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

Astroparticle school 2018, Obertrubach
Victoria Tokareva | October 6, 2018
Introduction:
The astroparticle physics data rate

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/year, which is comparable to the current LHC output*

Big data for deep learning

Modern astroparticle experiments

data rate [Gbytes/day]*

*Berghöfer T., Agrafioti I. et all. Towards a model for computing in European astroparticle physics, Astroparticle Physics European Coordination committee, 2016
Experiments improve and are measuring events with greater precision (large amount of data); but not too many events of our interest; ⇒ combined analysis of data from different experiments becomes topical; Astronomical Virtual Observatories (Auger & IceCube data).
Experiments improve and are measuring events with greater precision (large amount of data);
Experiments improve and are measuring events with greater precision (large amount of data);
But not too many events of our interest;
Experiments improve and are measuring events with greater precision (large amount of data);
But not too many events of our interest;
⇒ combined analysis of data from different experiments becomes topical;
Experiments improve and are measuring events with greater precision (large amount of data);
But not too many events of our interest;
⇒ combined analysis of data from different experiments becomes topical;
Astronomical Virtual Observatories (Auger & IceCube data).
German-Russian Astroparticle Data Life Cycle Initiative*

*Granted by RSF-Helmholtz Joint Research Groups

Introduction

Victoria Tokareva – Cosmic rays data center based on KCDC
KASCADE-Grande

- Proposed in 1989—disassembled in 2013;
- Aimed at studying high-energy (galactic) cosmic rays by observing extensive air showers (EAS);
- Consisted of:
  - scintillators detecting $e$, $\gamma$, $\mu$:
    - KASCADE—256 stations;
    - GRANDE—37 stations;
  - Hadronic calorimeter;
  - Digital radio array LOPES detecting $e$, $e^+$;
- Important features of cosmic-ray spectrum have been obtained. The data analysis is ongoing;
- KCDC (KASCADE Cosmic Ray Data Center, 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 continuously enhanced

**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 0.64m² each
- measures $e/\mu$ from EAS

**Tunka-IACT**
- Imaging Air Cherenkov Telescopes
- is being extended
Archiving and storage

TAIGA data

local_storage_1

local_storage_n

KASCADE data

User

request

request

request

Introduction

Victoria Tokareva – Cosmic rays data center based on KCDC

Data life cycle

Conclusion

October 6, 2018

8/17
Archiving and storage

TAIGA data

... local_storage_1 local_storage_n ...

KASCADE data

User

Introduction

Victoria Tokareva – Cosmic rays data center based on KCDC

Data life cycle

Conclusion

October 6, 2018
Archiving and storage

TAIGA data

local_storage_1

local_storage_n

KASCADE data

User

request

TAIGA main server

KASCADE main server
Archiving and storage

TAIGA data

... local_storage_1 local_storage_n ...

TAIGA main server

Gateway aggregator

KASCADE data

KASCADE main server

User request

TAIGA main server request

Gateway aggregator request

suitable data in a short-term storage?

yes

no

return data

short-term storage

add

Guideline for archiving and storage in a multi-server environment.
Proposed cosmic-ray metadata structure

Introduction

Data life cycle

Conclusion
Data workflow

KASCADE measurements

Simulations

TAIGA measurements

EAS simulation CORSIKA

Input
primary energy $E$
particle mass $A$
shower direction $\theta$, $\phi$

High energy models
QGSJet
EPOS
Sybill
Low energy model FLUKA

detector simulation CRES based on GEANT3

detector simulation for TAIGA

data reconstruction with KRETA
data calibration and correction
reconstruction of shower variables ($N_\text{e}$, $N_\mu$, $\theta$, $\phi$, etc.)
storage in ROOT files

data mapping
comparing distributions
data scaling

metadata database

final reconstruction data scaling

KCDC data access

shower detection
data acquisition
event building

final reconstruction data scaling

data reconstruction for TAIGA
data calibration and correction
reconstruction of shower variables ($N_\text{e}$, $N_\mu$, $\theta$, $\phi$, etc.)
storage in ROOT files

data mapping
comparing distributions
data scaling

metadata database

final reconstruction data scaling

Introduction

Victoria Tokareva – Cosmic rays data center based on KCDC

Data life cycle

Conclusion

October 6, 2018
**Data workflow**

**KASCADE measurements**
- shower detection
- data acquisition
- event building
- event storage

**Simulations**
- EAS simulation CORSIKA
  - Input
    - primary energy $E$
    - particle mass $A$
    - shower direction $\theta$, $\phi$
  - High energy models
    - QGSjet
    - EPOS
    - Sybill
  - Low energy model
    - FLUKA
- detector simulation CRES based on GEANT3
- detector simulation for TAIGA

**TAIGA measurements**
- shower detection
- data acquisition
- event building
- event storage

**Data reconstruction with KRETA**
- data calibration and correction
- reconstruction of shower variables ($N_e$, $N_\mu$, $\theta$, $\phi$, etc.)
- storage in ROOT files

**Final reconstruction**
- data mapping
  - comparing distributions
  - data scaling
- metadata database

**KASCADE**
- data access

**Introduction**
Victoria Tokareva – Cosmic rays data center based on KCDC

**Data life cycle**

**Conclusion**

October 6, 2018
Simulation: two steps

1. Simulating EAS:
   - CORSIKA, does not depend on detector features, depends on location and 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 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 distributed 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, GW Pilot, 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

Appendix

Victoria Tokareva – Cosmic rays data center based on KCDC

October 6, 2018
The German-Russian Astroparticle Data Life Cycle collaboration II

Sinp MSU—Skobelsyn 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,

- KCDC—KASCADE Cosmic Ray Data Center,
  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.