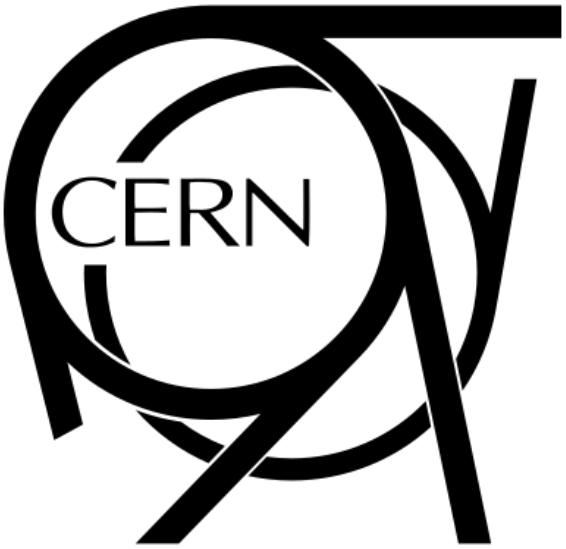

European Organization for Nuclear Research,

Summer student final report

CERN Open Data Portal

Supervisor: MS. Kati Lassila-Perini, PhD.
Author: Bc. Lukas Holub

Geneva, summer 2016

.

Abstract

The main goal of this report is to summarize my work at CERN during this summer. My first task was to transport code and dataset files from CERN Open Data Portal to Github that will be more convenient for users. The second part of my work was to copy environment from CERN Open Data Virtual MAchine and apply it in the analysis environment SWAN. The last task was to rescale X-axis of the histogram.

Keywords: CERN Open Data Portal, SWAN, CERN Virtual Machine, CERN environment

Introduction

The CERN Open Data Portal is the access point to a growing range of data produced through the research performed at CERN.[1] It disseminates the preserved output from various reaserch activities, including accompanying software and documentation which is needed to understand and analyse the data being shared.

Data produced by the LHC experiments are usually categorised in four different levels. The Open Data Portal focuses and uses two of these data:

- Simplified data formats for analysis in outreach and training exercises.
- Reconstructed data and simulations as well as the analysis level software to allow a full scietific analysis.

All datasets and other material available in this portal are minted with a persistent identifier, a so called Digital Object Identifier (DOI) that allows permanent linking to the records.

The validation benchmark analyses for CMS data

Several benchmark analyses have been provided on the CERN Open Data Portal in order to validate different primary datasets [2]

A analysis requires suitable and available software. Theis software is provided with CERN Open Data Virtual Machine with the CMS computing environment. In order to run analysis, some steps are required. The main goal of my work was to simplify this procedure. For this purpose the Git/Github.com were used.

Git [4] is a version control system that is used for software development and other version cobtrol tasks. Another words, Git allows a team of people to work together, all using the same files.

Github [5] is a web-based Git repository hosting service. It offers all of the distributed version control and source code management functionality of Git as well as adding its own features. Unlike Git, which is strictly a command-line tool, Gihub provides a Web-based graphical interface and desktop as well as mobile integration and much more.

On the Github.com, cms-opendata-validation organization was created. Repositories were created for validation benchmark analyses. I was working with these three repositories:

- 2010-commissioning-dimuon
- 2010-mu-mumonitor-dimuon
- 2010-minimumbias-trackvertex

Into these repositories, code and dataset files were downloaded from CERN Open Data Portal. Each repository includes almost the same code files that are written in `Python` and `C++` programming languages. There is also one `xml` document and folder `datasets` . This folder contains index files which are sources of the data and available on the Open Data Portal. The folder `datasets` was created here for convenience.

Another important task was to write a simple readme file for users. This file includes requirements and detailed instructions, so that users can handle everything without problems. There was also requirement from my supervisor in order all files from mentioned Github repositories could be situated in the different directories such as:

- Validation/Commissioning_dimuon_2010
- Validation/Mu-Mumonitor_dimuon_2010
- Validation/TrackVertex_2010

These directiories are created during cloning of repositories from Github. However, one problem appeared because when new directories were added the path and plug-in were changed. To repair the problem with paths I had to rewrite some lines in the original analysis code file. For solving of the problem with plug-in I had to rewrite declarations of classes and functions in the file `prefixAnalyzer.cc` which is situated in folder src/ and for every repository has different prefix.

Afterwards, I had to find out whether everything works fine and also how much time analysis takes approximately. During this testing I found out that there can be

small issue. A running analysis can kill itself when more programs run at the same time. Therefore it is better to run only one analysis at a time because this CERN Open Data Virtual Machine does not have to handle it.

Unresolved issues

At the end of my stay at CERN, I tried to solve two problems which I could not complet.

The first one is how to set an environment from CMS Open data Virtual Machine to SWAN. Another words, we would be able to run examples from CMS Open Data Virtual Machine in SWAN. SWAN is platform to perform interactive data analysis in the cloud. Thanks to this service we can analyse data without the need to install any software.

The second issue was not so difficult but there was not time to find out how to solve it. The main idea of the problem was how to change X-axis in logarithmic scale into the linear scale while histogram will be drawn in logarithmic scale. I was able to change it into the form which is shown in the picture below. (Fig:1) The only problem is that values on the X-axis do not correspond with the real physical values.

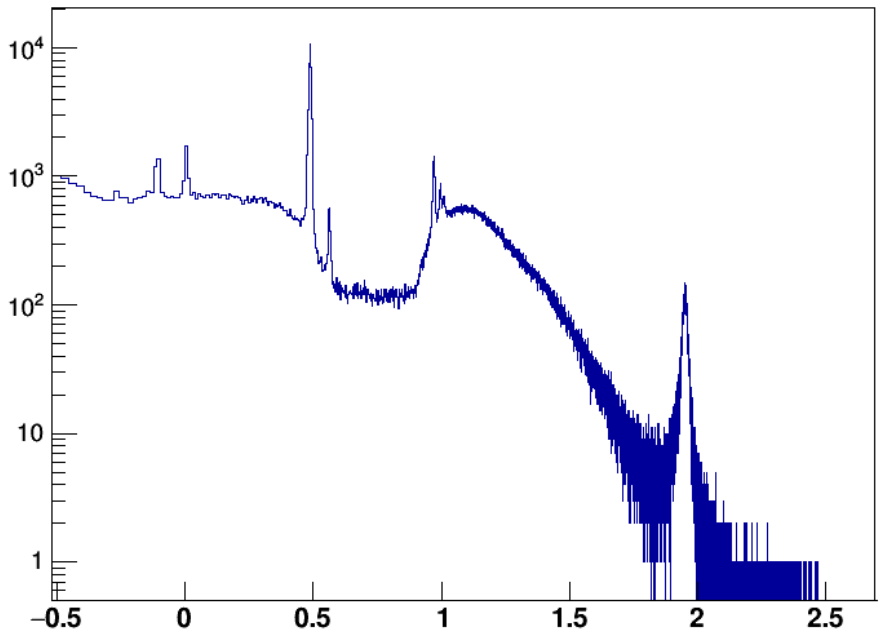

Figure 1: X-axis of histogram is in linear scale while histogram is drawn in logarithmic scale.

Bibliography

[1] CERN. *CERN Open Data Portal*. http://opendata.cern.ch/.

[2] CERN. *CERN Open Data Portal*. http://opendata.cern.ch/collection/CMS-Primary-Datasets.

[3] CERN. *SWAN*. https://swan001.cern.ch/hub/home.

[4] Git. *Git*. https://git-scm.com.

[5] Wikipedia. *Github*. https://de.wikipedia.org/wiki/GitHub.