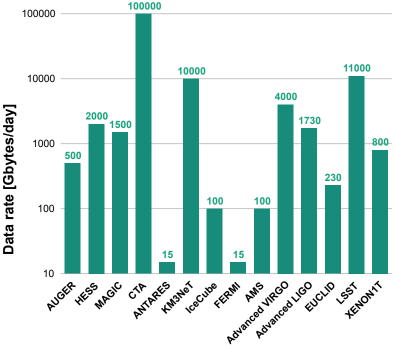
Data integration for various astroparticle data in German-Russian Astroparticle Data Life Cycle project

Astroparticle school 2018, Obertrubach

Victoria Tokareva | October 6, 2018

INSTITUTE FOR NUCLEAR PHYSICS (IKP)

Introduction:
The astroparticle physics data rate

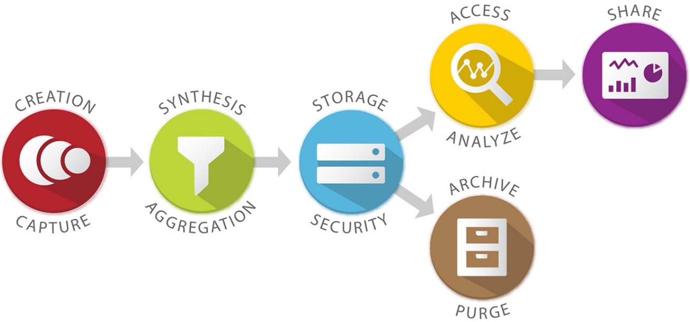


Modern astroparticle experiments data rate [Gbytes/day]*

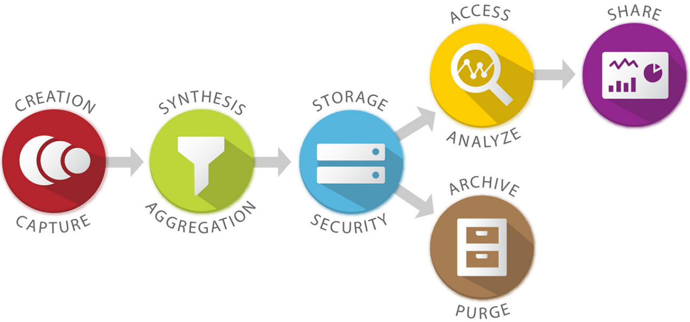
- More than hundred years of cosmic particle measurements;
- Looking at the same sky with different detectors;
- Common data rate for astrophysical experiments all together is a few PBytes/yeary, which is comparable to the current LHC output*
- Big data for deep learning

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

Data life cycle

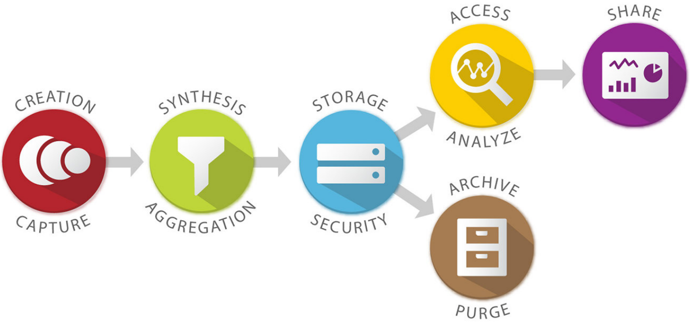


Data life cycle



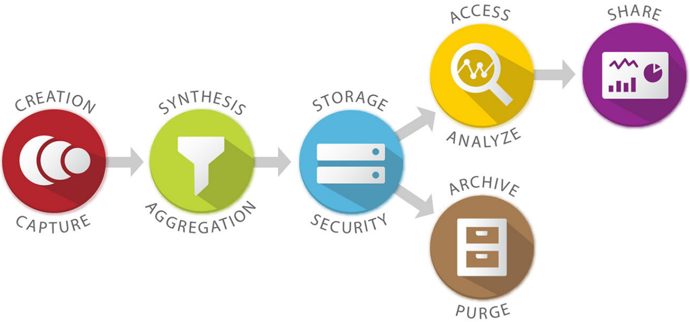
- Experiments improve and are measuring events with greater precision (large amount of data);

Data life cycle



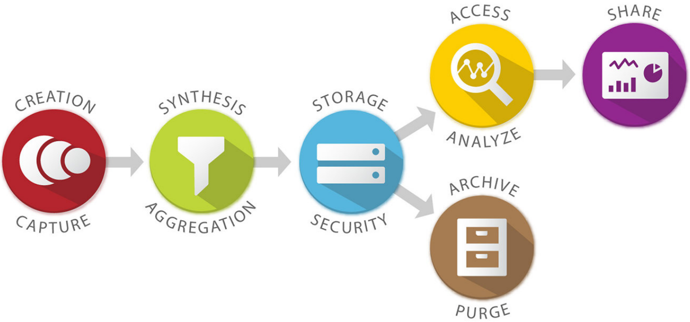
- Experiments improve and are measuring events with greater precision (large amount of data);
- But not too many events of our interest;

Data life cycle



- Experiments improve and are measuring events with greater precision (large amount of data);
- But not too many events of our interest;
  $\Rightarrow$ combined analysis of data from different experiments becomes topical;

Data life cycle



- Experiments improve and are measuring events with greater precision (large amount of data);
- But not too many events of our interest;
  $\Rightarrow$ combined analysis of data from different experiments becomes topical;
- Astronomical Virtual Observatories (Auger & IceCube data).

German-Russian Astroparticle Data Life Cycle Initiative*

KIT Karlsruhe Institute of Technology

НИИЯФ МГУ

TAIGA

KIT Karlsruher Institut für Technologie

IKP Institut für Kernphysik

ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ 1918

RSF Russian Science Foundation

HELMHOLTZ ASSOCIATION

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

KASCADE-Grande

- Proposed in 1989—disassembled in 2013;
- Aimed at studying high-evergy (galactic) cosmic rays by observing extensive air showers (EAS);
- Consisted of:
  - scintillators detecting *e*, γ, μ:
    - KASCADE—256 stations;
    - GRANDE—37 stations;
  - Hadronic callorimeter;
  - Digital radio array LOPES detecting *e*, e^+;
- Important features of cosmic-ray spectrum have been 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.

TAIGA

- Started in the mid 90s, is still operating and continiously enhancend

Tunka-133

- 133 photomultipliers
- measures EAS Cherenkov light

Tunka-Rex

- 63 antennas
- measures EAS radio-emission

Tunka-HiSCORE

- 47×4 photomultipliers
- measures EAS Cherenkov light

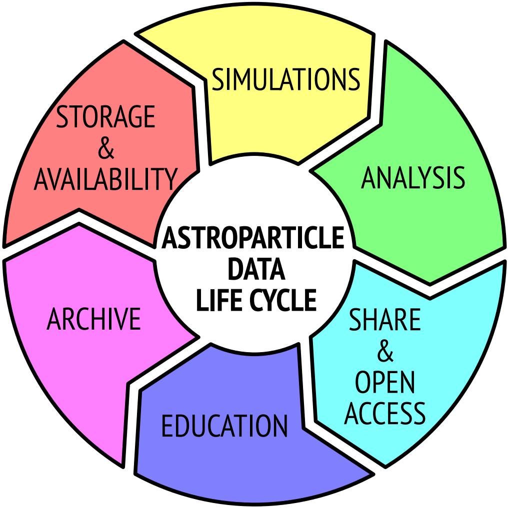
Tunka-Grande

- 380 scintillators $0.64 m^2$ each
- measures e/μ from EAS

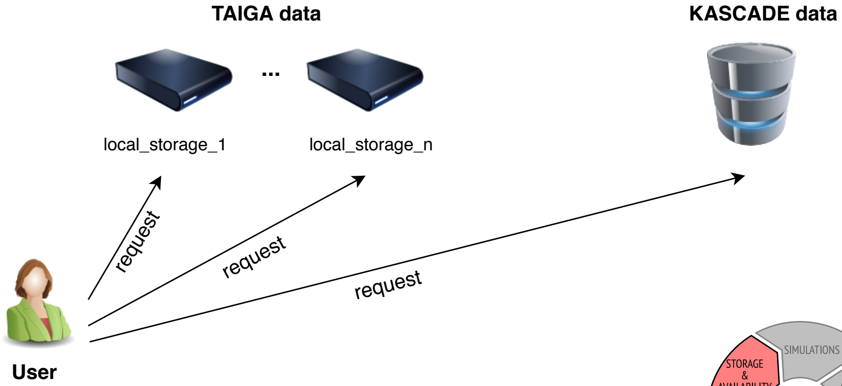
Tunka-IACT

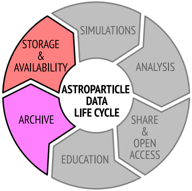
- Imaging Air Cherenkov Telescopes
- is being extended

Data life cycle scheme

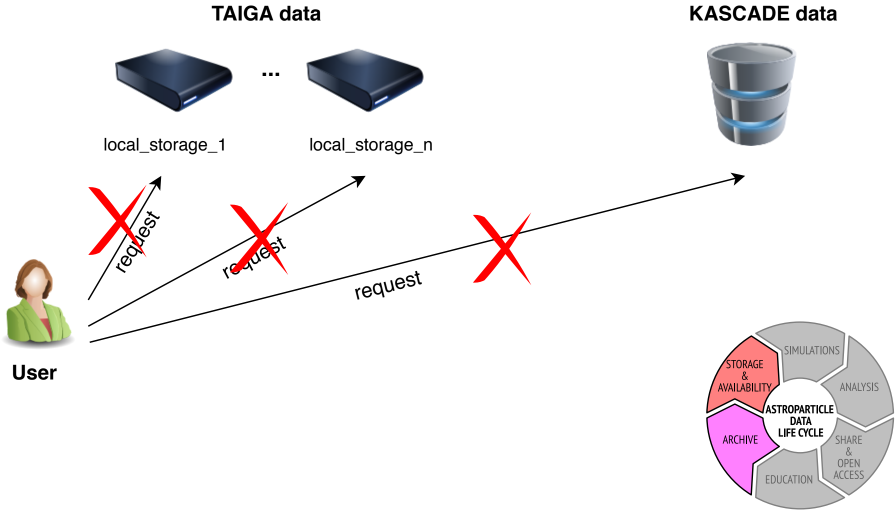


Archiving and storage

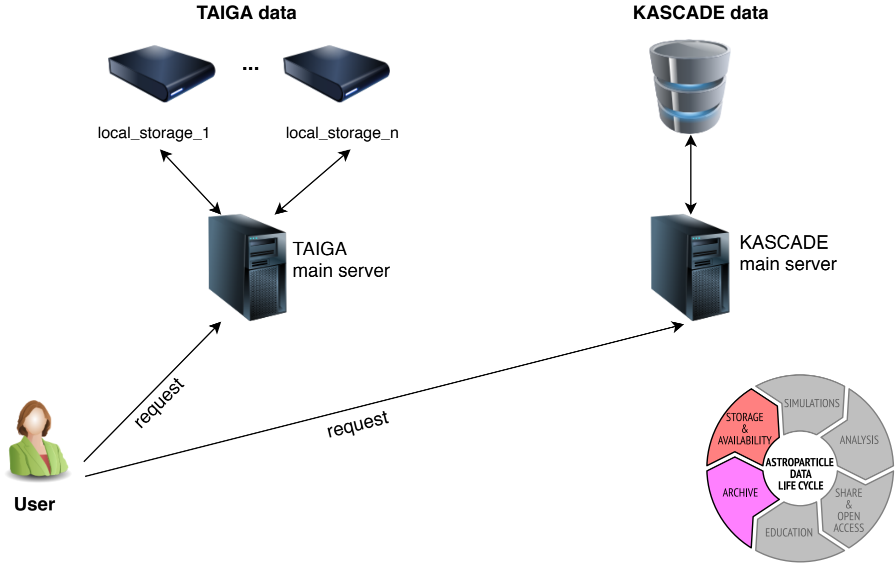




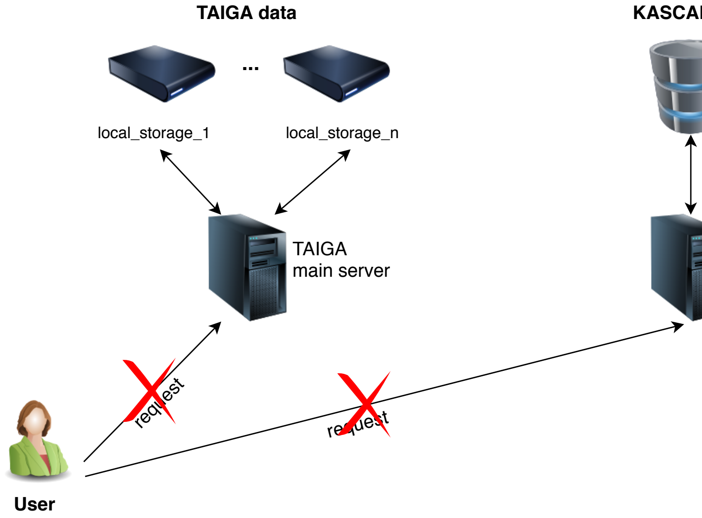
Archiving and storage

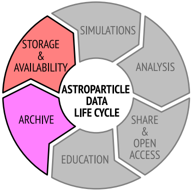


Archiving and storage

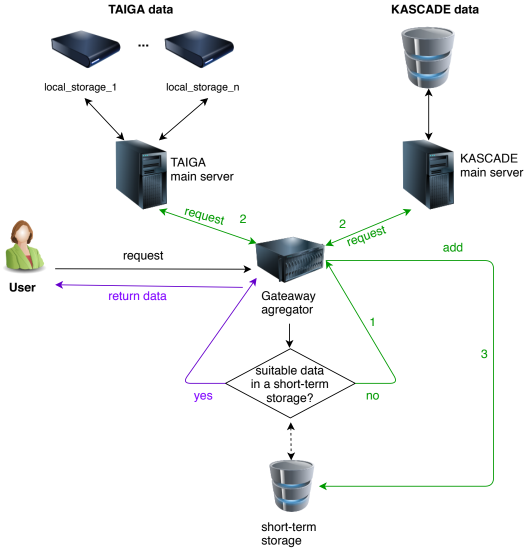


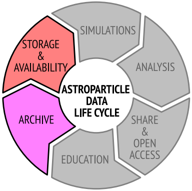
Archiving and storage



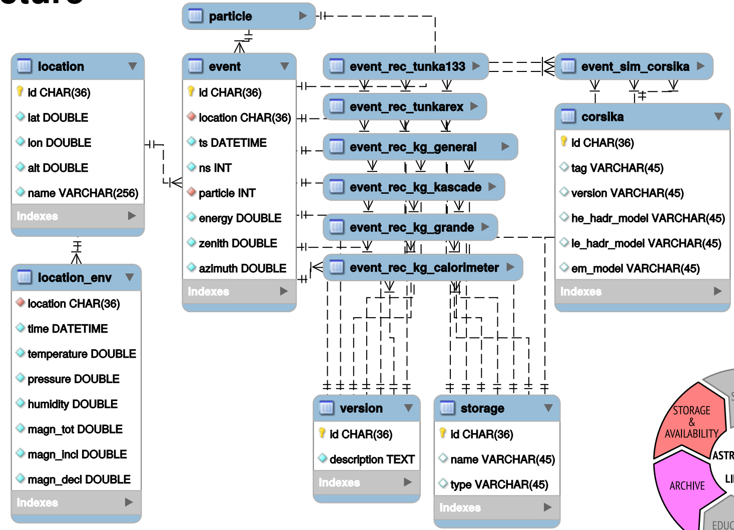


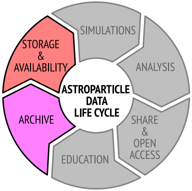
Archiving and storage



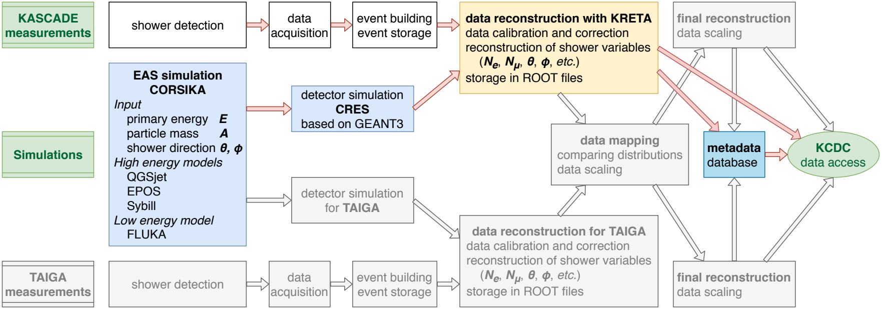


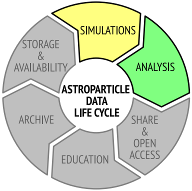
Proposed cosmic-ray metadata structure



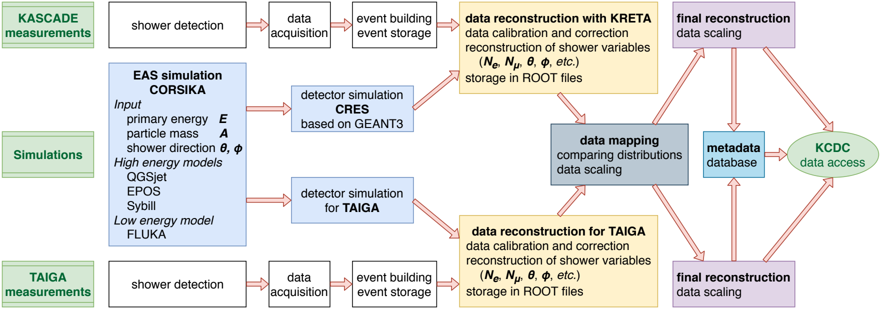


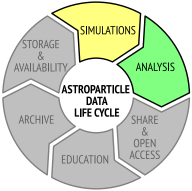
Data workflow





Data workflow

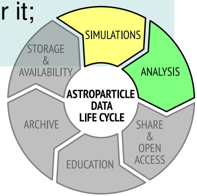




Simulation

Simulation: two steps

1. Simulating EAS:
   - CORSIKA, does not depend on detector features, depends on location ans atmospheric conditions;
   - requires large computing power with a standard environment;
   - a small amount of input data and a large amount of output data;
2. Simulating detector output:
   - depends on detector features;
   - requires dedicated software and special environment for it;
   - large amount of both input and output data;

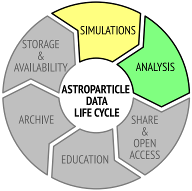


Analysis

Karlsruhe Institute of Technology

Event building → Data calibration and correction → Preliminary reconstruction → Data mapping and scaling → Reconstruction → Joint data analysis

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



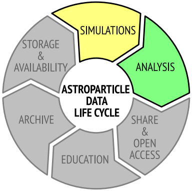
Analysis features

Software for data analysis depends on a particular experiment

- Problem: It may even require dedicated system environment
- **Solution: Virtualization**†

Data analysis requires huge amounts of input data

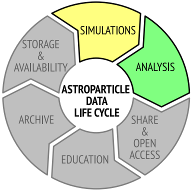
- Problem: It is often more optimal to perform it on the same site the data are stored
- **Solution: Job management**



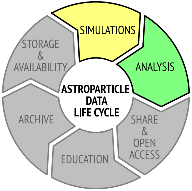
† "The act of creating a virtual (rather than actual) version of something, including virtual computer hardware platforms, storage devices, and computer network resources". © Wikipedia

WMS—workload management system

- 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 (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 distibuted sites;
- The same system for the data access, analysis and simulation.

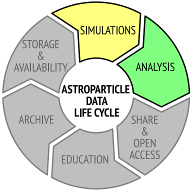


WMS for astroparticle data management



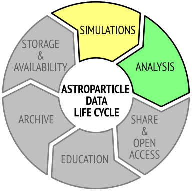
WMS for astroparticle data management

- IceCube ?



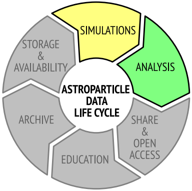
WMS for astroparticle data management

- IceCube ? PanDa



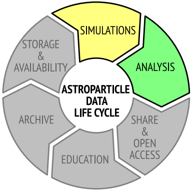
WMS for astroparticle data management

- IceCube ? PanDa
- Auger ?



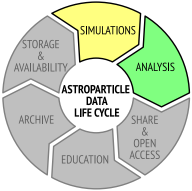
WMS for astroparticle data management

- IceCube ? PanDa
- Auger ? DiRAC



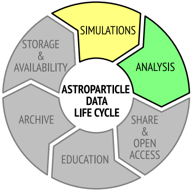
WMS for astroparticle data management

- IceCube ? PanDa
- Auger ? DiRAC
- Other WMS ?



WMS for astroparticle data management

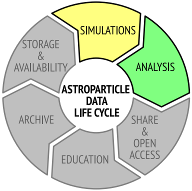
- IceCube ? PanDa
- Auger ? DiRAC
- Other WMS ? VCondor, MyCluster, GWPilot, BigJob, ...



WMS for astroparticle data management

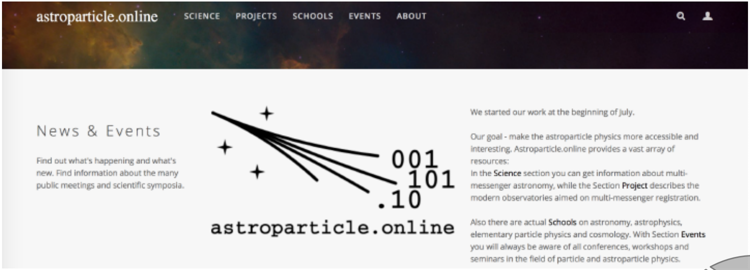
- IceCube ? PanDa
- Auger ? DiRAC
- Other WMS ? VCondor, MyCluster, GWPilot, BigJob, ...

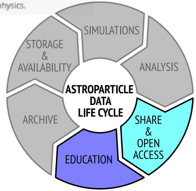
- **APPDC - ?**



Open access and education

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





Thank you
for your attention!

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 (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

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.