Contribution ID: **186** Type: **Sectional reports**

Data analysis platform for stream and batch data processing on hybrid computing resources

Tuesday, 6 July 2021 14:00 (15 minutes)

The modern Big Data ecosystem provides tools to build a flexible platform for processing data streams and batch datasets. Supporting both the functioning of modern giant particle physics experiments and the services necessary for the work of many individual physics researchers generate and transfer large quantities of semi-structured data. Thus, it is promising to apply cutting-edge technologies to study these data flows and make the services ' provisioning more effective.
In this work, we describe the structure and implementation of our data analysis platform, built around an Apache Spark cluster. With the official support for GPU computing now available in Spark version 3, we propose a change in architecture to utilize these more performant resources while keeping the platform's functionality provided by using mainstream Big Data software. Furthermore, wanting GPU support necessitated a change of computing resource management infrastructure from Apache Mesos to Kubernetes. Finally, to show the features and operation of the system, we used the task of network packet analysis for security monitoring and anomaly detection in both batch and stream mode.

Summary

Primary authors: KADOCHNIKOV, Ivan (JINR, PRUE); BELOV, Sergey (Joint Institute for Nuclear Research, PRUE); KORENKOV, Vladimir (JINR, PRUE); SEMENOV, Roman (JINR, PRUE); ZRELOV, Petr (JINR, PRUE)

Presenter: KADOCHNIKOV, Ivan (JINR, PRUE)

Session Classification: Big data Analytics and Machine learning.

Track Classification: 9. Big data Analytics and Machine learning