Karlsruhe Institute of Technology

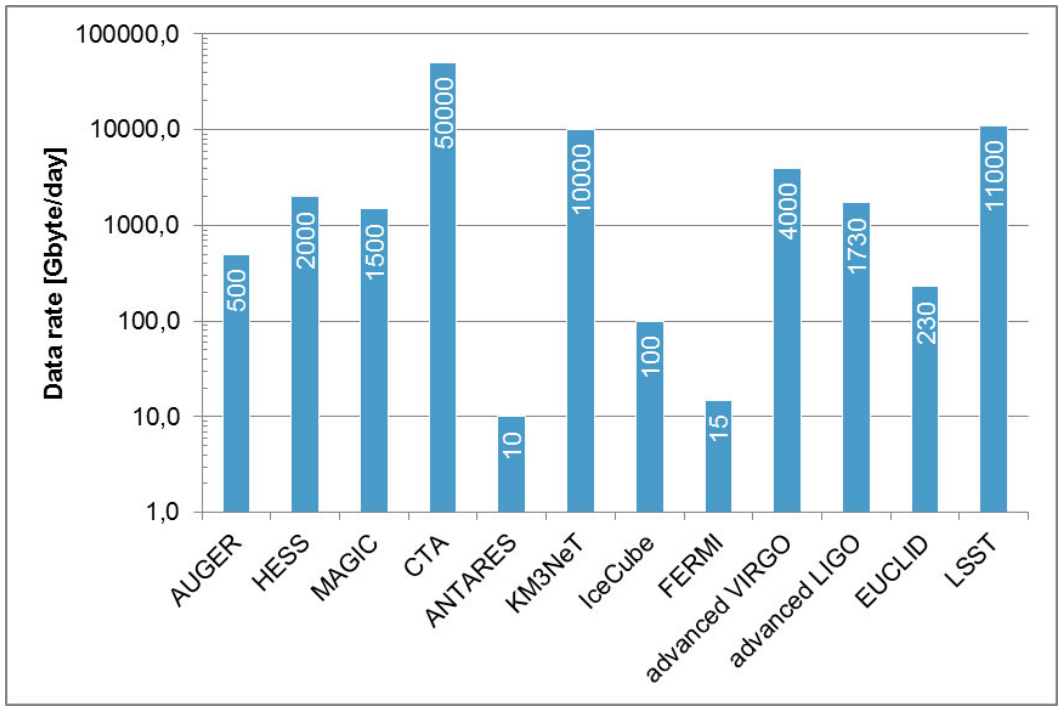
Data engineering for joint analysis of different astroparticle data in GRADLC project

Big Data Science in Astroparticle Research, Aachen University

Victoria Tokareva for GRADLC Consortium | 18-20 February 2019

INSTITUTE FOR NUCLEAR PHYSICS (IKP)

Big data in astroparticle physics (APP)



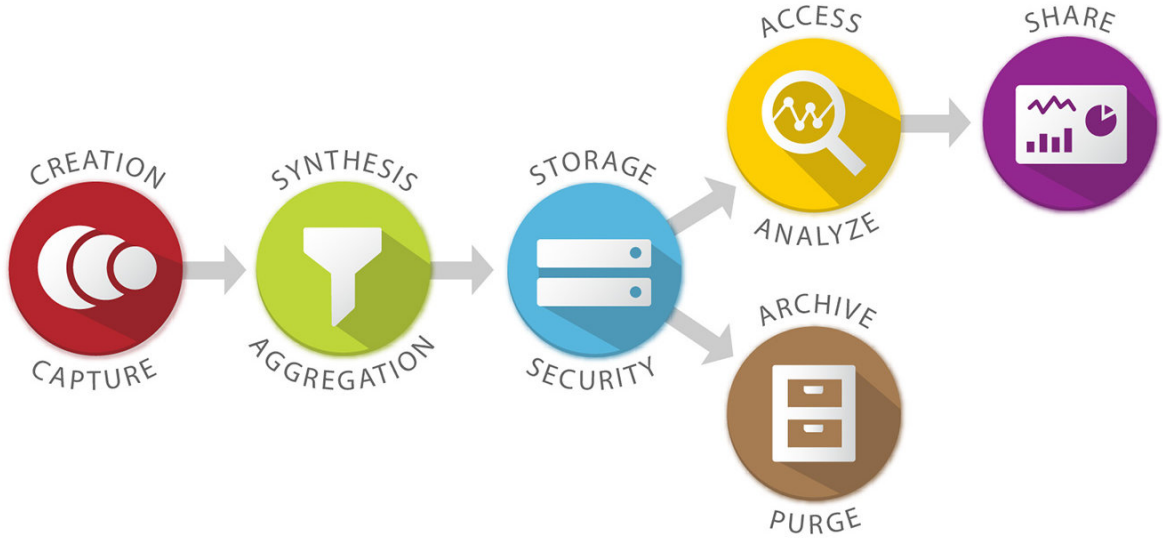
Modern astroparticle experiments data rate [Gbytes/day]*

- 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 all. Towards a model for computing in European astroparticle physics, Astroparticle Physics European Coordination committee, 2016

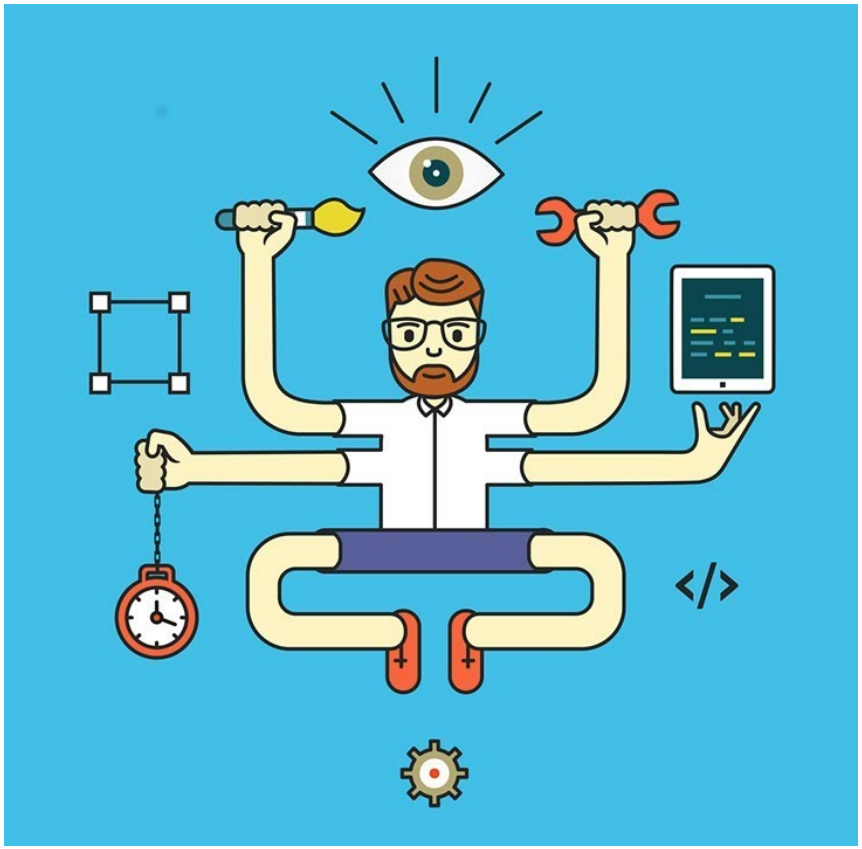
Software data life cycles

- **Data engineering** includes development and support of data architecture and a data pipeline for a platform that enables data analysts, scientists, and other personnel to query the data.
- **Data Life Cycle (DLC)** is 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. The typical data life cycle includes data gathering, processing, storage, analysis, sharing and archiving.

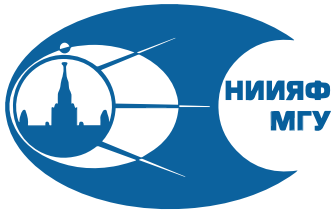


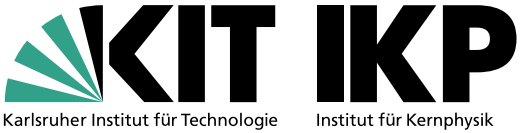
Data engineering in APP

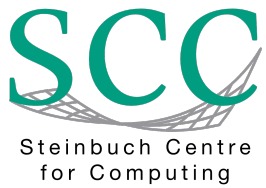
- GRADLC initiative and main objectives
- Features of DLC in APP
- KASCADE Cosmic-ray Data Center
- Proposed DLC architecture
- Data aggregation server: metadata, data summary and workflows
- Application server: conception, computing sources and analysis
- Outlook



German-Russian Astroparticle Data Life Cycle Initiative*







Matrosov Institute
for System Dynamics
and Control Theory

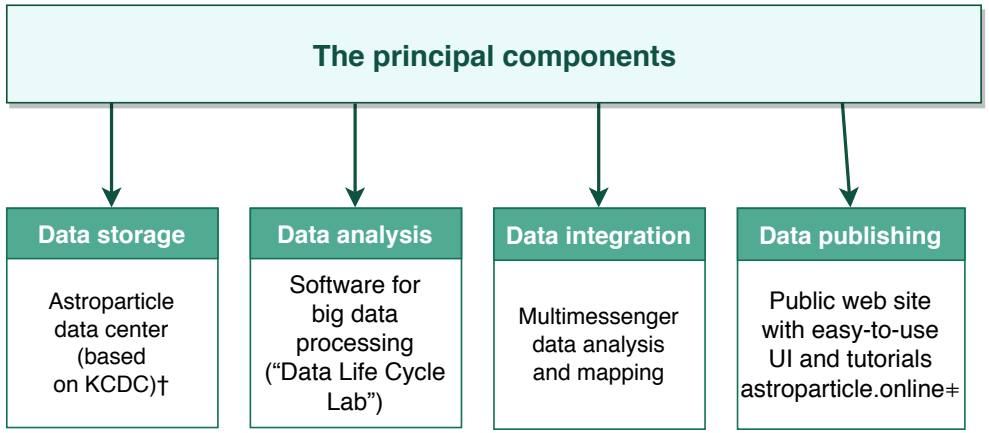
KASCADE - Grande

KArlsruhe Shower Core and Array DEtector - Grande

*Granted by RSF-Helmholtz Joint Research Groups

The main objectives

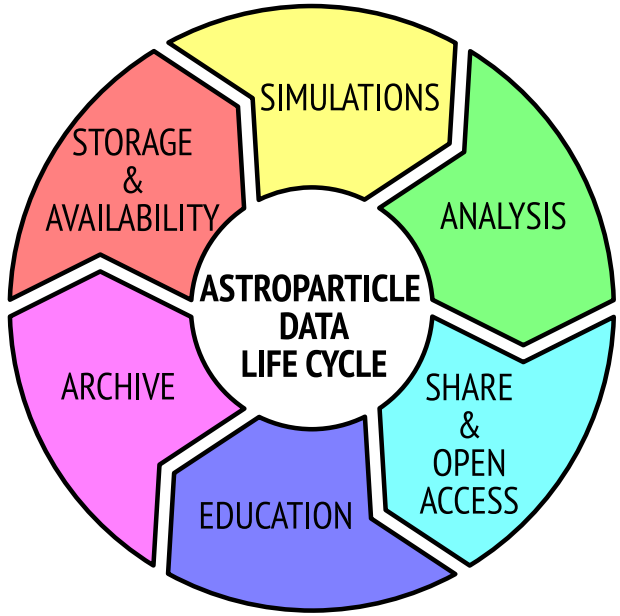
- Provide sustainable access to scientific data
- Archiving of Data and Metadata
- Providing analysis tools
- Education in Big Data Science



- Development area for multi-messenger analyses (e.g. Deep Learning)
- Platform for communication and exchange within Astroparticle Physics

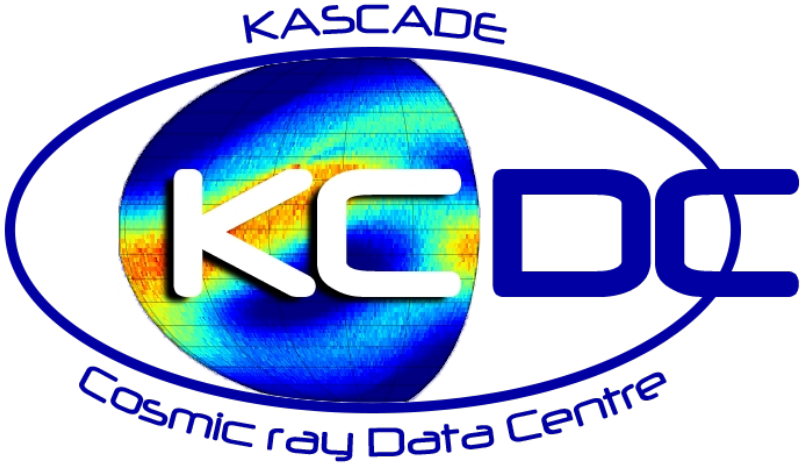
Features of DLC in APP

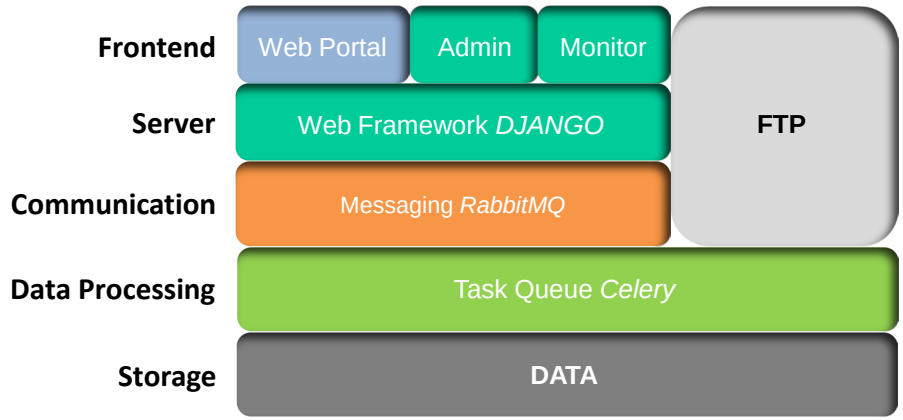
- Constantly growing precision and data amounts;
- Rare events and low statistics;
- Call for multi-messenger astrophysics;
- Need for various data in analysis;
- Data mining in astroparticle data;
- Need for advanced storage architectures and smart data selection queries.



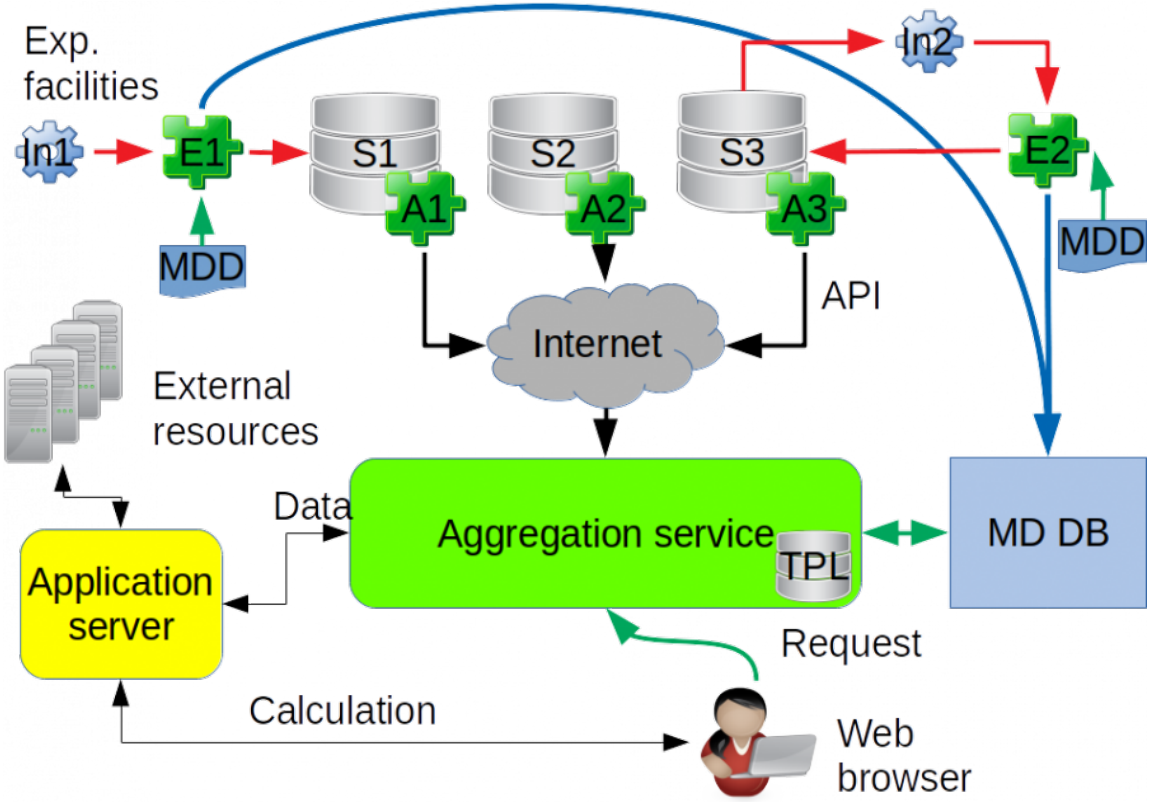
KASCADE Cosmic-ray Data Center (KCDC)

- 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.





DLC Architecture



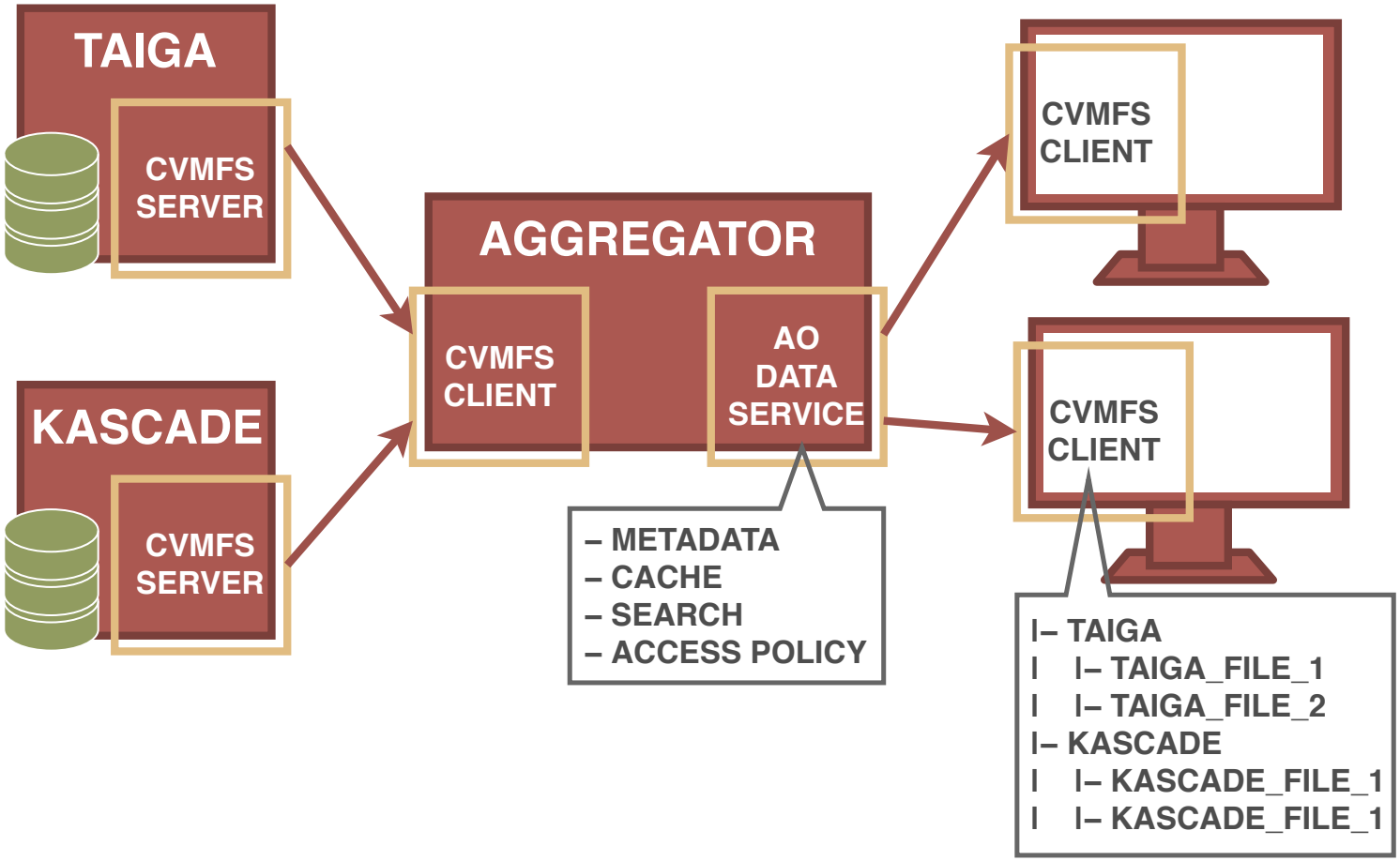
- **Si** — local data storages;
- **Ini** — data sources of different types;
- **MDD** — metadata description;
- **Ei** — metadata extractors;
- **Ai** — adapters, provide API for data access;
- **TPL** — template library;
- **MD DB** — metadata database.

Requirements for the data warehouse

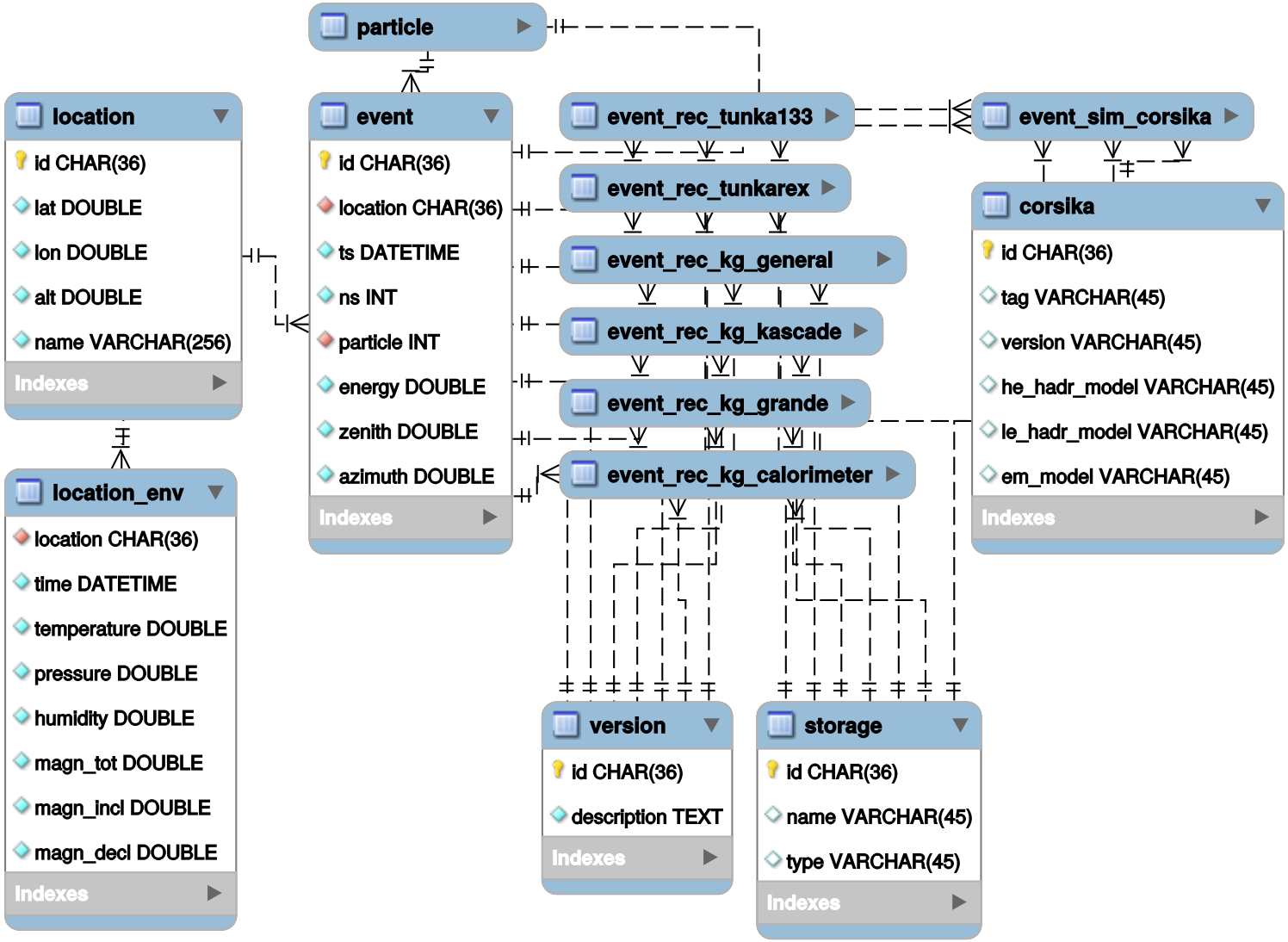
- Multiple experiments (TAIGA, KASCADE, etc.)
- More than hundreds of terabytes of raw data at each site
- Remote access to query results as local file systems
- On-demand data transfer by requests only
- Automatic real-time updates
- No changes to existing site infrastructure, only add-ons

Proposed solution for data aggregation

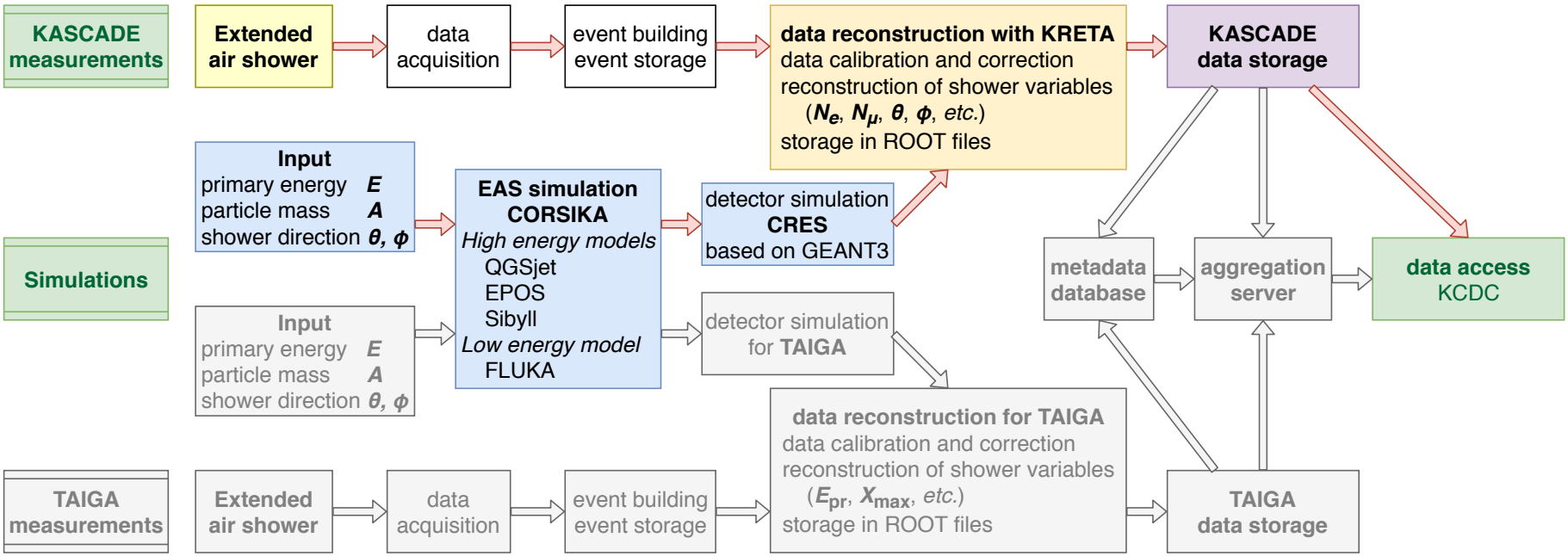
Main solution components: CVMFS, PostgreSQL, TPL



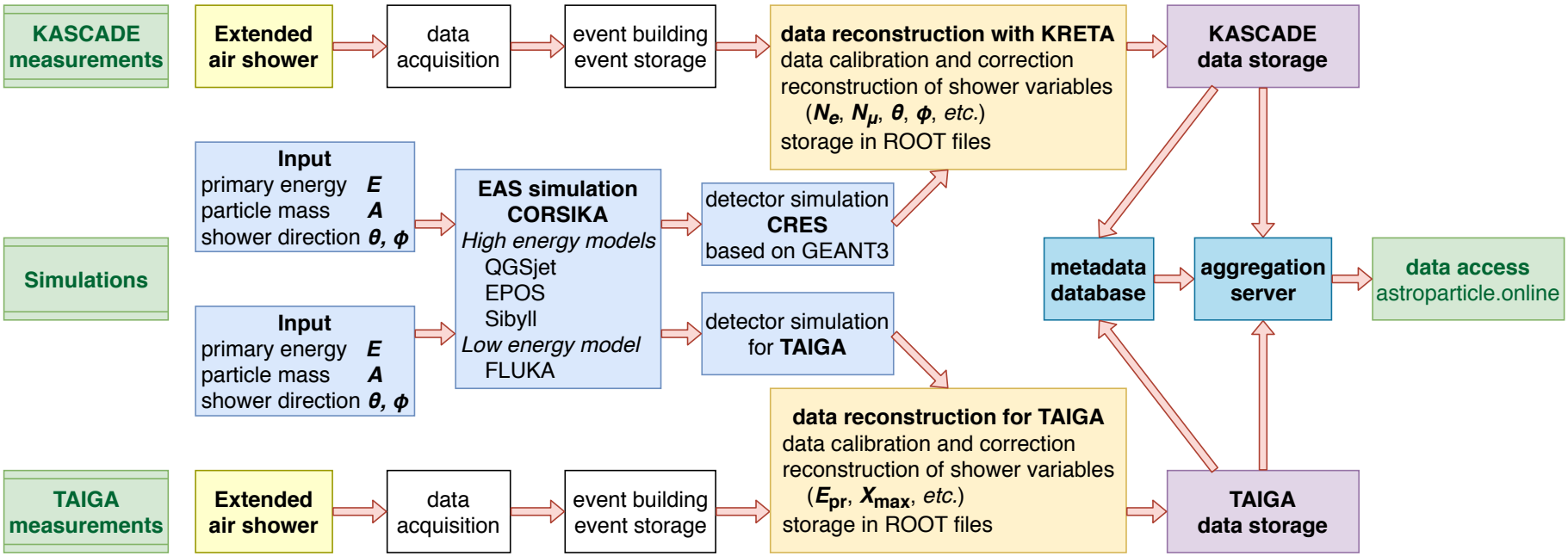
Proposed cosmic-ray metadata structure



Data workflow



Data workflow



Application server concept in GRADLC

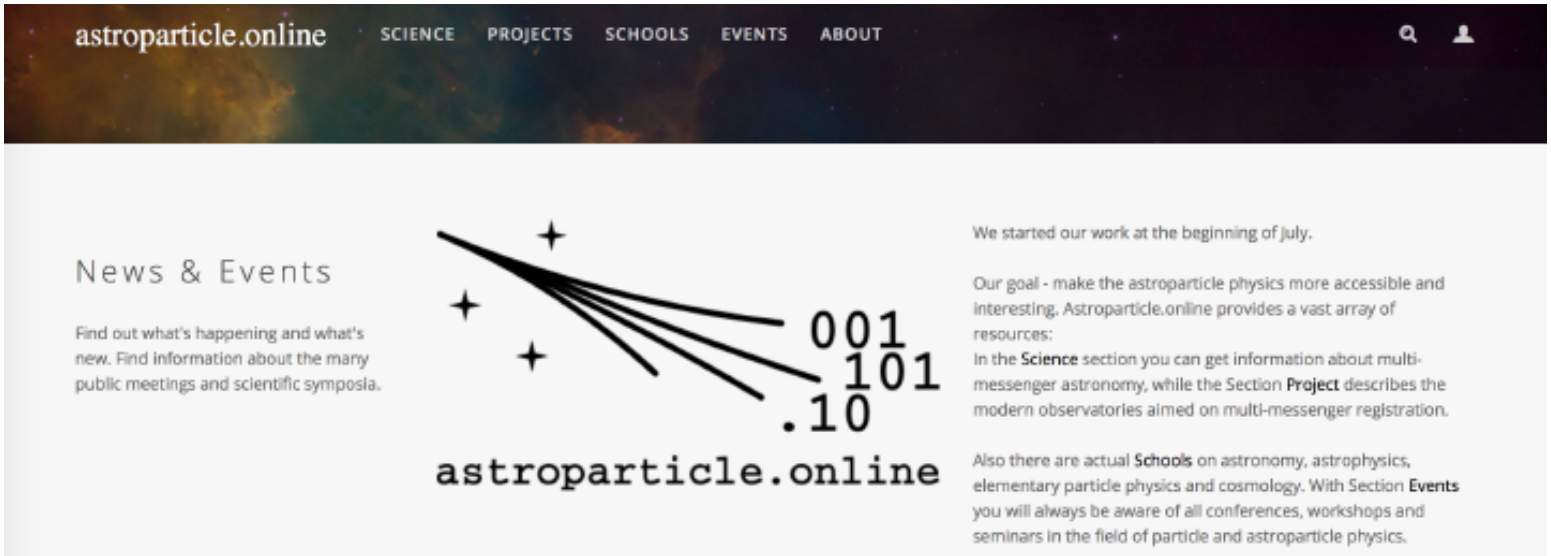
Aim: to provide user the opportunity to analyze the data selected remotely.
Computing sources: CR local network, clusters (GRIDKa, BW-HPC, etc.), external clouds (Exoscale, OpenNebula, Amazon, Google, etc.).
Possible solutions: HTCondor or HTCondor-based workload management systems: VCondor, Panda, Dirac.

Analysis use-cases

- Analysis could be either algorithmic or machine learning;
- Machine learning requires large enough statistics in order to work properly.

Open access and education

- Open access: a dedicated portal planned
- Education: `astroparticle.online`



Outlook

- There are no standard solutions so far for careful sharing and managing large data volumes from various APP experiments, but they are in a high demand;
- The Astroparticle Data Life Cycle data ecosystem is in development by participants of GRADLC Consortium, and based on already existing data portal KCDC and a well-known instruments widely used in particle physics;
- The first stage of development includes building aggregation data server and performing the proof-of-principle joint analysis of data from KASCADE and TAIGA experiments are currently being developed;
- Access to the project resource is now partially provided through the astroparticle.online web-portal;
- Most of the data of KASCADE-Grande are available on KCDC web-portal: https://kcdc.ikp.kit.edu/

The German-Russian Astroparticle Data Life Cycle collaboration I

TAIGA—Tunka Advanced Instrument for cosmic ray physics and Gamma Astronomy (see taiga-experiment.info);

KASCADE - Grande
KArlsruhe Shower Core and Array DEtector - Grande

KASCADE-Grande—KArlsruhe Shower Core and Array DEtector—Grande (see www-ik.fzk.de/KASCADE_home.html);

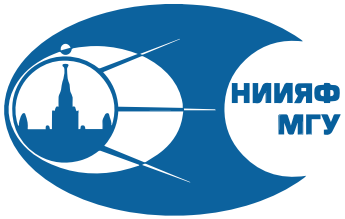
KIT IKP
Karlsruher Institut für Technologie
Institut für Kernphysik

KIT-IKP—Institute for Nuclear Physics Karlsruhe Institute of Technology

SCC
Steinbuch Centre for Computing

SCC—Steinbuch Centre for Computing Karlsruhe Institute of Technology

The German-Russian Astroparticle Data Life Cycle collaboration II



SINP MSU—Skobeltsyn Institute Of Nuclear Physics Lomonosov Moscow State University

ISU—Irkutsk State University

ISDCT—Matrosov Institute for System Dynamics and Control Theory

References

- Berghöfer T., Agrafioti I. *et al.* Towards a model for computing in European astroparticle physics, Astroparticle Physics European Coordination committee, 2016, web-source: `http://appec.org/wp-content/uploads/Documents/Docs-from-old-site/AModelForComputing-2.pdf`;
- KCDC—**K**ASCADE **C**osmic Ray **D**ata **C**enter, web-source: `http://kcdc.ikp.kit.edu`;
- KASCADE-Grande official site, web-source: `http://www-ik.fzk.de/KASCADE_home.html`;
- TAIGA collaboration official site, web-source: `http://taiga-experiment.info`;
- Astroparticle.online—outreach resource, web-source: `http://astroparticle.online`.