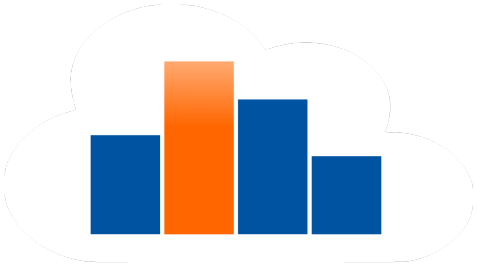

Contribution ID: **29** Type: **not specified**

Integrating CMSSW in SWAN

Friday 11 October 2019 11:45 (20 minutes)

The TOTEM and CMS experiments use the software framework for offline data processing, namely CMSSW. The framework provides necessary data formats for accessing and analysing RECO and AOD objects. In our study we investigated the feasibility of using SWAN, RDataFrame and Spark technologies to examine, analyse and reduce 500 TB data sample generated by the TOTEM detectors. The sample itself is a collection of ROOT files in RECO format, stored temporarily on EOS instance in CERN.

We successfully managed to implement the analysis code using the RDataFrame in C++ and run it on lxplus cluster with the standard CMSSW setup. Next, we attempted to run it on SWAN in order to use the Spark cluster. This attempt was not successful yet, due to the following reasons:

- Data in RECO format is not the NTUPLE, so it contains CMSSW specific classes (i.e. vectors of tracks)
- The analysis requires ROOT dictionaries to process the data and perform calculations
- The above requires loading relevant subset of CMSSW libraries to the analysis program running on SWAN.

The detailed description of the issues we found is the following. SWAN relies on CVMFS for software releases, letting the user choose from the latest LCG releases and nightlies. The core functionalities thus make use of software under the sft.cern.ch CVMFS namespace, while the current analysis use case was originally developed within the CMSSW framework and still relies on software and libraries inside it, especially those regarding the AOD data format. Keeping in mind the final goal, that is to exploit the potential of the Spark clusters to run the analysis, the first simplest step to take was to run the code directly in the SWAN terminal. This is possible via some modifications to PATH environment variables in the SWAN user session. The next step involved using a Jupyter notebook to run the same code. This could be achieved by creating an "environment script" in bash, but still some silent errors started to show up and could only be seen in the SWAN logs. The final step, connecting to the Spark clusters to run the analysis, was simply impossible with the current status of the platform. The main problem revolves around having two CVMFS software stacks, namely SFT and CMS, clashing at different levels. For instance, ROOT relies also on external tools such as the gcc compiler and the python interpreter, which were picked from the corresponding CMS stack. At the same time, Jupyter and its extensions as well as the whole Spark framework are in the SFT stack but not in the CMS one, making the connection to the clusters simply incompatible with the environment needed for the analysis.

In summary, we observed that integrating CMSSW in SWAN is only possible if considering it as a terminal interface, that is picking only ROOT and the CMS data format libraries via modifying some environment variables. Using the Jupyter notebook and the Spark clusters is incompatible at this moment with the environment needed for the analysis. Such setup would involve picking ROOT from the CMS repository and the Jupyter extensions and Spark framework from the SFT repository.

Authors: PIPARO, Danilo (CERN); CERVANTES VILLANUEVA, Javier (CERN); BAK, Karol Remigiusz (AGH University of Science and Technology (PL)); GRZANKA, Leszek (AGH University of Science and Technology (PL)); MALAWSKI, Maciej (AGH University of Science and Technology (PL)); AVATI, Valentina (AGH University of Science and Technology (PL)); PADULANO, Vincenzo Eduardo (Universita & INFN, Milano-Bicocca (IT))

Presenter: GRZANKA, Leszek (AGH University of Science and Technology (PL))

Session Classification: User's use cases