Karlsruhe Institute of Technology

Open data and machine learning in German-Russian Data Life Cycle initiative

Big Data Science in Astroparticle Research, Aachen University

Victoria Tokareva for GRADLCI | 17-19 February 2020

INSTITUTE FOR NUCLEAR PHYSICS (IKP)

German-Russian Astroparticle Data Life Cycle Initiative*

The international initiative aiming at automatisation the maintenance of astroparticle-physics data throughout their entire life cycle.

НИИЯФ
МГУ

TAIGA

KIT
Karlsruher Institut für Technologie

IKP
Institut für Kernphysik

ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
1918

SCC
Steinbuch Centre
for Computing

Matrosov Institute
for System Dynamics
and Control Theory

KASCADE - Grande
KArlsruhe Shower Core and Array DEtector - Grande

*Granted by RSF-Helmholtz Joint Research Groups

Features

- The system of aggregated data selection and retrieval;
- Flexible horizontal expansion allowing connection of heterogeneous storages of astroparticle data;
- Usage of modern virtualization and machine-learning technologies;
- Online data analysis capability;
- Access to scientific data for the general public.

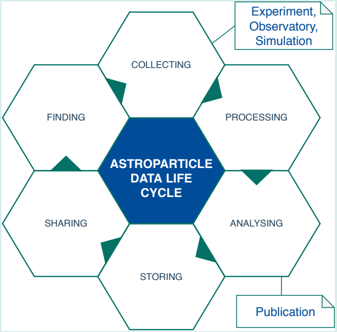
Astroparticle Data Life Cycle

Data Life Cycle (DLC)

The sequence of stages that a particular unit of data goes through from its initial generation or capture to its eventual archival and/or deletion.

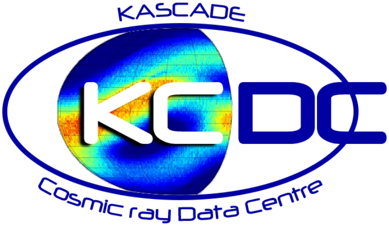
Features of DLC in APP

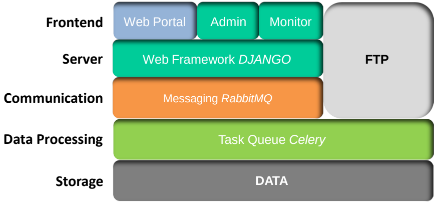
- Constantly growing precision and data amounts;
- Rare events and low statistics;
- Call for multi-messenger astrophysics;
- Need for various data in analysis;
- Data mining in astroparticle data;
- Need for advanced storage architectures and smart data selection queries.



KASCADE Cosmic-ray Data Center (KCDC)

- providing free, unlimited, reliable open access to KASCADE cosmic ray data at https://kcdc.ikp.kit.edu;
- almost all KASCADE data is available;
- selection of fully calibrated quantities and detector signals;
- information platform: physics and experiment backgrounds, tutorials, meta information for data analysis;
- archive of KASCADE software and data;
- uses modern and open source web technologies.

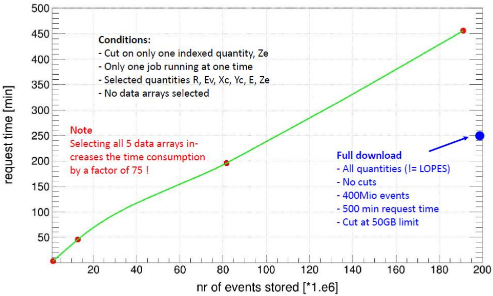




KCDC upgrade

- KCDC web portal has been updated to release OCEANUS 1.0 (November 25th 2019)
- Add data of radio detector component LOPES

- More simulation data sets including gamma simulations
- Most of the software packages and the KCDC-Manuals were updated
- Event were indexed with uuids
- Processing speed increased by a factor of 10 to 50



Conditions:
- Cut on only one indexed quantity, Ze
- Only one job running at one time
- Selected quantities R, Ev, Xc, Yc, E, Ze
- No data arrays selected

Note
Selecting all 5 data arrays increases the time consumption by a factor of 75 !

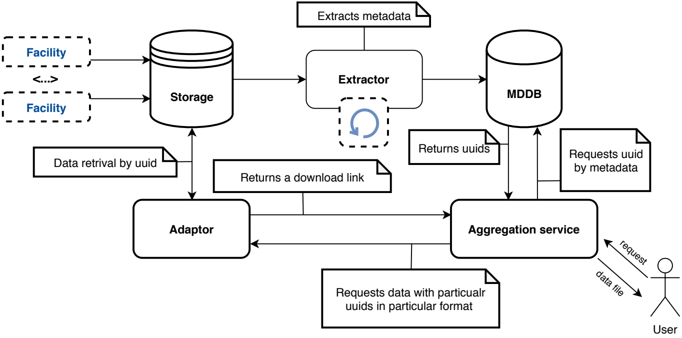
Full download
- All quantities (!= LOPES)
- No cuts
- 400Mio events
- 500 min request time
- Cut at 50GB limit

Data aggregation

Metadata definition

We introduce 2 level metadata model:

1. physical level metadata: file size, file type, last changed, etc.
2. event level: event_id, datetime, setup, atmosphere, etc.



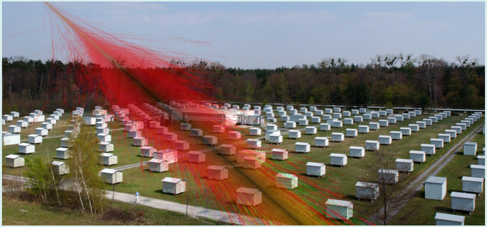
Detectors / Data storages

- KASCADE
- 252 scintillators
- 450 000 000 events

- KASCADE-GRANDE
- 37 scintillators
- 35 310 393 events

- LOPES
- radio antennas array
- 3 058 events

- MongoDB
- a total data volume is ≈ 20 TB



Detectors / Data storages

Tunka-133

- 133 photomultipliers
- MySQL
- 7 421 630 events +
- 0.5 GB

Tunka-Rex

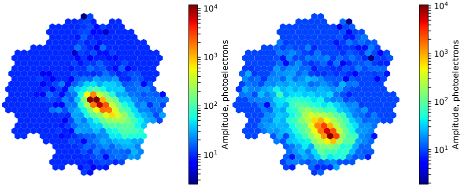
- 63 radio antennas
- MySQL
- 107 360 524 events +
- ≈ 3 TB +

TAIGA-IACT

- 2 Imaging Air Cherenkov Telescopes (being extended)
- binaries
- 2 700 000 000 events +
- 605 GB +

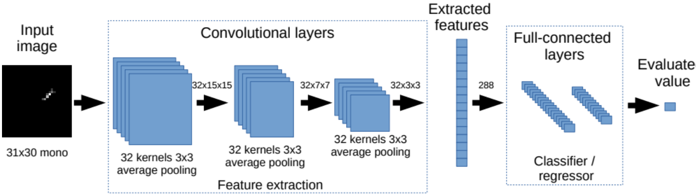
Gamma-hadron separation at TAIGA-IACT

- 3 Convolutional layers
- 2 Full Connected layers
- Activation ReLU
- Dropout 25% and 50%
- Output : sigmoid
- 150 epoch training
- Trained on GPU NVIDIA Tesla P100
- Overfit Hillas ≈ 2 times
- ROC AUC is 0.9647



Examples of the TAIGA-IACT simulation images: gamma-ray (left) and proton (right)

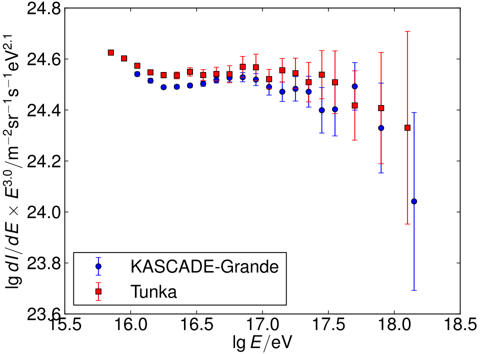
Available at astroparticle.online



CNN for classification (regression).

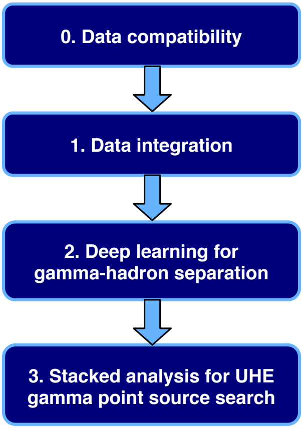
I.Bychkov et all., Russian-German Astroparticle Data Life Cycle Initiative, International Journal on Data Science and Technology, Vol.5, No.2, 2019

Joined analysis of KASCADE and Tunka-133



Energy spectra of cosmic rays from KASCADE-Grande and Tunka-133: normalized flux per energy.

Analysis pipeline



0. Data compatibility

1. Data integration

2. Deep learning for gamma-hadron separation

3. Stacked analysis for UHE gamma point source search

W.D. Apel et al., Tunka-Rex and LOPES Collaborations, Phys. Lett. B 763, 2016, 179

Tunka-Rex Virtual Observatory: Structure

Data Layers (DL)

- `DL0`: raw traces recorded by the ADCs
- `DL1`: traces containing voltages at the antenna stations
- `DL2`: traces containing values of electrical field at the antenna stations
  ⇒ `DL2-AIRSHOWER`, `DL2-ASTRONOMY`, `DL2-OTHER`
- `DL3+` will contain high-level reconstruction of radio data

Antenna station data

- `Trace ID`
- `Antenna ID`
- `Timestamp`
- `Version`
- `Traces`

Calibration data

- `Commission`
- `Decommission`
- `Antenna ID`
- `LNA ID`
- `Filter ID`

Air-shower data

- `UUID`
- `Timestamp`
- `Theta, Phi`
- `X, Y, Z`
- `Energy`

Tunka-Rex Virtual Observatory: Status

Application

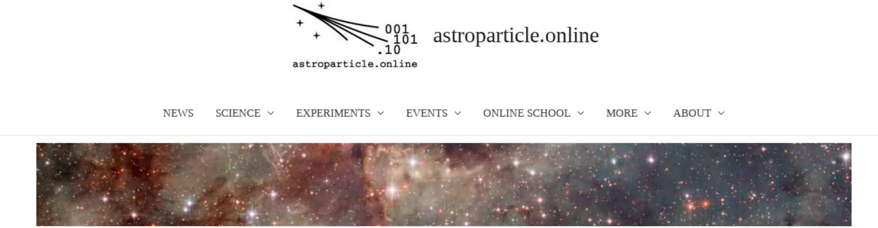
- Studies of the radio background in the frequency band of 30-80 MHz
- Searching for radio transients
- Training of neural networks for RFI tagging
- Outreach and education

Implementation & performance

- 3 TB MySQL database with 100M events (DL1) deployed at IKP KIT
- Processing of 1k events/s
- Almarac (Tien-Shan radio array) DB is deployed at API ISU
- Integration with GRADLCI services

Open access and education

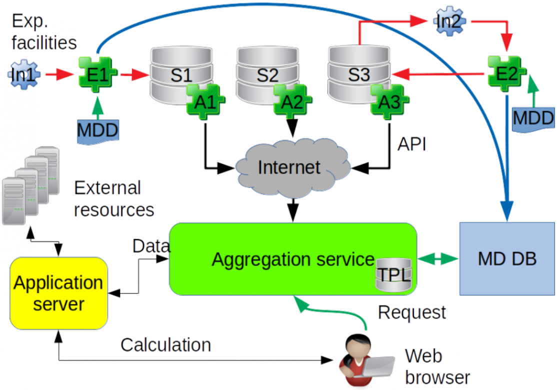
- In web: astroparticle.online
- Outreach: lectures, exercises, quizzes, etc.
- News and events
- Online neural-network analysis (alpha version)
- Aggregated data search (alpha version)



Outlook

- GRADLCI is the international initiative aiming at automatisation the maintenance of astroparticle-physics data throughout their entire life cycle
- The key components: metadata management, data aggregation, data analysis (employing deep learning), go for public
- KCDC upgrade: enlarged dataset, uuids, more simulations, processing speed increased up to a factor of 50
- 2 level metadata management model was introduced
- The system of data aggregation was developed and being tested
- The storages attached: KASCADE, KASCADE-Grande, LOPES, Tunka-133, TAIGA-IACT, Tunka-Rex
- Deep learning is being used for primary gammas identification on TAIGA-IACT and KASCADE+Tunka-133 datasets

DLC Architecture



- **Si** — local data storages;
- **Ini** — data sources of different types;
- **MDD** — metadata description;
- **Ei** — metadata extractors;
- **Ai** — adapters, provide API for data access;
- **TPL** — template library;
- **MD DB** — metadata database.