Development of a Data Infrastructure for a Global Data and Analysis Center in Astroparticle Physics ⋆

Victoria Tokareva[1][0000−0001−6699−830X]⋆⋆, Andreas Haungs[1][0000−0002−9638−7574], Donghwa Kang[1][0000−0002−5149−9767], Dmitriy Kostunin[2][0000−0002−0487−0076], Frank Polgart[1][0000−0002−9324−7146], Doris Wochele[1][0000−0001−6121−0632], Jürgen Wochele[1][0000−0003−3854−4890]

[1] Karlsruhe Institute of Technology, Institute for Nuclear Physics, 76021 Karlsruhe, Germany
[2] Deutsches Elektronen-Synchrotron, 15738 Zeuthen, Germany
victoria.tokareva@kit.edu

Abstract. Nowadays astroparticle physics faces a rapid data volume increase. Meanwhile, there are still challenges of testing the theoretical models for clarifying the origin of cosmic rays by applying a multimessenger approach, machine learning and investigation of the phenomena related to the rare statistics in detecting incoming particles. The problems are related to the accurate data mapping and data management as well as to the distributed storage and high-performance data processing. In particular, one could be interested in employing such solutions in study of air-showers induced by ultra-high energy cosmic and gamma rays, testing new hypotheses of hadronic interaction or crosscalibration of different experiments. KASCADE (Karlsruhe, Germany) and TAIGA (Tunka valley, Russia) are experiments in the field of astroparticle physics, aiming at the detection of cosmic-ray air-showers, induced by the primaries in the energy range of about hundreds TeVs to hundreds PeVs. They are located at the same latitude and have an overlap in operation runs. These factors determine the interest in performing a joint analysis of these data. In the German-Russian Astroparticle Data Life Cycle Initiative (GRADLCI), modern technologies of the distributed data management are being employed for establishing a reliable open access to the experimental cosmic-ray physics data collected by KASCADE and the Tunka-133 setup of TAIGA.

Keywords: big data · data engineering · astroparticle physics · KASCADE · TAIGA · GRADLC.

1 Introduction

The AstroParticle Physics European Consortium (APPEC) [8] considers the following challenges for the future usage of information technologies and computing in astroparticle physics: adapting the architecture of computer networks to the rapid growth of the received data amount, usage of distributed data storage and processing systems that have now found their widespread use in both industry and particle physics experiments, and problems of experimental data access and open data.

According to the Berlin Open Data Declaration [9], research data produced with taxpayer money must be publicly available. Currently, the need for open access is recognized everywhere and there are several initiatives aimed at providing access to data [17, 22, 23].

KASCADE [1] was one of the first experiments in the ultra-high energy field that provided access to nearly all of its data according to the principles of FAIR (Findability, Accessibility, Interoperability, and Reusability) [36].

At present, other astrophysical experiments are also moving towards publishing their data, what led to estblishing several global virtual observatories [6, 21, 35]. Following this trend, the TAIGA [12, 32] experiment has shown interest to employ the experience gained in KASCADE for this purpose, what led to the forming of GRADLCI [15] aiming in developing a single center for the analysis and processing of astrophysical data with pilot datasets from both experiments KASCADE and TAIGA. This article discusses the challenges of the data integration from various experiments, the organization of distributed access and processing, i.e. about the important stages of the data processing cycle, also called the data life cycle.

2 Experiments

2.1 KASCADE experiment and data ecosystem

The KArlsruhe Shower Core and Array DEtector (KASCADE) [1] is an experiment in astroparticle physics that was running on Campus North of the Karlsruhe Institute of Technology (KIT) in Germany from October 1996 till December 2013, corresponding to a total of 4383 days of observation. During this time about 450 million events were collected, which resulted in about 4 TB of reconstructed data.

The data was collected for the purpose to study the spectrum of cosmic rays in the energy range of 10^{14}–10^{18} eV.

To achieve this goal, 252 scintillation detectors were placed on the area of 200×200 m^2. Later the setup was extended to KASCADE-Grande [2] and LOPES [26] experiments. The high accuracy of the data collection and the large amount of accumulated statistics made it possible to obtain important results [3, 5] in the field of ultra-high energy astroparticle physics, acknowledged by the community.

There are several levels of reconstructed KASCADE data, starting from the original raw data stored in the CERN ZEBRA [37] format, ending with the high-level of reconstruction shared to the general public.
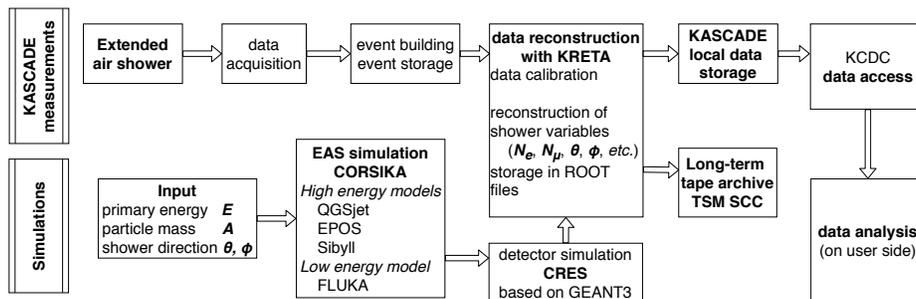

Fig. 1. KASCADE data processing workflow (data life cycle).

Data processing is performed by means of special software developed for the experiment: a data reconstruction program KRETA [31], a program for detector output simulation CRES [18] based on GEANT3 [11] and a program for detailed EAS simulation CORSIKA [19,28]. A scheme of the data reconstruction process is presented in fig. 1.

The open access data are stored locally on KASCADE servers. Data storage on magnetic tapes is used as a long-term storage. It is employing the Tivoli Storage Manager (TSM) of the Steinbuch Centre for Computing (SCC) at KIT.

2.2 TAIGA detector and data engineering

TAIGA (Tunka Advanced Instrument for cosmic ray physics and Gamma Astronomy) is a complex hybrid detector system, which is intended for for cosmic ray studies from 100 TeV to several EeV as well as for a ground-based gamma-ray astronomy for energies from a few TeV to several PeV.

The experiment infrastructure includes several setups observing air showers in a broad energy range. They are wide-angle atmospheric Cherenkov timing arrays Tunka-133 [7] for higher energies and TAIGA-HiSCORE [24] for lower energies, an array of imaging atmospheric Cherenkov telescopes TAIGA-IACT [38], a radio extension Tunka-Rex [10], and a surface scintillator array Tunka-Grande [13].

Currently, all installations together have collected about 50 TB of raw data. Estimates of the current annual data rate and its increase expected in the coming years are given in table 1.

The data collected by the experiment are stored in a distributed way on the servers of the TAIGA project in the Tunka Valley and Irkutsk, as well as on the servers of Moscow State University. Data are stored in four specific binary data

Table 1. Current and expected data rates of TAIGA setups, TB/year

Setup	Current data rate	Expected data rate
TAIGA-HiSCORE	6.4	18
TAIGA-IACT	0.5	1.5
Tunka-Grande, Tunka-133 and Tunka-Rex	0.5	0.5
Total	7.4	20

formats developed specifically for the experiment. After being collected by different setup clusters, the events are preprocessed and merged using timestamps of the single packets. Then the data are calibrated and stored to the server for user access. Parsing and verifying the raw experimental data is performed using the specifications defined with FlexT and Kaitai Struct languages [14].

3 German-Russian Astroparticle Data Life Cycle Initiative

As shown in Ref. [4], a joint analysis of data from the certain setups of the TAIGA and KASCADE experiments is possible and of particular interest, since the experiments are at the same latitude and observe the same region of the celestial sphere, and measure the same range of the energy spectrum of cosmic rays. Thus, a joint analysis of the data from the TAIGA and KASCADE experiments using advanced methods, including machine learning, can be significant in finding the answers to fundamental questions in astroparticle physics. The GRADLC project was created to coordinate the joint work of two independent observatories to join efforts in building a joint data and analysis center [25] for Multi-Messenger Astroparticle Physics.

The main goals of the project include the extension of the KCDC data center of the KASCADE experiment by adding access to the TAIGA data, software development for collaborative data analysis, providing data analysis capabilities on the data center side, and implementing solutions for visualizing analysis results.

3.1 KASCADE Cosmic-ray Data Center

The KASCADE Cosmic-ray Data Center (KCDC) [27, 30] was established in 2013 to provide users a reliable access to the cosmic-ray data collected by the KASCADE experiment. These data include measured and reconstructed parameters of more than 433 million air showers, metadata information, simulations data for three different high energy interaction models, published spectra from various experiments, and detailed educational examples. All together enable users outside the community of experts to perform their own data analysis.

With the last release, named NABOO [29], more than 433 million events are provided from the whole measuring time of KASCADE-Grande.

The KCDC software architecture is presented in fig. 2. Adhering to the ideas of open access, KCDC relies only on non-commercial open source software.
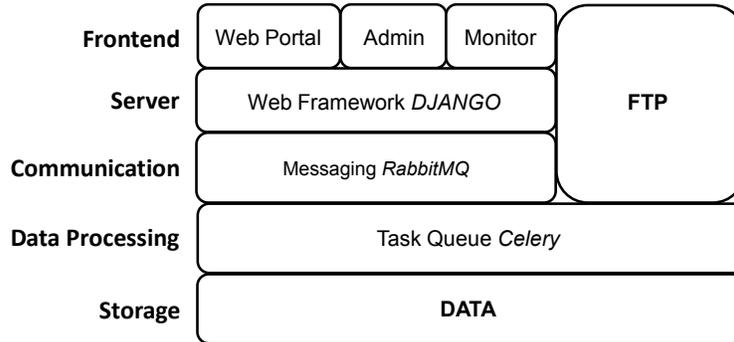
Fig. 2. KCDC IT structure [27].

Expanding the experimental data by adding new detector components could require to change the structure of a stored event. In order to do this without the restraint of a fixed database schema, a NoSQL database MongoDB has been chosen to store the experimental data. MongoDB uses JSON-like documents with schemata. It supports field, range query, and regular expression searches, and indexing with primary and secondary indices. MongoDB scales horizontally using sharding and can run over multiple servers. Currently KCDC is running MongoDB on a single server, but we are aiming at a sharded cluster for better performance.

The full KCDC system runs on an `nginx` [33] server and communicates with the database server and worker nodes via the `RabbitMQ` [34] open-source message-broker software. The KCDC web site is built using `Django` web framework [20], that is Python-based and follows the model-view-template architectural pattern. Each worker node is managed and monitored via the `Celery` [16] open source asynchronous task queue based on distributed message passing. Python tools on the worker nodes process data selections issued by users. The selections are stored on a dedicated FTP server, where they can be retrieved by a registered user, after the processing of their jobs has been successfully finished.

3.2 Extending KASCADE data center with TAIGA data

In the process of the data center extension the following challenges appeared.

Increasing data access speed. With the increase of the data amount, there arises a challenge of maintaining the speed of searching data on the server and providing search and selection results to the user. At the same time, the number of requests to the data center increases, which increases the probability of a server access-denial error. To solve this problem, a data aggregation server is introduced as an intermediate node for communication with users. Such solution allows one to cache data which are most frequently requested for, and to reduce the server

load by shielding user requests on the aggregation server, thereby helping to maintain stability and performance.

In addition, on the aggregation server one can perform a primary data search using the so-called event level metadata database. Requests related to data analysis usually correspond to the event level, and data selections are being performed using certain criteria, such as the reconstructed energy of the event, the zenith angle, the maximum shower depth, the number of electrons, etc.

Selecting database type for the metadata database. In modern high-load projects, NoSQL databases are in a wide use. Due to the lack of strict requirements for a structure of the stored data, the databases of this family make it possible to easily scale the system over time, adding data of an arbitrary structure to it. Also, these types of databases make it easy to distribute files on servers, thereby reducing the load on unified storage and facilitating data backup. An example of such a database is MongoDB that is used to store KASCADE data. On the other hand, SQL databases allow for very fast searches on data of fixed structure. This is a proven database type with a well-documented standard. The main advantage of an SQL database is declarativeness: with the help of SQL, the programmer describes only what data to extract or modify, and the exact way to do this is chosen by the database management system when processing the SQL query. At the moment we are considering PostgreSQL as an intermediate solution: an SQL database, which allows additional XML fields to be entered into its structure.

Providing a common interface for data access. The KASCADE data are high-level data, while the TAIGA data are stored in a binary format. At the same time, access should be provided at high-level of data reconstruction for all users. To achieve this effect, we are introducing an intermediate level of data search on the TAIGA server side using file level metadata. At the same time, events are not reconstructed at the binary level; so for the search we can use only basic information presented in the catalogs: setup, data collection season, month, day, file size, file type, etc. The raw data found using such criteria is then transferred to the aggregation server and reconstructed there using special software for the subsequent download by the user.

4 Outlook

The GRADLC project was initiated to provide the public with access to data from two experiments, KASCADE and TAIGA. Joint analysis of these data can bring us closer to answering fundamental questions in the field of astroparticle physics.

However, infrastructure development for data curation and joint data analysis is associated with overcoming a series of challenges, in particular, organizing a common interface for data access, expanding the current data center of KCDC

while maintaining stability and performance, finding solutions for data aggregation, and some others beyond the scope of this article.

In the process of working on the project, we are trying to use proven solutions for big data processing, which are in use in the industry and particle physics. In particular, these are solutions for working with metadata, data caching and aggregation, distributed storage and processing.

References

1. Antoni, T., et al.: The cosmic-ray experiment KASCADE. Nucl. Instrum. Methods Phys. Res. Sect. A **513**(3), 490–510 (2003)
2. Apel, W.D., et al.: The KASCADE-Grande experiment. Nucl. Instrum. Methods Phys. Res. Sect. A **620**(2), 202–216 (2010)
3. Apel, W.D., et al.: Kneelike structure in the spectrum of the heavy component of cosmic rays observed with KASCADE-Grande. Phys. Rev. Lett. **107**(17), 171104 (2011)
4. Apel, W.D., et al.: A comparison of the cosmic-ray energy scales of Tunka-133 and KASCADE-Grande via their radio extensions Tunka-Rex and LOPES. Phys. Lett. B **763**, 179–185 (2016)
5. Apel, W.D., et al.: KASCADE-Grande limits on the isotropic diffuse gamma-ray flux between 100 TeV and 1 EeV. Astrophys. J. **848**(1), 1 (2017)
6. AstroGrid: UK's Virtual Observatory Service. http://www.astrogrid.org
7. Berezhnev, S., et al.: The Tunka-133 EAS cherenkov light array: Status of 2011. Nucl. Instrum. Methods Phys. Res. Sect. A **692**, 98–105 (2012)
8. Berghöfer, T., et al.: Towards a model for computing in european astroparticle physics. arXiv:1512.00988 [astro-ph.IM] (2015)
9. Berlin declaration on open access to knowledge in the sciences and humanities. https://openaccess.mpg.de/Berlin-Declaration, published: January 2015
10. Bezyazeekov, P.A., et al.: Measurement of cosmic-ray air showers with the Tunka Radio Extension (Tunka-Rex). Nucl. Instrum. Methods Phys. Res. Sect. A **802**, 89–96 (2015)
11. Brun, R., Bruyant, F., Maire, M., McPherson, A.C., Zanarini, P.: GEANT 3: user's guide Geant 3.10, Geant 3.11. Tech. rep., CERN, Geneva (1987)
12. Budnev, N., et al.: The TAIGA experiment: from cosmic ray to gamma-ray astronomy in the Tunka valley. J. Phys. Conf. Ser. **718**(5), 052006 (2016)
13. Budnev, N., et al.: The Tunka-Grande experiment. J. of Instr. **12(06)**, C06019–C06019 (2017)
14. Bychkov, I., et al.: Using binary file format description languages for documenting, parsing, and verifying raw data in TAIGA experiment. In: Proceedings of the VIII International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2018). Dubna, Russia (2018), https://arxiv.org/abs/1812.01324
15. Bychkov, I., et al.: Russian-German Astroparticle Data Life Cycle Initiative. Data J. **3**(4), 56 (2018)
16. Celery distributed task queue. http://www.celeryproject.org/
17. CERN Open Data. http://opendata.cern.ch
18. Cosmic Ray Event Simulation (CRES). https://kcdc.ikp.kit.edu/static/pdf/kcdc_mainpage/kcdc-Simulation-Manual.pdf, p. 13

19. COsmic Ray SImulations for KAscade (CORSIKA).
    https://www.ikp.kit.edu/corsika
20. Django web framework documentation. https://docs.djangoproject.com/en/2.2/
21. Euro-VO: the European Virtual Observatory (a partnership of VOs including AstroGrid, the French-VO, ESO, ESA, etc.). http://www.euro-vo.org
22. European Open Science Cloud (EOSC).
    https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
23. European Science Cluster of Astronomy & Particle physics (ESCAPE).
    https://www.escape2020.eu
24. Gress, O., et al.: Tunka-HiSCORE – a new array for multi-TeV γ-ray astronomy and cosmic-ray physics. Nucl. Instrum. Methods Phys. Res. Sect. A **732**, 290–294 (2013)
25. Haungs, A.: Towards a global analysis and data center in Astroparticle Physics. these proceedings (2019)
26. Haungs, A., et al.: Air shower measurements with the LOPES radio antenna array. Nucl. Instrum. Methods Phys. Res. Sect. A **604**(1-2), S1–S8 (2009)
27. Haungs, A., et al.: The KASCADE Cosmic-ray Data Centre KCDC: granting open access to astroparticle physics research data. Eur. Phys. J. C **78**(9), 741 (2018)
28. Heck, D., Schatz, G., Knapp, J., Thouw, T., Capdevielle, J.: CORSIKA: a Monte Carlo code to simulate extensive air showers. Tech. rep., Forschungszentrum Karlsruhe GmbH, Karlsruhe (1998)
29. Kang, D., et al.: A new release of the KASCADE cosmic ray data centre (KCDC). In: 35th International Cosmic Ray Conference (ICRC2017). p. 452. Proceedings of Science (2017)
30. KASCADE Cosmic-ray Data Center (KCDC). https://kcdc.ikp.kit.edu
31. KASCADE Reconstruction for ExTensive Airshowers (KRETA).
    https://kcdc.ikp.kit.edu/static/pdf/kcdc_mainpage/kcdc-Simulation-Manual.pdf, p. 13
32. Kostunin, D., et al.: Tunka Advanced Instrument for cosmic rays and Gamma Astronomy. In: 18th International Baikal Summer School on Physics of Elementary Particles and Astrophysics: Exploring the Universe through multiple messengers (ISAPP-Baikal 2018) Bolshie Koty, Lake Baikal, Russia, July 12-21, 2018 (2019)
33. NGINX documentation. https://docs.nginx.com/
34. RabbitMQ message-broker. https://www.rabbitmq.com/
35. SPASE: Space Physics Archive Search and Extract. http://www.spase-group.org
36. Wilkinson, M.D., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**, 160018 (Mar 2016)
37. ZEBRA reference manual. https://cdsweb.cern.ch/record/2296399/files/zebra.pdf
38. Zhurov, D., et al.: First results of the tracking system calibration of the TAIGA-IACT telescope. J. Phys.: Conf. Ser. **1181**, 012045 (2019)