

Big Data Technologies and Physics Analysis with Apache Spark (lecture 1)

Tuesday, 5 March 2019 13:30 (1 hour)

The Large Hadron Collider is scheduled to shut down for a 2 years maintenance period since December 2018. However, the already collected data -which are stored in a dedicated custom storage service- between April 2015 and November 2018, exceed 150 PBs in total. To analyse these data, more and more teams at CERN decide to use Big Data Technologies to perform Physics Analysis and "Data Reduction", i.e. produce smaller reusable datasets for frequent access. These technologies show great potential in speeding up the existing procedures.

This lecture will provide an overview of the latest trending big data technologies in the Hadoop and Spark ecosystems with focus on their main architecture characteristics, and then will target a number of important questions: How can we perform Physics Analysis with Big Data Technologies? What are the problems faced? What are the challenges and the available data sources? What are the other domain in which Big Data Analytics are applied at CERN?

Presenter: MOTESNITSALIS, Evangelos (CERN)

Track Classification: Lectures and exercises