Contribution ID: **14** Type: **not specified**

Evolution of interactive data analysis for HEP at CERN –SWAN, Kubernetes, Apache Spark and RDataFrame

Wednesday 27 March 2019 14:50 (25 minutes)

This talk is focused on recent experiences and developments in providing data analytics platform SWAN based on Apache Spark for High Energy Physics at CERN.

The Hadoop Service expands its user base for analysts who want to perform analysis with big data technologies - namely Apache Spark –with main users from accelerator operations and infrastructure monitoring. Hadoop Service integration with SWAN Service offers scalable interactive data analysis and visualizations using Jupyter notebooks, with computations being offloaded to compute clusters - on-premise YARN clusters and more recently to cloud-native Kubernetes clusters. The ROOT framework is most widely used tool for high-energy physics analysis. Its integration with SWAN allows physicists to perform web-based interactive analysis using standard tools and libraries, in the cloud.

The first part of presentation will focus on integration of Spark on Kubernetes into SWAN service, which allows to offload computations to elastic, virtualized and container-based infrastructure in the private or public clouds, compared to complex to manage and operate on-premise Hadoop clusters.

The second part will focus on evolutions in exploiting analytics infrastructure - namely new developments in ROOT framework –Distributed RDataFrame - which would allow interactive, parallel and distributed analysis on large physics datasets by transparently exploiting dynamically pluggable resources in SWAN, e.g. Hadoop or Kubernetes clusters.

Author: MROWCZYNSKI, Piotr (CERN)

Co-authors: KOTHURI, Prasanth (CERN); TEJEDOR, Enric (CERN)

Presenter: MROWCZYNSKI, Piotr (CERN)

Session Classification: Computing & Batch Systems

Track Classification: Computing & Batch Services