Distributed Data Storage for Modern Astroparticle Physics Experiments[*]

Alexander Kryukov[1][0000−0002−1624−6131], Minh-Duc
Nguyen[1][0000−0002−5003−3623], Igor Bychkov[2], Andrey
Mikhailov[2][0000−0001−5572−5349], Alexey Shigarov[2], and Julia
Dubenskaya[1][0000−0002−2437−4600]

[1] Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University,
Moscow 119992, Russia
kryukov@theory.sinp.msu.ru, nguyendmitri@gmail.com, jdubenskaya@gmail.com
[2] Matrosov Institute for System Dynamics and Control Theory,
Siberian Branch of Russian Academy of Sciences, Lermontov st. 134, Irkutsk, Russia
shigarov@icc.ru

Abstract. The German-Russian Astroparticle Data Life Cycle Initiative is an international project launched in 2018. The Initiative aims to develop technologies that provide a unified approach to data management, as well as to demonstrate their applicability on the example of two large astrophysical experiments - KASCADE and TAIGA. One of the key points of the project is the development of a distributed storage, which, on the one hand, will allow data of several experiments to be combined into a single repository with unified interface, and on the other hand, will provide data to all participants of experimental groups for multi-messenger analysis. Our approach to storage design is based on the single write-multiple read (SWMR) model for accessing raw or centrally processed data for further analysis. The main feature of the distributed storage is the ability to extract data either as a collection of files or as aggregated events from different sources. In the last case the storage provides users with a special service that aggregates data from different storages into a single sample. Thanks to this feature, multi-messenger methods used for more sophisticated data exploration can be applied. Users can use both Web-interface and Application Programming Interface (API) for accessing the storage. In this paper we describe the architecture of a distributed data storage for astroparticle physics and discuss the current status of our work.

Keywords: Astroparticle physics · Distributed storage · Open science · CERNVM-FS · Timeseries DB.

1 Introduction

Currently, a number of experimental facilities in the field of particle astrophysics of the mega-sciences class are under construction or are already operating around

---

the world. Among them there are such installations as LSST [1, 2], MAGIC [3, 4], CTA [5, 6], VERITAS [7], HESS [8], and others. These facilities collect a tremendous volume of data. For example the annual (reduced) raw data of the CTA project have a volume of about 4 PB. The total volume to be managed by the CTA archive is of the order of 25 PB per year, when all data-set versions and backup replicas are considered.

In addition to the huge flow of data produced, an important feature of this class of projects is the participation of many organizations and, as a result, the distributed nature of data processing and analysis. All this presents a real challenge to developers of the data analytics infrastructure.

To meet a similar challenge in high energy physics, the WLCG grid was deployed as part of the LHC project [9]. This solution, on the one hand, proved to be highly efficient, but on the other hand, it turned out to be a rather heavy one requiring high administrative costs, highly qualified staff and a very homogeneous environment on which applications operate. The success of the WLCG is based primarily on the fact that thousands of physicist users actually solve one global problem using a highly centralized system management.

Taking into account the tendency to a multi-messenger analysis [10] of data with its potential for a more accurate exploration of the Universe and mordern trend to open science [11, 12], it is very important to provide users from geographically distributed locations with access to the data of different astrophysics facilities. Today, open access to data or, more generally, open science is becoming increasingly popular. This is due to the fact that the amount of data received in some experiment often exceeds the capabilities of the relevant collaboration to process and analyze these data. And only the involvement of all scientists interested in research in this area allows for a comprehensive analysis of the data in full.

Please note that most existing collaborations have a long history and apply methods for data processing they are accustomed to. So, our approach to the design of data storage for astroparticle physics should be based on two main principles. The first principle is that there is no interference with the existing local storage. And the second principle is the processing of user requests in a special service outside the local storage using metadata. The interaction between local storages and any user of the system should be provided by special adaptors which define a unified interface for data exchange in the system. Our approach to storage design is based on a single write-multiple read model (SWMR) for accessing raw data or centrally processed data for further analysis. The motivation for the solution is that both raw data and data after the initial processing (for example, calibration) should be stored unchanged and presented to users as is upon request. A similar approach is being discussed in the HDF5 community [13].

The main ideas of the proposed approach are as follows:

- no changes in the inner structure of local storage;
- unification of access to local storage based on corresponding adapter modules;
- use of local data access policies;

- search of the requested data using the only metadata on a special service;
- aggregating the requested data into a new collection and providing the user with access to it;
- data transfer only at the moment of actual access to them.

Based on the above principles and ideas, we propose a concept of distributed storage for astrophysical experiments, which we call APPDS (abbreviated from AstroParticle Physics Distributed Storage). The prototype of such distributed storage is developed in the framework of the German-Russian Astroparticle Data Life Cycle Initiative [17]. This initiative aims to develop a distributed data storage system in the field of astrophysics of particles by the example of two experiments KASCADE [14] and TAIGA [15, 16], as well as to demonstrate its viability, stability and efficiency.

Below we discuss the architecture of the distributed data storage and briefly report the current status of the project and the nearest plans.

2 Architecture of the data storage

One of the main ideas of the distributed data storage architecture for the physics of astroparticles is that we do not interfere with the work of local storages S1 ... S3 (see Fig. 1). This is achieved by using special adapter programs A1 ... A3 that allow local storages to interact with the data aggregation service. As adapters, we use the CERNVM-FS [18] file system to export local file systems to the aggregation service in a read-only mode. First, it provides a transparent way for users to interact with local storages. Secondly, the actual transfer of data will only occur when a user actually accesses these data. Additionally, reducing network traffic can be achieved through the use of CVMFS caching properties.

To retrieve the necessary files, a user forms a request through the web interface provided by the Data Aggregation Service. When the Data Aggregation Service receives the user request, it requests a response from the Metadata Catalogue (MDC). After the Metadata Catalogue responds, the Data Aggregation Service forms the corresponding resulting response and delivers it to the user.

The proposed system offers two types of search conditions for user requests: a file-level search and an event-level search.

In the case of a file-level search, the user requests a set of files, imposing conditions on the metadata (that is, on the data about the files). An example of such a condition is the range of dates of observation of gamma sources in the sky. It is important to note that the user will receive in response the corresponding set of files with the same directory structure as in the original repository. Thus, the application software can be run without modification, as if the user runs the program locally.

In the case of a event-level search, the user wants to select from the files only some events that satisfy the search conditions, for example, some energy range of the air flow. In this case the events are selected from the files and the aggregation service prepares a new one which contains only the necessary events.
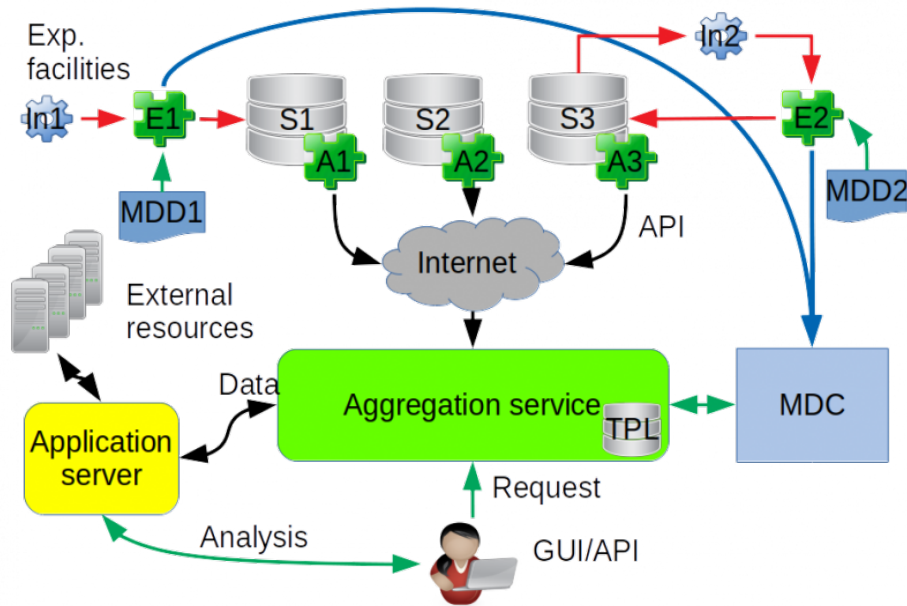
Fig. 1. The architecture of APPDS

The new file is transferred to the user. However, the directory structure will be preserved too.

The processing of user requests is performed by the metadata catalogue, the main purpose of which is to specify, up to an event, in which files and where the data requested by the user are contained. The MDC service is built around TimescaleDB [20–22].

The extractors E1, E2 play a key role in the architecture of APPDS. All data stored in the local storages must pass through the extractors. The extractors take off metadata from the data and store the metadata in MDC. The type of the extracted metadata is defined by the metadata description file (MDD) which is used as input for the extractor. The MDD file is written in Kaitai Struct [23, 24] format with special marks pointing to elements of binary data which are metadata and should be extracted.

The extractor E1 takes out metadata from raw data, while the extractor E2 takes out metadata from centrally processed data (for example, from data after calibration or calculation of the shower energy). Thus, the information needed to process user requests is collected in the MDC service.

It is important to note that all services in APPDS are built as microservices [25] and have a well–defined REST API [26]. Some services are running in Docker containers [27].

A more detailed description of the aggregation service and the metadata catalogue service can be found in the papers by M-D.Nguyen [28] and I.Bychkov [29] in these proceedings.

3   Status

Currently a prototype of APPDS was deployed in Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University. The prototype consists of two local storages interconnected via a local network for modelling distributed storage, an aggregation service and a metadata service based on TimescaleDB. The next version of the system will also include KCDC [30] storage at KIT and storage at Irkutsk State University.

Most of the components of the system are written in Python. As the first-time example of the production use of the system, users of the KASCADE and TAIGA/TUNKA collaborations will gain access to the data of these experiments, as well as the Monte Carlo simulation data. It should be mentioned that the system is developed for broad general use and is not limited to astrophysics applications.

References

1. Large Synoptic Survey Telescope. https://www.lsst.org/
2. Kahn, S. M. Project Status. https://project.lsst.org/groups/sac/sites/lsst.org.groups.sac/files/Kahn_projectstatus.pd
3. MAGIC. https://doi.org/10.15161/oar.it/1446204371.89
4. Ricoa, J. for the MAGIC Collaboration: Overview of MAGIC results. In. 37th International Conference on High Energy Physics, 2-9 July 2014 ● Valencia, Spain, Nuclear and Particle Physics Proceedings, **273–275**, 328-333 (2016)
5. Cherenkov Telescope Array. Exploring the Universe at the Highest Energies. https://www.cta-observatory.org/. Last accessed 24 Jan 2019
6. The Cherenkov Telescope Array Consortium: Science with the Cherenkov Telescope Array. Arxiv: 1709.07997, https://arxiv.org/pdf/1709.07997. Last accessed 24 Jan 2019
7. VERITAS. https://veritas.sao.arizona.edu/. Last accessed 24 Jan 2019
8. HESS, https://www.mpi-hd.mpg.de/hfm/HESS/. Last accessed 24 Jan 2019
9. Worldwide LCH Computing GRID. http://wlcg.web.cern.ch/
10. Franckowiak, A.: Multimessenger Astronomy with Neutrinos. J. Phys.: Conf. Ser., **888**, 012009 (2017)
11. Voruganti, A., Deil1, Ch. , Donath, A., and King, J.: gamma-sky.net: Portal to the Gamma-Ray Sky. Arxiv: 1709.04217, https://arxiv.org/pdf/1709.04217. Last accessed 24 Jan 2019
12. Wagner, S.: Gamma – Ray Astronomy in the 2020s. https://www.eso.org/sci/meetings/2015/eso-2020/eso2015_Gamma_Ray_Wagner.pdf. Last accessed Jan. 24 2019
13. HDF5 Single–writer/Multiple–reader User's Guide. https://support.hdfgroup.org/HDF5/docNewFeatures/SWMR/HDF5_SWMR_Users_Guide.pdf. Last accessed June 06, 2019.

14. W.D.Apel and etc. The KASCADE-Grande experiment. Nuclear Instruments and Methods in Physics Research, Section A, **620**(2010), pp.202–216, https://doi.org/10.1016/j.nima.2010.03.147

15. TAIGA. https://taiga-experiment.info/. Last accessed 24 Jan 2019

16. Budnev, N. and etc. The TAIGA experiment: From cosmic-ray to gamma-ray astronomy in the Tunka valley. Nuclear Instruments and Methods in Physics Research. Section A, **845**(2017), pp.330–333, https://doi.org/10.1016/j.nima.2016.06.041

17. Bychkov, I., et al.: Russian–German Astroparticle Data Life Cycle Initiative. Data, **4**(4), 56 (2018). DOI: 10.3390/data3040056.

18. Blomer, J. , Buncic, P. , Ganis, G. , Hardi, N. , Meusel, R., and Popescu, R.: New directions in the CernVM file system. In. 22nd International Conference on Computing in High Energy and Nuclear Physics (CHEP2016), 10–14 October 2016, San Francisco, USA. Journal of Physics: Conf. Series, **898**, 062031 (2017)

19. MariaDB home page. https://mariadb.org/. Last accessed Jan. 24, 2019

20. Freedman, M.J.: TimescaleDB: Re-engineering PostgreSQL as a time-series database. https://www.percona.com/live/18/sites/default/files/slides/TimescaleDB-Percona-2018-main.pdf. Last accessed 24 Jan 2019

21. Yang,Ch., et. al.: AstroServ: Distributed Database for Serving Large-Scale Full Life-Cycle Astronomical Data. ArXiv: 1811.10861. https://arxiv.org/pdf/1811.10861. Last accessed 24 Jan 2019.

22. Stefancova, E.: Evaluation of the TimescaleDB PostgreSQL Time Series extension. https://cds.cern.ch/record/2638621/files/evaluation-timescaledb-postgresql.pdf. Last accessed 24 Jan 2019

23. Kaitai Struct. http://doc.kaitai.io/. Last accessed 24 Jan 2019

24. Bychkov, I. et al.: Using binary file format description languages for documenting, parsing and verifying raw data in TAIGA experiment. In. International Conference "Distributed Computing and Grid-technologies in Science and Education" 2018 (GRID'2018), Dubna, Russia, September 10-14, 2018. CEUR Workshop Proceedings, **2267**, 563-567 (2018).

25. Sill, A.: The Design and Architecture of Microservices. IEEE Cloud Computing, **3**(5), 76-80 (2016)

26. Fielding, R. Th.: Architectural Styles and the Design of Network-based Software Architectures. https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf, PhD Thesis (2000). Last accessed 24 Jan 2019

27. Docker home page. https://www.docker.com/. Last accessed 24 Jan 2019

28. Nguyen, M.-D. and etc. Data aggregation in the Astroparticle Physics Distributed Data Storage. In Proc. of 3-d Int Workshop DLC-2019 (this book).

29. Bychkov, I. and etc. Metadata extraction from raw astroparticle data of TAIGA experiment. In Proc. of 3-d Int Workshop DLC-2019 (this book).

30. KASCADE Cosmic Ray Data Centre (KCDC). https://kcdc.ikp.kit.edu/. Last accessed 24 Jan 2019