

Current status of data center for cosmic rays based on KCDC

GRID-2018, Dubna

Victoria Tokareva, Dmitriy Kostunin, Andreas Haungs

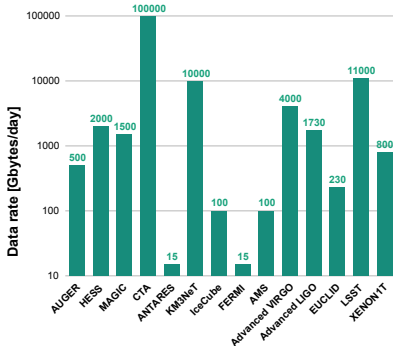
for the German-Russian Astroparticle Data Life Cycle Group | September 12, 2018

INSTITUTE FOR NUCLEAR PHYSICS (IKP)



Introduction:

The astroparticle physics data rate



- Wide range of experiments;
- Looking at the same sky with different eyes: different detectors, different phenomena under the study;
- Common data rate for astrophysical experiments all together is a few PBytes/yearly, which is comparable to the current LHC output*
- Big data for deep learning

Modern astroparticle experiments

data rate [Gbytes/day]*

* APPEC brochure on Computing, 2016

<http://www.appec.org/wp-content/uploads/Documents/Docs-from-old>

German-Russian Astroparticle Data Life Cycle Initiative*



**Matrosov Institute
for System Dynamics
and Control Theory**

KASCADE - Grande
Karlsruhe Shower Core and Array DEtector - Grande

* Granted by RSF-Helmholtz Joint Research Groups

- Proposed in 1989—disassembled in 2013;
- Aimed at studying high-energy (galactic) cosmic rays by observing extensive air showers (EAS);
- Consisted of:
 - scintillators detecting e, γ, μ :
 - KASCADE - 256 stations;
 - GRANDE - 37 stations;
 - Hadronic calorimeter;
 - Radiodetector LOPES detecting e, e^+ ;
- Recognized astrophysical results were obtained. The data analysis is ongoing;
- KCDC (**K**ASCADE **C**osmic Ray **D**ata **C**enter, <http://kcdc.ikp.kit.edu>) is a dedicated portal where all the data collected are available online.



- Started in the mid 90s, is still operating and continuously enhanced;

Tunka-133



- 133 photomultipliers
- measures EAS Cherenkov light

Tunka-Rex



- 63 antennas
- measures EAS radio-emission

Tunka-HiSCORE



- 47 photomultipliers
- measures EAS Cherenkov light

Tunka-Grande



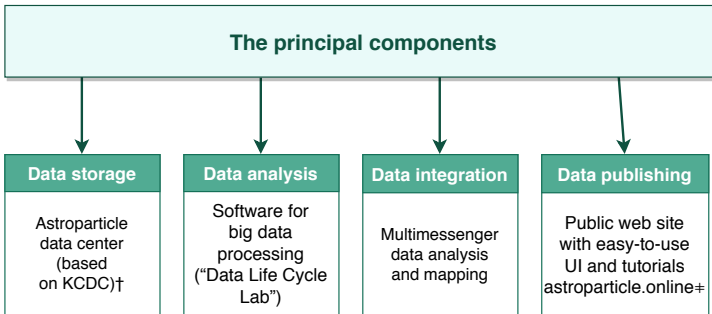
- 380 scintillators 0.64m² each
- measures e/μ from EAS

Tunka-IACT



- Imaging Air Cherenkov Telescopes
- is being extended

The project objectives



[†]Minh Duc Nguyen, *A distributed data warehouse system for astroparticle physics*, GRID2018 session 10

[‡]Yu. Kazarina, *Application of Hubzero platform for the educational process in astroparticle physics*, GRID2018 poster

Deep into KASCADE-Grande and Tunka data formats

Different

- Data format (depends on available detectors)
- Dedicated software for analyzing data
- Special system environment for the software

Common

- Metadata format (e.g. time, location, atmospheric conditions)
- Software for EAS simulation (e.g. CORSIKA)
- Shower parameters
- Theoretical models

Current state

- Separate APIs and UIs for different experiments

Our objective

- Unified API and UIs for different experiments

- The basic idea is to provide a central queue for all users and make all the distributed sites look like local ones;
- Starting from mid 90's are widely used in collider experiments (AliEn, Dirac, PanDA);
- Dedicated for:
 - Unified usage of the distributed remote data and common data analysis;
 - Conceal various low-level software and provide unified high-level interface;
- Provide the common way to issue tasks to different types of the distributed sites;
- The same system for the data access, analysis and simulation.

What data do we work with?

- Data types:

- Raw detector readouts;
- Pre-analyzed events;
- Metadata

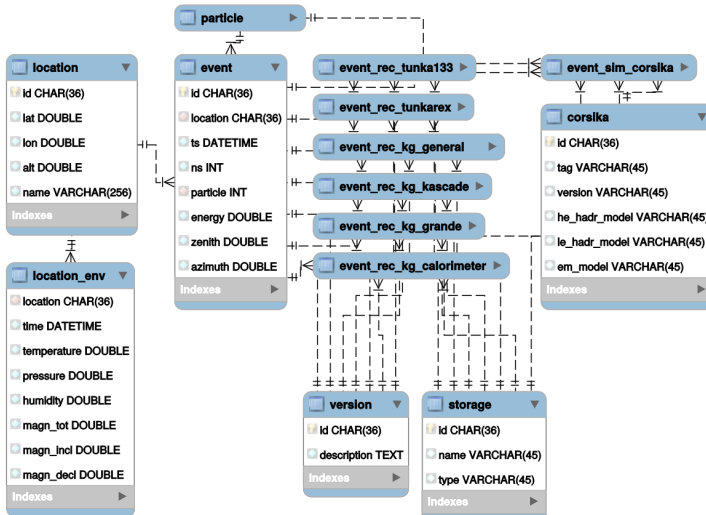
- Data structure:

- Different formats;
- Different messengers;
- Common metadata

Our approach:

- It is proposed to store unique event id and metadata in the unified database
- With growing data sizes, distributed storage for events could be useful

Proposed cosmic-ray metadata structure



- Software for data analysis depends on a particular experiment
 - Problem: It may even require dedicated system environment
 - ▶ Solution: Virtualization could be useful
- Data analysis requires huge amounts of input data
 - Problem: It is often more optimal to perform it on the same site the data are stored
 - ▶ Solution: Job management could handle the task

Feature

Consequence

The software for EAS simulation (e.g. CORSIKA) does not depend on a particular experiment

⇒

Simulations require standardized system environment

Simulations require small amounts of input data

Simulations can be done independently for different events

⇒

Simulations are easily scalable

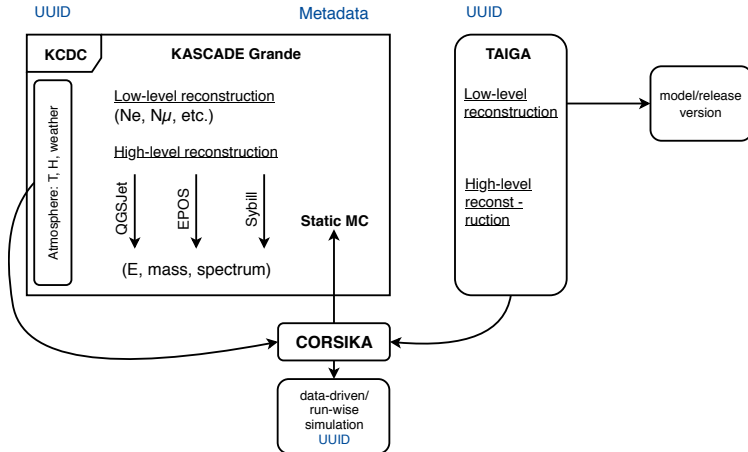
Simulations require a lot of computing resources

⇒

HPC sites are needed

Distributed computing could be useful

Distributed analysis and simulation scheme



- The KASCADE project has a data center called KCDC, that is planned to serve as the basis for the future common center for data access;
- The differences in the data formats were analyzed and solutions for organizing storage and distributed data processing were proposed;
- A scheme of a relational database for the future data center is designed using a metadata-based approach;
- The possibilities to apply the results of the project to educational and outreach activities are being explored.

The joint resource [astroparticle.online](https://www.astroparticle.online) is created to provide access to KASCADE and TAIGA data and metadata.

- The constantly growing amount of accumulated astroparticle data and the request for the multi-messenger astronomy and machine learning, enable us to develop a unified system for astroparticle data storage and processing;
- KASCADE is the only astroparticle experiment so far that has fully published its data and has a software infrastructure for data access and online analysis (KCDC);
- The peculiarities of data format and acquisition make it impossible to utilize 'from scratch' the solutions widely used in collider experiments;
- We are developing a new approach to the astroparticle data life cycle for combined analysis of the KASCADE and TAIGA data;
- The built-up infrastructure will be used to analyze combined data sets with large statistics, allowing to study galactic sources of high-energy γ -rays, which could be a notable step forward in multi-messenger astroparticle physics.

The German-Russian Astroparticle Data Life Cycle collaboration

- **TAIGA** - Tunka Advanced Instrument for cosmic ray physics and Gamma Astronomy (see taiga-experiment.info);
- KASCADE-Grande – Karlsruhe Shower Core and Array Detector - Grande (see www-ik.fzk.de/KASCADE_home.html);
- KIT-IKP - Institute for Nuclear Physics Karlsruhe Institute of Technology
- SCC - Steinbuch Centre for Computing Karlsruhe Institute of Technology
- SINP MSU - Skobeltsyn Institute Of Nuclear Physics Lomonosov Moscow State University
- ISU – Irkutsk State University
- ISDCT – Matrosov Institute for System Dynamics and Control Theory

The German-Russian Astroparticle Data Life Cycle Initiative



- 133 photomultipliers
- measures EAS
Cherenkov light