process.

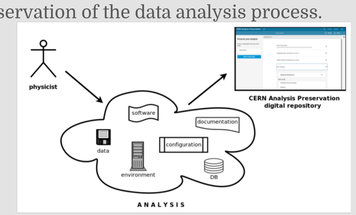

Figure 1: Composition of an analysis for CAP

Capturing knowledge and assets of individual physics analyses in a digital repository in view of facilitating their future reuse.

ReANA Platform

reana[2] is a reusable and reproducible research data analysis platform. It structures input data, analysis code, containerized environments and computational workflows. The ReANA platform supports multiple scenarios, such as: computing clouds, running environments, resource orchestration tools, workflow engines, and shared storage systems.

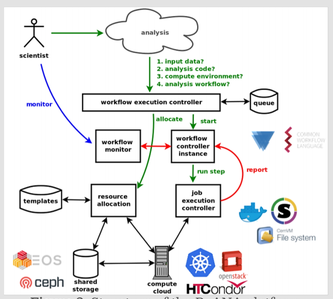

Figure 2: Structure of the ReANA platform

ReANA is applicable to any scientific discipline, but it is of vast interest for physics analysis done at CERN. ReANA, in essence, builds a system to instantiate preserved analysis on the cloud.

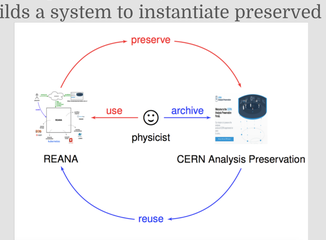

Figure 3: Integration of the CAP and ReANA platforms

Higgs-to-four-lepton analysis example using 2011-2012 data

A simplified reimplementation of the Higgs discovery[3] using CMS Open Data inputs available with appropriate CMS Software environment. ReANA, captures the analysis workflows and runs the commands to obtain the output plot.

The characteristics that make this analysis reproducible on the ReANA platform are:

- **Inputs**: 2011-2012 CMS collision and simulated data
- **Code**: Configuration files and analyzer code (Python, C++)
- **Environment**: CMS Software image from Docker container
- **Workflows**: CWL (Common Workflow Language) layout

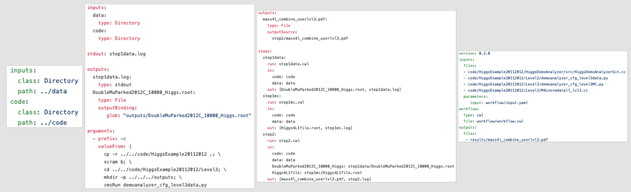

Figure 4: Workflows for Higgs Analysis. (Left to right) Workflows stating inputs, environment, steps, workflow sequence, and ReANA interpreter.

On ReANA, the analysis will produce an output file with one Higgs and signal distributions. Then, both files are analyzed and produce the output plot.

Figure 5: (Left) The output plot from running the analysis on the ReANA cluster. (Right) The resulting plot published at the original paper of the Higgs discovery.

Differences in the data points and margins are most noticeable yet both plots present the expected Higgs mass at 125 GeV. Attributions from Monte Carlo simulated values, already weighted by luminosity, cross-section and number of events, are separated by ZZ, a pair of heavier bosons, ttbar, a pair of top and anti-top quarks, and some irreducible background from singular Z bosons. At the end, adding these contributions plus data result in a distribution of the four-lepton reconstructed mass in the low mass region for the sum of the $4e$, 4μ and $2e\,2\mu$ channels.

Running the analysis on ReANA cluster

This CERN project interprets and executes workflows for any complex or simple analysis. In addition to the existing code, only a set of descriptive workflow files are needed. To use within the cluster, the user must install a ReANA command-line client.

```
$ # install REANA client:
$ mkvirtualenv reana-client
$ pip install reana-client
$ # connect to some REANA cloud instance:
$ export REANA_SERVER_URL=https://reana.cern.ch/
$ export REANA_ACCESS_TOKEN=XXXXXXX
$ # create new workflow:
$ reana-client create -n my-analysis
$ export REANA_WORKON=my-analysis
$ # upload input code and data to the workspace:
$ reana-client upload ./code ./data
$ # start computational workflow:
$ reana-client start
$ # ... should be finished in about a minute:
$ reana-client status
$ # list workspace files:
$ reana-client list
$ # download output results:
$ reana-client download results/mass4l_combine_userlvl3.pdf
```

Figure 6: Commands to test client-to-server connection, create a new workflow setting, upload code and start the workflow

Reprocessing AOD from 2010-2012 RAW samples

Validation examples[4] prove that a result can be obtained with the legacy or open data tools. The inputs are datasets specific to every example, and the code for the analysis used with the CMS Open Data Virtual Machine environment. The ReANA setup allows the reproducibility of CMS Open Data large-scale analyses and helps to confirm if instructions are correct.

- **Inputs**: Selected RAW datasets

/MinimumBias/Run2010B-v1/RAW	/SingleMu /Run2011A-v1/RAW
/Electron/Run2010B-v1/RAW	/Jet /Run2011A-v1/RAW
/Mu/Run2010B-v1/RAW	/DoubleElectron/Run2012B-v1/RAW
/Jet /Run2010B-v1/RAW	/SingleElectron /Run2012B-v1/RAW
/DoubleElectron/Run2011A-v1/RAW	/DoubleMuParked/Run2012B-v1/RAW
/SingleElectron /Run2011A-v1/RAW	/SingleMu /Run2012B-v1/RAW
/DoubleMu /Run2011A-v1/RAW	/JetHT/Run2012B-v1/RAW

- **Code**: Configuration file created using a CMS tool
- **Environment**: CMS Software embedded in a Virtual Machine

After data reconstruction, a plotting code[5] loops over different physics objects, such as tracks, electrons, muons, photons, jets, taus and missing et, and fills histograms with p, pt, eta and phi of these objects.

The newly reprocessed AOD from RAW files are compared to the results given from the corresponding AOD files from the Open Data Portal, in order to validate the results. Taking in consideration the histograms showing the electron momentum for all the samples tested, it was easy to determine their relation.

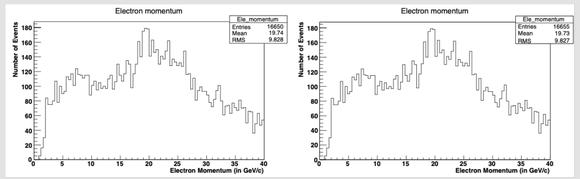

Figure 7: (Left) Resulting histogram of electron momentum using reconstructed AOD file from /Electron/Run2010B-v1/RAW sample. (Right) Output histogram of electron momentum from the Open Data AOD corresponding file.

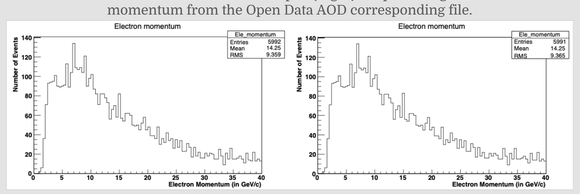

Figure 8: (Left) Resulting histogram of electron momentum using reconstructed AOD file from /Jet/Run2010B-v1/RAW sample. (Right) Output histogram of electron momentum from the Open Data AOD corresponding file.

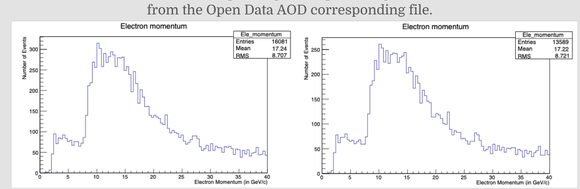

Figure 9: (Left) Resulting histogram of electron momentum using reconstructed AOD file from /DoubleMu/Run2011A-v1/RAW sample. (Right) Output histogram of electron momentum from the Open Data AOD corresponding file.

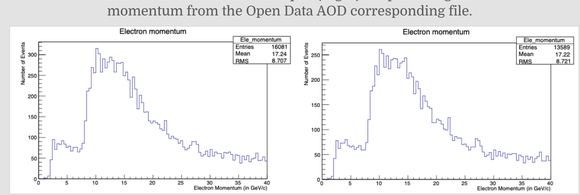

Figure 10: (Left) Resulting histogram of electron momentum using reconstructed AOD file from /DoubleElectron/Run2011A-v1/RAW sample. (Right) Output histogram of electron momentum from the Open Data AOD corresponding file.

Working Examples on the ReANA Platform

The Higgs-to-four-lepton analysis example and the data reconstruction of Open Data process have been implemented in ReANA., and runnable for public testing. Also, these CMS Open Data examples are up and running on the CERN Open Data Portal.

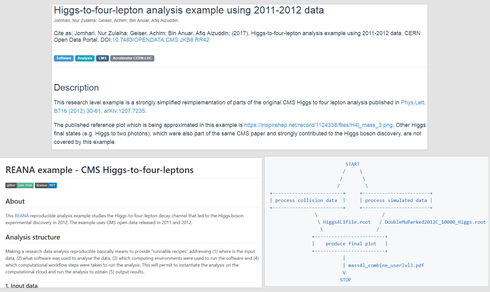

Figure 11: Published CMS examples found in the CODP and GitHub page for ReANA analyses[6]

Conclusion

In this work, CMS analysis examples from CERN Open Data Portal are taken as examples to connect the three data preservation services, CODP, CAP and ReANA. The preservation of these physics analysis examples ensures that the complete information is structured and reproducible. Workflow steps for the examples were tested locally and on a remote server, proving to be successful in producing the expected outputs. In ReANA, access to the experimental datasets, analysis software area, operating system environment and computational workflow steps remains to be tested. Overall, the implementation of preservation and reproducibility through the CERN Analysis Preservation and ReANA looks promising as more examples continue to be evaluated.

References

[1] https://analysispreservation.cern.ch/
[2] http://www.reana.io/
[3] http://doi.org/10.7483/OPENDATA.CMS.JKB8.RR42
[4] CMS Open Data Validation Analyses. http://opendata.cern.ch/search?page=1\&size=20\&q=validation\%20cms&subtype=Validation\&type=Software\&experiment=CMS
[5] http://doi.org/10.7483/OPENDATA.CMS.11RI.SDX7
[6] ReANA CMS Examle. https://github.com/reanahub/reana-demo-cms-h4l

Acknowledgements

- This project was funded by the CERN Non-Member State Summer Program, and the Physics Department of University of Michigan.
- Special thanks to the National Science Foundation for supporting this poster.
- Many thanks to Dr. Sudhir Malik for supporting me while I further my research experience.