Karlsruhe Institute of Technology

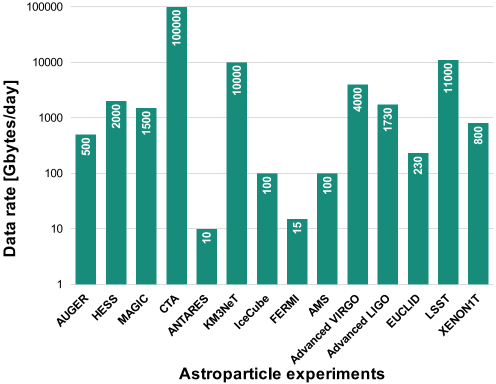
Data retrieval and analysis opportunities in GRADLCI

DLC-19, Irkutsk

Victoria Tokareva

INSTITUTE FOR NUCLEAR PHYSICS (IKP)

Big data in astroparticle physics (APP)



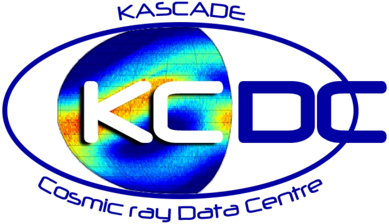
Modern astroparticle experiments data rate [GBytes/day]*

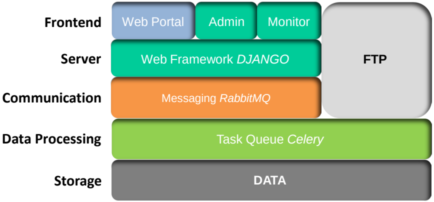
- Wide range of experiments;
- More than hundred years of cosmic particle measurements;
- Looking at the same sky with different detectors;
- Common data rate for astroparticle physics experiments all together is a few PBytes/year, which is comparable to the current LHC output*
- Big data for deep learning

*Berghöfer T., Agrafioti I. et al. Towards a model for computing in European astroparticle physics, Astroparticle Physics European Coordination committee, 2016

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.





Frontend: Web Portal | Admin | Monitor

Server: Web Framework *DJANGO*

Communication: Messaging *RabbitMQ*

FTP

Data Processing: Task Queue *Celery*

Storage: DATA

KASCADE and TAIGA data rates

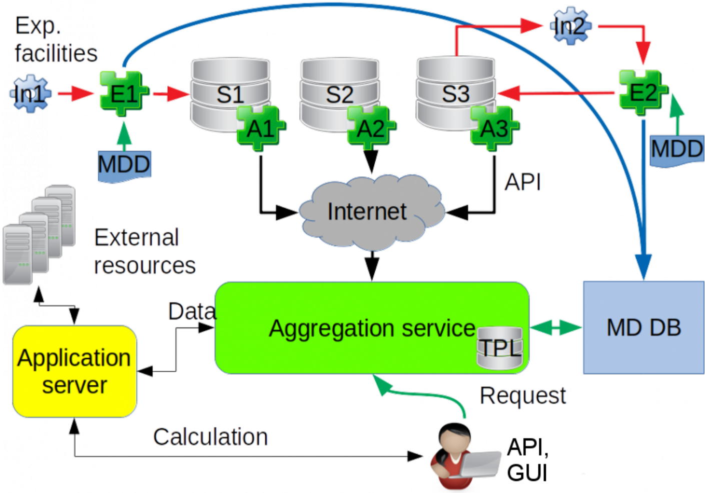
- KASCADE:
  - 450 000 000 events
  - $\sim$ 4 TB of measured data
- planned TAIGA rate: $\sim$ 20 TB/year
  - HiSCORE: $\sim$ 18 TB/year
  - IACT: $\sim$ 1.5 TB/year
  - others: $\sim$ 0.5 TB/year

- current TAIGA rate:
  - $\sim$ 50 TB of raw data;
  - $\sim$ 8 TB/year of reconstructed data:
    - HiSCORE: $\sim$ 6.4 TB/year
    - IACT: $\sim$ 1 TB/year
    - others: $\sim$ 0.5 TB/year

What would we like to get?

- Employing already developed environment and instruments: KCDC update;
- Open access to high-level data;
- Support of data analysis in APP: cut-supporting interface;
- MetadataDB includes information about the events to make basic cuts and decrease the amount of data user has to download;
- Integration of TAIGA data access into existing environment.

DLC Architecture



Aggregation server concept in GRADLC

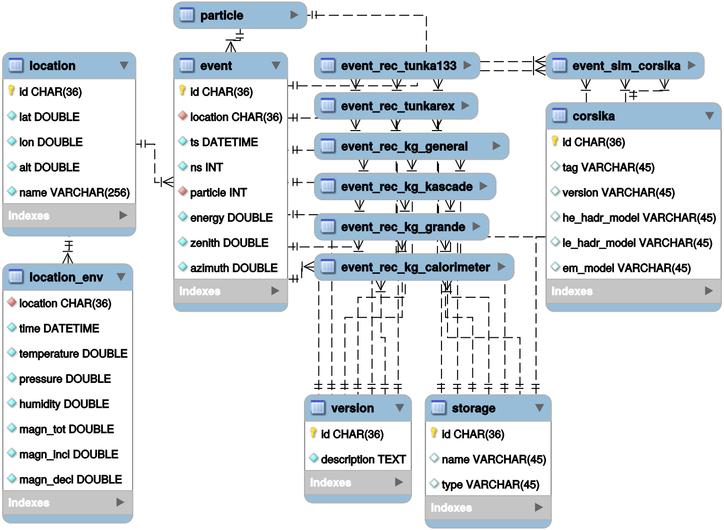
Aim: to promote fast reliable access for accumulated data.

Possible solutions: CVMFS, PostgreSQL (or TimeScale - ?), TPL.

Features:

- Data caching
- Fast metadata DB search

Medatata database



particle

location
- id CHAR(36)
- lat DOUBLE
- lon DOUBLE
- alt DOUBLE
- name VARCHAR(256)
- Indexes

location_env
- location CHAR(36)
- time DATETIME
- temperature DOUBLE
- pressure DOUBLE
- humidity DOUBLE
- magn_tot DOUBLE
- magn_incl DOUBLE
- magn_decl DOUBLE
- Indexes

event
- id CHAR(36)
- location CHAR(36)
- ts DATETIME
- ns INT
- particle INT
- energy DOUBLE
- zenith DOUBLE
- azimuth DOUBLE
- Indexes

event_rec_tunka133

event_rec_tunkarex

event_rec_kg_general

event_rec_kg_kascade

event_rec_kg_grande

event_rec_kg_calorimeter

event_sim_corsika

corsika
- id CHAR(36)
- tag VARCHAR(45)
- version VARCHAR(45)
- he_hadr_model VARCHAR(45)
- le_hadr_model VARCHAR(45)
- em_model VARCHAR(45)
- Indexes

version
- id CHAR(36)
- description TEXT
- Indexes

storage
- id CHAR(36)
- name VARCHAR(45)
- type VARCHAR(45)
- Indexes

Application server concept in GRADLC

Aim: to provide user the opportunity to analyze the data selected remotely.

Computing sources:

- CR local network, clusters (GRIDKa, BW-HPC, etc.), external clouds (Exoscale, OpenNebula, Amazon, Google, etc.).
- HTCondor or HTCondor-based workload management systems: VCondor, Panda, Dirac.

Features:

- Data mapping plugins;
- Vispa-like interface.

KASCADE-TAIGA data mapping

The possibility to map Tunka-133 and KASCADE-Grande spectra was shown at: W.D. Apel et al., Tunka-Rex and LOPES Collaborations, Phys. Lett. B 763 (2016) 179
in two ways:

- simulations;
- radio extensions LOPES and Tunka-Rex.

- KASCADE - TAIGA-HiSCORE data mapping - ?
- proton-γ separation: Xmax or - ?

VISPA project

- access an experimental data;
- example analyses;
- user-determined algorithms, popular software libraries are available
- results visualization

Outlook

- Done so far:
  - Metadata DB filled out with sample data of KASCADE and Tunka-133;
  - The filling with the main amount is ongoing;
- ToDo:
  - Metadata extractor for KCDC;
- Open for discussion:
  - 2nd level MD search - the place in the general scheme;
  - user interface;
  - data mapping;