



Contribution ID: 477

Type: Poster

Setting up technical requirements for an astroparticle data life cycle

Increasing data rates opens up new opportunities for astroparticle physics by improving the precision of data analysis and by deploying advanced analysis techniques that demand relatively large data volumes, e.g. deep learning. One of the ways to increase statistics is to combine data from different experimental setups for joint analysis. Moreover, such data integration provides us with an opportunity to carry out multi-messenger astroparticle studies and to search for hidden patterns within the data.

A data life cycle (DLC), namely the data processing pipeline, that is clearly defined and maximally automated at each step, from receiving data to obtaining final results of the analysis, allows us to facilitate and speed up data processing and to make the calculations more reliable and reproducible at each stage. The German-Russian Astroparticle Data Life Cycle Initiative (GRADLCI) is aimed to develop such DLC for combined analysis of data from the KASCADE-Grande (Karlsruhe, Germany) and Tunka-133 (Tunka valley, Russia) experiments. The important features of astroparticle DLC include scalability for handling large amounts of data, heterogeneous data integration, and exploiting parallel and distributed computing at every possible stage of the data processing. This demands special technical requirements necessary to perform the analysis in the hardware and software environment of a computing cluster.

In the talk we discuss the plans on DLC organization that are being implemented in the GRADLCI. This include accelerating the KASCADE-Grande database by the use of end-to-end indexing; developing a database for TAIGA/Tunka-133; organizing fast search within distributed data with a help of a proxy server and metadata database; developing access infrastructure and interfaces for scientists and general public to interact with data; performing a joint data analysis; configuring DLC for safe handling client analysis requests on the server side; and maximum automation of distributed run-wise simulations. The talk addresses the choice of a distributed data storage system, and of virtualization tools.

Authors: TOKAREVA, Victoria (Karlsruhe Institute of Technology (KIT)); HAUNGS, Andreas; KOSTUNIN, Dmitriy (KIT)

Presenter: TOKAREVA, Victoria (Karlsruhe Institute of Technology (KIT))

Session Classification: Poster Session

Track Classification: Track 1: Computing Technology for Physics Research