Contribution ID: **240** Type: **presentation**

Using Big Data Technologies for HEP analysis

Tuesday, 10 July 2018 11:15 (15 minutes)

The HEP community is approaching an era were the excellent performances of the particle accelerators in delivering collision at high rate will force the experiments to record a large amount of information. The growing size of the datasets could potentially become a limiting factor in the capability to produce scientific results timely and efficiently. Recently, new technologies and new approaches have been developed in industry to answer to the necessity to retrieve information as quick as possible by analyzing PB and EB datasets. Providing the scientists with more modern computing tools will lead to rethinking the principles of data analysis in HEP, making the overall scientific process faster and smoother.

In this talk, we are presenting the latest developments and the most recent results on the usage of Apache Spark for HEP analysis. The study aims at evaluating the efficiency of the application of the new tools both quantitatively, by measuring the performances, and qualitatively, focusing on the user experience. The first goal is achieved by developing a data reduction facility: working together with CERN Openlab and Intel, CMS replicates a real physics search using Spark-based technologies, with the ambition of reducing 1 PB of public data collected by the CMS experiment to 1 TB of data in a format suitable for physics analysis in 5 hours.

The second goal is achieved by implementing multiple physics use-cases in Apache Spark using in input preprocessed datasets derived from official CMS data and simulation. By performing different end-analyses up to the publication plots on different hardware, feasibility, usability and portability are compared to the ones of a traditional ROOT-based workflow.

Authors: CREMONESI, Matteo (Fermi National Accelerator Lab. (US)); GUTSCHE, Oliver (Fermi National Accelerator Lab. (US))

Presenter: CREMONESI, Matteo (Fermi National Accelerator Lab. (US))

Session Classification: T6 - Machine learning and physics analysis

Track Classification: Track 6 –Machine learning and physics analysis