Contribution ID: **246** Type: **Sectional**

Architecture of the computing system for experiments with large amount of data streams

Friday, 4 October 2019 10:45 (15 minutes)

The emergence of a new series of experiments in nuclear physics with large amount of data streams requires a review of the general idea of computing. The well-established concept of LHC data processing involves a huge amount of data in which rare events need to be highlighted. Such an approach is determined by the physics of the phenomenon under study with low densities and high energies. New experiments are aimed at a different physics, when the energy is not so high, and the density is much higher. This generates a huge data stream that needs to be processed entirely.

Selecting tools for working with large amounts of data is a separate task for the development team. Not infrequently, the architecture had to be extremely drastically changed because of increased data loads and control of stored data was lost and the collection of statistics became more and more difficult. There is a need for a solution that allows not only to store all sorts of information with the ability to download from different sources but also has a set of tools to analyze the collected information (Big Panda, Informatica, etc.).

A data lake is a concept, an architectural approach to centralized storage that allows you to store all structured and unstructured data with the possibility of unlimited scaling. A data lake can store structured data from relational databases (rows and columns), semi-structured data (CSV, journals, XML, JSON), unstructured data (emails, documents, PDF files) and binary data (images, audio, video). Quite popular is the approach in which incoming data is converted into metadata. This allows you to store data in its original state, without special architecture or the need to know which questions you may need to answer in the future, without the need to structure the data and have various types of analytics - from dashboards and visualizations to big data processing, real-time analytics and machine learning to make the right decisions.

We believe that this technology is well suited as a basis for new experiments computing. As a result of the analysis of existing solutions, the following functional modules were identified that are the most necessary and need to be developed in the universal solution:

• Storage for all data with the ability to create separate storage for hot/cold data, for ever-changing data or to handle fast streaming
• Security module
• Databases for structured data
• The module of tools for working with data (analysis, data engines, dashboards, etc.)
• Machine learning module
• Services for the development of add-ons, modifications and deployment of storage

Primary author: Prof. DEGTYAREV, Alexander (Professor)

Co-author: Prof. BOGDANOV, Alexander (St.Petersburg State University)

Presenter: Prof. DEGTYAREV, Alexander (Professor)

Session Classification: Computing for Large Scale Facilities (LHC, FAIR, NICA, SKA, PIC, XFEL, ELI, etc.)