Some Thoughts on Data Strategies for Large-Scale Scientific Facilities in Frontier Physics (Postprint)

Authors: Chen Gang

Date: 2023-03-19T00:00:00+00:00

Abstract

Frontier physics large-scale scientific facilities constitute national strategic assets for fundamental physics research and addressing national strategic needs. In particular, China has increased its investment in the construction and operation of such facilities in recent years, providing crucial platforms for both basic and applied research. The data generated by these facilities pose enormous challenges to scientific computing, data analysis, and data management. Beyond focusing on the software and hardware technologies required for building high-level data processing and scientific computing platforms, it is essential to address strategic issues necessary for maximizing data utility. This article analyzes the data characteristics of frontier physics large-scale scientific facilities and presents a brief discussion on three aspects: data sharing, long-term data preservation and reuse, and data talent strategy, with the aim of providing references for data application of large-scale scientific facilities in China.

Full Text

Preamble

Large research facilities for physics frontiers represent national strategic assets that serve both fundamental research and national strategic needs. In recent years, China has significantly increased investment in the construction and operation of these facilities, providing essential platforms for basic and applied research. However, the massive data generated by these facilities poses enormous challenges to scientific computing, data analysis, and data management. Beyond the hardware and software technologies required to build high-performance data processing and scientific computing platforms, equal attention must be paid to strategic issues that maximize the utility of these data. This article analyzes the characteristics of data from large physics frontier facilities and discusses three

key aspects—data sharing, long-term data preservation and reuse, and talent strategy—aiming to provide references for data utilization at China's major scientific facilities.

Keywords: large facilities, big data, data management, data sharing, data preservation, physics frontier
DOI: 10.16418/j.issn.1000-3045.2018.08.015

Introduction

Large research facilities for physics frontiers constitute the largest proportion of China's major scientific facilities and are particularly crucial for national scientific research and strategic needs. These facilities generate the most extensive and structurally complex datasets. Our goal is to efficiently maximize the scientific return from these data. While the application of data in scientific research is not the focus of this paper, we aim to discuss policy-related issues in data management and application, hoping to provide references for data producers, users, and investors in large facilities when formulating policies and strategies.

These facilities represent the most important tools and conditions for contemporary fundamental and applied physics research both domestically and internationally, encompassing major particle physics experiments, neutrino experiments, heavy ion accelerators, tokamak experiments, astronomical telescopes such as FAST and LAMOST, ground-based and space-based cosmic ray and astrophysics observatories, synchrotron radiation platforms, spallation neutron sources, and steady high magnetic field facilities. Particle physics experiments, fusion experiments, astronomical telescopes, and cosmic ray observatories are primarily dedicated facilities used to study fundamental structures of matter and cosmic evolution from microscopic to cosmological scales. Synchrotron radiation, spallation neutron sources, and steady high magnetic field facilities serve as national public experimental platforms open to research and industry communities for microscopic studies in life sciences, materials science, chemistry, and physics.

International high-energy physics experiments, represented by the Large Hadron Collider (LHC) at CERN, generate dozens of petabytes of data annually. The Beijing Electron-Positron Collider represents China's most important high-energy physics experimental facility, producing over 10 petabytes of data in recent years. Cosmic ray and astrophysics observation platforms fall into two categories: ground-based cosmic ray observatories and space science satellites. The Yangbajing Cosmic Ray Observatory and the Large High Altitude Air Shower Observatory (LHAASO) under construction in Daocheng are among the world's most important ground-based observatories, collecting cosmic ray data at the petabyte scale annually.

China's synchrotron radiation facilities include the operational Beijing Synchrotron Radiation Facility, Shanghai Synchrotron Radiation Facility, and Hefei Synchrotron Radiation Facility, as well as the Beijing High Energy Photon

Source (to be constructed) and Shanghai Hard X-ray Free Electron Laser Facility (under construction). Additionally, China's Spallation Neutron Source has been completed and put into operation. These public experimental platforms attract thousands of scientists from various disciplines annually, generating petabyte-scale data. All data from these large facilities represent primary scientific materials and the source of scientific discoveries.

Data Sharing and Utilization

Large facilities for physics frontiers are characterized by massive scale, long construction and operation cycles, and scientific goals targeting international frontiers while making strategic, fundamental, and forward-looking contributions to national economic and social development. The data they produce are gold mines for scientific discovery, though data sharing and application models vary by facility type.

Particle Physics Experiments

Current particle physics experiments involving Chinese scientists include foreign-based experiments such as CERN's LHC experiments and China-based, China-led experiments such as the BESIII experiment at the Beijing Electron-Positron Collider, the Daya Bay neutrino experiment, and the LHAASO experiment. All these particle physics experiments adopt international collaboration models, with partners sharing responsibilities for construction, operation, and management. Consequently, experimental data are freely shared and jointly utilized among collaborating institutions, with scientific results published collectively and ownership shared.

Nevertheless, competition exists among collaboration members, with each party striving to achieve research results first. Therefore, beyond deploying top scientists, optimal conditions must be created for data transmission and computing to produce scientific outcomes rapidly. While China has made important contributions to LHC construction and upgrades, insufficient investment in data transmission, sharing, and scientific computing has adversely affected Chinese scientists' LHC physics research. As LHC undergoes high-luminosity upgrades with data production rates increasing by dozens of times, this poses enormous challenges for data transmission and processing. We recommend that the state provide strong support for LHC China groups in networking and distributed computing to facilitate first-rate physics results. Simultaneously, in China-led particle physics experiments where we hold management authority, we should adopt appropriate strategies and technical means to gain initiative in data sharing and utilization under fair cooperation principles.

Astronomical Observations

Astronomical observations, particularly large general-purpose telescopes, predominantly adopt delayed public release for data sharing internationally. Ob-

servers obtain exclusive access to data during a protection period to analyze and achieve scientific results quickly, after which data are publicly released. Typically, this delay lasts 1-2 years. After the protection period, data are stored on database servers for access and use by astronomers worldwide. This sharing model deserves emulation by other fields: data public release enables more scientists to utilize the data for additional research findings, while peer access facilitates verification of results. Currently, space science satellites and cosmic ray observation experiments are also adopting this model, releasing satellite observation data and cosmic ray data in batches for scientific research.

Synchrotron Radiation and Spallation Neutron Source Facilities

These facilities are public experimental platforms constructed with national investment. Scientists from academic institutions can apply to conduct experiments, with generated data used for scientific research. International synchrotron facilities have corresponding policies for experimental data. The European Synchrotron Radiation Facility (ESRF) stipulates that ESRF preserves all experimental raw data and metadata with a 3-year protection period, extendable when necessary. During the protection period, experimenters have full usage rights; afterward, ESRF makes data available to registered ESRF users under appropriate licensing conditions, requiring citation in publications.

Domestic synchrotron radiation and spallation neutron source facilities currently offer free access to scientists from universities and research institutions. However, the absence of unified data policies for these domestic facilities hinders maximizing experimental data benefits. Since these facilities are constructed and operated with national investment, the state should have joint ownership of the generated data. Therefore, we recommend establishing national data policies similar to international ones that both protect experimenters' priority usage rights and maximize data utility through sharing. Public experimental platforms can adopt two data sharing models: (1) establishing a 2-3 year data protection period to ensure experimenters' priority usage rights; (2) for external users urgently needing experimental data, establishing data sharing mechanisms through cooperation agreements with experimenters, enabling these users to conduct scientific research with data even during the protection period.

Data Preservation and Reuse

The construction, maintenance, and data acquisition of large facilities for physics frontiers consume substantial human and material resources, making experimental data extremely valuable. Scientists' utilization of data does not cease when data acquisition ends; many experiments continue data analysis and publish related papers for several years post-acquisition. Data from different large facilities are unique, and new scientific discoveries may emerge from old experimental data as theoretical research advances and analytical methods improve. Additionally, joint analysis and cross-validation of new and old data from different experiments can enhance the precision and credibility of scientific discoveries.

Another important use of large facility data is for teaching and science popularization in universities and schools. Thus, long-term data preservation is of paramount importance.

Data preservation encompasses not only experimentally acquired data but also knowledge bases. Knowledge bases include parameters describing experimental conditions, software for data classification, documentation, and other materials required for data analysis—essential information for correct data reuse and analysis. Different types of subsequent data analysis impose different requirements: some require original experimental data, while others need only processed high-level data. In high-energy physics, for instance, the international community has established the DPHEP (Data Preservation in High Energy Physics) collaboration group (with the Institute of High Energy Physics, Chinese Academy of Sciences as one of the initiating units) and compiled a technical white paper on data preservation. This white paper provides detailed descriptions of data and knowledge base preservation, related technologies, and strategies. China's large facilities for physics frontiers lack systematic planning and strategies for long-term data preservation and reuse, making this white paper an excellent reference for national policy formulation. Furthermore, since China's funding mechanisms are primarily project-based, obtaining follow-up funding for data preservation after facility operation ends is difficult. Therefore, corresponding funding mechanisms should be established to ensure long-term preservation and efficient reuse of data after large facility operations conclude.

Talent Strategy

Large facilities for physics frontiers are unprecedented in scale and complexity. The data analysis process is complex and voluminous, requiring not only physics professionals but also their collaboration with computer scientists. Most physics professionals lack sufficient training in computer technology, particularly young master's and doctoral graduates and postdocs who face challenges with data analysis tools, software, and programming languages. Therefore, large facility projects should provide on-the-job computer technology training for these physics professionals. CERN annually hosts high-level summer schools on computing technology, selecting outstanding young students and scientists worldwide to participate in scientific computing training courses and internships. China should establish high-level training courses tailored to different large facilities or scientific computing methods, encouraging scientists to participate in computing technology training to significantly advance researchers' software and data analysis capabilities and promote scientific output.

At the Institute of High Energy Physics, Chinese Academy of Sciences, for example, a team of computer experts has been assembled to cooperate and communicate with physicists, optimizing data analysis software. Physicists, in turn, engage in deep communication with computer experts based on the requirements and characteristics of physics analysis computing, optimizing computer hardware platforms, data management systems, and middleware systems. Com-

puter experts help physicists optimize software to improve data access and software execution efficiency while designing and constructing data storage systems and computing cluster architectures according to physicists' data access patterns and CPU utilization characteristics, achieving maximum data processing efficiency. This two-way communication ensures that data analysis computing systems meet scientific computing requirements with optimal efficiency.

High-quality software is key to the success of large facilities. Recognizing software developers' work and ensuring their career advancement and compensation are necessary conditions for attracting high-level software talent to stably engage in software development and maintenance for large facility data and computing. In high-energy physics, for instance, CERN maintains a high-level team for computer and physics software development. For decades, this team has dedicated itself to developing large-scale general-purpose physics and data analysis software. Their "World Wide Web technology" has become the world's most important network technology, significantly advancing internet development. Additionally, CERN's physics simulation software "GEANT4" has become fundamental for calculations in particle physics, nuclear physics, nuclear medicine, and radiation technology, while "ROOT" has become core technology for data analysis. This demonstrates the importance of ensuring scientists can fully dedicate themselves to software research and development.

To enhance software developers' visibility, we should encourage them to publish articles on software development techniques and achievements. Simultaneously, we should encourage or require domain scientists to appropriately cite the software used in their publications. This is particularly important for properly recognizing software developers' contributions.

References

1. 中国科学院条件保障与财务局. 中国科学院重大科技基础设施. [2018-08-14]. http://lssf.cas.cn/dzzRegisterController.do?outerelss&flag=99.
2. European Organization for Nuclear Research (CERN). Large Hadron Collider. [2018-08-14]. https://home.cern/topics/large-hadron-collider.
3. European Synchrotron Radiation Facility. ESRF Data Policy. [2018-08-14]. http://www.esrf.eu/datapolicy.
4. DPHEP Study Group. Towards a Global Effort for Sustainable Data Preservation in High Energy Physics. [2018-08-14]. https://arxiv.org/pdf/1205.4667v1.pdf.
5. European Organization for Nuclear Research (CERN). CERN School of Computing. [2018-08-14]. http://csc.web.cern.ch.

Author Information

CHEN Gang is a Professor and Deputy Director of the Institute of High Energy Physics (IHEP), Chinese Academy of Sciences (CAS). He received his BSc degree from Nanjing University in 1982 and his Ph.D. from IHEP in 1994.

In the 1990s, he worked as an experimental physicist on the L3 experiment at CERN, the Alpha Magnetic Spectrometer (AMS) project, and the BES experiment on the Beijing Electron Positron Collider (BEPC). Since 2003, he has been in charge of providing high-performance computing infrastructure for high-energy physics projects. In 2005, he became a member of the International High Energy Physics Computing Coordination Committee (IHEPCCC). Since 2006, he has led EU FP6/FP7 projects on grid and cloud computing as the Chinese coordinator and chair of the project management committee. In 2017, he became the Principal Investigator of the Project of High Performance Application Software System for High Energy Physics (HEPHPC), an HPC project under the National Key R&D Plan of the Ministry of Science and Technology.
E-mail: Gang.Chen@ihep.ac.cn

Responsible Editor: YUE Lingsheng

Note: Figure translations are in progress. See original paper for figures.