Contribution ID: **138** Type: **Presentation**

Running Parallel, Distributed ROOT Analysis with PyRDF on Public Cloud - AWS Lambda Case Study

Tuesday 26 January 2021 09:30 (15 minutes)

ROOT framework is a standard tool for HEP data storage and analysis. Current approaches for parallel and distributed processing in ROOT require either batch systems such as HTCondor, or big data frameworks like Apache Spark supported by the recently added PyRDF interface. However, these approaches are heavily reliant on existing scientific infrastructure. On the other hand, cloud computing allows provisioning resources on demand in increasingly elastic and fine-grained ways, as in example of serverless computing and function-as-a-service model.

To make it possible to run parallelized ROOT without any dependency on existing resources in research computing centers, we explore the possibility of using public clouds with serverless paradigm, creating our infrastructure on-demand. The goal is to allow deploying and running the ROOT analysis on public cloud resources. As a proof of concept, we developed a prototype of a new distributed processing backend to PyRDF interface, to support running ROOT analysis workflows on AWS Lambda in the same way as on the existing Apache Spark backend.

We report on our experience with serverless infrastructure, starting with compiling and deployment of the entire ROOT environment with its dependencies, through usage of CERN resources without any CERN-specific software, up to the point of connecting multiple Serverless Functions using PyRDF to mimic the existing PySpark environment. The technologies we employed include Terraform for deployment of our application, boto for client-AWS integration, Docker for simpler installation, and AWS S3, Lambda and EFS services for underlying infrastructure. We outline the issues, the possibilities, technical limitations and current roadblocks waiting to be solved for the tool to be used easily by anyone.

Acknowledgments:

We would like to thank the ROOT team for their support and discussions, in particular to Vincenzo Padulano, Enric Tejedor and Vassil Vassilev.

This work was supported by the Polish Ministry of Science and Higher Education, grant DIR/WK/2018/13.

Primary authors: Mr KUSNIERZ, Jacek (AGH University of Science and Technology (PL)); Mr PASTERNAK, Piotr (AGH University of Science and Technology (PL)); MALAWSKI, Maciej (AGH University of Science and Technology (PL))

Presenter: Mr KUSNIERZ, Jacek (AGH University of Science and Technology (PL))

Session Classification: Novel Data Science Environments

Track Classification: Main session: User Voice: Novel Applications, Data Science Environments & Open Data