Scientific Big Data Management Technologies and Systems Postprint

Authors: Li Jianhui, Li Yuepeng, Wang Huajin, Chen Mingqi

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

Abstract

Given that modern scientific discoveries increasingly depend on the analysis and processing of large-scale scientific data, the efficient management of scientific big data has become a pressing issue. This paper analyzes the application scenarios and requirements of scientific big data, and expounds on the challenges in four aspects: scale dynamism, pipeline management, unified access, and data sharing (SPUS). It proposes an architecture for scientific big data management systems consisting of four modules: computing and storage management, data pipeline management, data fusion query management, and data sharing management, and analyzes the key technical issues within the system. Finally, it introduces the research progress and future research directions of the National Key R&D Program project "Scientific Big Data Management System".

Full Text

Technology and Methodology

ChinaXiv Cooperative Journal: Scientific Big Data Management Technology and System
Chinese Academy of Sciences Computer Network Information Center, University of Chinese Academy of Sciences, Chinese Academy of Sciences

As modern scientific discoveries increasingly rely on the analysis and processing of large-scale scientific data, efficient management of scientific big data has become an urgent challenge. This paper analyzes the application scenarios and requirements of scientific big data, elaborates on four key challenges in scientific big data management—Scale Dynamic, Pipeline Management, Unified Access, and Sharing management (SPUS)—and proposes a system architecture comprising four modules: computing and storage management, data pipeline management, data fusion query management, and data sharing management. Key technical issues within the system are analyzed, and the research progress

and future directions of the National Key R&D Program project "Scientific Big Data Management System" are introduced.

Keywords: scientific big data, integrated query, pipeline, data sharing, elastic scaling
DOI: 10.16418/j.issn.1000-3045.2018.08.005

Abstract

As modern scientific discoveries heavily depend on the analysis and processing of large-scale scientific data, efficient management of scientific big data has become an urgent challenge. This paper first introduces the application scenarios and requirements of scientific big data. It then summarizes four key challenges in scientific big data management (SPUS): Scale Dynamic, Pipeline Management, Unified Access, and Sharing Management. Following this, we present a proposed scientific big data management system consisting of four components: computing and storage management, data processing management, data fusion management, and data sharing management, and analyze the key technical issues within the proposed system. Finally, we introduce the ongoing Big Scientific Data Management System (BigSDMS) program, a national key research and development initiative.

Author Information

LI Jianhui
Professor, Computer Network Information Center, Chinese Academy of Sciences (CAS). Executive Member of CODATA and Director of the Beijing Engineering Laboratory for Big Data Application Service Technologies. Prof. Li has long been dedicated to promoting scientific data openness, sharing, and application services. His major responsibilities include constructing the CAS Data Cloud Service System, developing cloud service platforms, and innovating data-intensive scientific applications. His current research focuses on big data resource sharing, management technologies, and computing and analysis techniques.
E-mail: lijh@cnic.cn

LI Yuepeng
(Affiliation information not fully provided in the source text)

WANG Huajin
(Affiliation information not fully provided in the source text)

CHEN Mingqi
Deputy Director of the Informatization Office and Director of the Informatization Division, General Office, Chinese Academy of Sciences (CAS). He received

his doctorate in Information and Signal Processing from Beijing University of Posts and Telecommunications in 2000. Since 2007, he has been responsible for planning and implementing CAS information system construction, as well as daily operation and maintenance. He coordinates and supports the acquisition, development, and provision of state-of-the-art information infrastructure resources, tools, services, and application systems essential to CAS. He is a member of the Editorial Board of *China's Blue Book on Scientific Research Informatization*. His main research areas include information technology strategy, scientific research informatization, network and information security, and signal and information processing.
E-mail: mqchen@cashq.ac.cn

Corresponding author

---

1. Introduction

Jim Gray [1] proposed the fourth paradigm of scientific research—data-intensive scientific discovery—arguing that massive data would become a primary driver of future scientific breakthroughs. On July 4, 2012, the European Organization for Nuclear Research (CERN) announced the discovery of the "God particle" by analyzing two years of Large Hadron Collider (LHC) experimental data; the following year, the particle's predictors were awarded the Nobel Prize in Physics. The Laser Interferometer Gravitational-Wave Observatory (LIGO) Scientific Collaboration announced the first detection of gravitational waves on February 11, 2016, after accumulating 500 PB of data and 14 years of model and system improvements, confirming the final prediction of general relativity; in 2017, three key LIGO contributors received the Nobel Prize in Physics.

Today's major scientific research facilities and projects—including the Large Synoptic Survey Telescope (LSST) in astronomy, the LHC in high-energy physics, the Human Genome Project (HGP) in life sciences, and the Integrated Research on Disaster Risk (IRDR) in geoscience—all continuously collect data from large scientific instruments or observation devices, then achieve scientific discoveries through data analysis. By around 2020, LSST will be fully operational, completing one sky survey every three days and generating 15 TB of data daily for research objectives such as new star discovery and dark matter detection. The Square Kilometre Array (SKA) will generate 200 GB of raw data per second, require exascale computing, and demand data transmission speeds ten times faster than the current Internet—all awaiting breakthroughs and challenges for researchers. These large-scale scientific projects are crucial for understanding the origin of the universe, discovering natural laws, and driving technological innovation. The ability to effectively manage, process, and utilize this data will be a key factor in determining whether China can achieve international leadership in science and technology in the new era.

2. Scientific Big Data Application Scenarios and Management Requirements

2.1 Scientific Big Data Application Scenarios and Typical Characteristics

Scientific data serves as the input, output, and asset of research activities, forming the foundation for verifying or falsifying scientific discoveries through facts, evidence, and reasoning. It includes digitized observations and measurements from instruments or sensors, computational simulation and model outputs, descriptions of scenarios or phenomena, behavioral observations or qualitative descriptions, and statistical data for management or commercial purposes [2]. Scientific big data now pervades research across all fields, particularly prominent in "big science" domains such as astronomy, high-energy physics, and microbiology [3].

In astronomy, the Ground-based Wide Angle Camera (GWAC)—a key ground-based instrument for the Sino-French SVOM gamma-ray burst detection satellite—generates 32 MB sky images per camera every 15 seconds. It must complete point source extraction and cross-identification before the next image arrives, ultimately performing insertions of 1-10 million star catalog entries and JOIN operations on 1-10 billion star catalog entries within 3-5 seconds [4].

In high-energy physics, the LHC built by CERN performs 600 million collision experiments per second, generating 6 PB of event data. After event filtering, approximately 1 GB of experimental data is stored. Currently, LHC experimental data exceeds 200 PB, and will surpass 1 EB within five years, with event counts reaching quadrillion levels—requiring the selection of one in a million events within 10 seconds [5].

In microbiology, the World Data Center for Microorganisms (WDCM) at the Chinese Academy of Sciences' Institute of Microbiology performs entity recognition, disambiguation, and ontology construction on 36 data sources including Taxonomy, GenBank, and Gene, building a knowledge graph structure containing 8.3 million nodes and 130 million edges. Over the next five years, WDCM will aggregate open biological, resource, literature, sequence, and disease data from over 10,000 sources to construct a knowledge graph with 10 billion connections, requiring six-step associative queries on 10 billion connected records within one second.

These characteristics of scientific big data pose enormous challenges to data management systems. The accuracy of scientific discoveries depends on repeated computational verification of massive experimental data. For instance, the discoveries of the "God particle" and dark matter required multiple computations on hundreds of petabytes of data, with conclusions only published after repeated verification. Scientific experiments generate large volumes of observational data in short timeframes for pipeline processing, with data continuously entering long-term persistent storage devices. Scientific phenomena are quan-

tified through diverse metrics in forms such as images, audio, and time series, distributed across different countries and institutions, requiring integration of multi-source heterogeneous data. Scientific data originates from large scientific facilities, the Internet, and national agencies, relating to national interests and personal privacy, making data sharing and mining crucial for societal advancement while raising complex governance issues.

2.2 Challenges in Scientific Big Data Management

Scientific big data exhibits the common "4V" characteristics—volume, velocity, variety, and value—but also possesses unique properties that present four major challenges throughout its lifecycle (SPUS):

(1) Scale Dynamic. Scientific experiments continuously generate massive data requiring long-term persistent storage. As mentioned, most scientific research projects (GWAC, LHC, etc.) produce gigabytes of data per second with no expiration date. However, research institutions cannot predetermine optimal storage and computing resource configurations to meet scientific application demands. Therefore, elastically and dynamically allocating storage space and processing resources for this data represents a major challenge.

(2) Pipeline Management. Scientific experiments follow rigorous procedures, with massive raw scientific data from instruments undergoing extensive feature extraction, transformation, and analysis operations before yielding research outcomes. Using GWAC's new star discovery application as an example, raw scientific data entering the system requires completion of feature extraction, cross-identification, and other rigorous processing operations. When a new star alert occurs, the system must trace back to the characteristic records, sky images, and lenses that generated the alert for repeated verification. Additionally, numerous similar experimental workflows exist under the same scientific instrument. Effectively creating, executing, and managing these experimental steps and data will greatly improve scientific research efficiency.

(3) Unified Access. Large-scale scientific applications frequently perform fusion mining and analysis on heterogeneous data from different domains and institutions. The Digital Belt and Road (DBAR) international scientific program initiated by Chinese scientists, for instance, requires integrated Earth big data platforms combining satellite remote sensing data, climate observation station data, biological observation station data, and public opinion hotspot data from social networks to provide decision-making references for regional and national policies. Therefore, unified access to multi-source heterogeneous data will significantly enhance the value and scale of scientific discoveries.

(4) Sharing Management. Scientific experimental results and intermediate data are openly shared through the Internet to harness the collective power of scientists worldwide for experimental verification, model improvement, and subsequent research. For example, physicists globally access data from the LHC via the Internet and share research outcomes. However, open scientific data

raises critical issues: how to fairly divide research achievements between data providers and researchers, how to authenticate data provider copyrights, how to establish incentive mechanisms, and how to protect privacy in shared data. Without proper solutions, these issues will affect researcher enthusiasm and the health of the scientific ecosystem.

3. Scientific Big Data Management System Architecture

The scientific big data management system consists of four core components: computing and storage management, data pipeline management, data fusion query management, and data sharing management [Figure 2: see original paper].

(1) Computing and Storage Management Component. This component enables elastic scaling of computing and storage resources with changes in upper-layer application load to optimize the ratio of processing time to resource investment. Current elastic scaling approaches include progressive and quantitative methods. Progressive scaling monitors resource contention from upper-layer applications on underlying computing and storage resources, dynamically adding or reducing resources. For example, on AWS E-MapReduce clusters, MapReduce job resource contention is measured by remaining available cluster memory, with automatic cluster expansion triggered when contention exceeds thresholds. Quantitative scaling estimates target application resource requirements in advance to determine application scale. Compared with progressive scaling, quantitative scaling has shorter reaction times but heavily depends on accurate estimation of target application computing and storage needs [7], such as through workload modeling. In practice, hybrid approaches combining both methods are often employed.

(2) Data Pipeline Management Component. By abstracting data processing workflows into logical processing units within a pipeline, this component enables standardized and unified management of data processing. Typically, one processing unit in a pipeline represents a function, WebService, or SQL statement , with outputs serving as inputs to one or more subsequent units. Through branching, looping, and other constructs, these processing units are assembled to complete scientific discovery workflows. Pipeline management shares formal representations with workflows and instruction streams, such as Pi-calculus and Petri nets [8], which theoretically guarantee execution accuracy and enable exception handling. Beyond ensuring correct pipeline operation, pipeline management must address core issues including data ingestion, data provenance, and intermediate data transformation. Common pipeline management tools include Apache NiFi and StreamSets.

(3) Data Fusion Query Management Component. This component provides unified interfaces for querying and analyzing multi-source heterogeneous data. Current data fusion approaches include federated databases, multi-model databases, polystore databases, and data integration [9]. Federated databases transparently map data from multiple autonomous heterogeneous or homoge-

neous databases into a global view, featuring autonomy, heterogeneity, and distribution—such as the Federate function in SQL Server 2000 and MySQL 5.0. Multi-model databases store multiple data types in a single database backend, like OrientDB and ArangoDB. Polystore architectures lack a unified global view, instead comprising local and intermediate views queried through unified languages, with typical examples including BigDAWG and Myria. Data integration can be categorized as online or offline based on data transformation methods. Offline integration converts data from different sources via ETL and stores it in a global view for unified management and analysis, such as data warehouses, data lakes, and DataHub. Online integration parses query statements to convert local view data to global view in real-time, as seen in Spark SQL, Impala, and Presto [10-13].

(4) Data Sharing Management Component. This component's fundamental task is to facilitate the flow, dissemination, and reuse of data resources between owners and users. Current research on scientific data sharing mechanisms focuses on four aspects: data submission mechanisms, data publishing mechanisms, data alliance mechanisms, and service incentive mechanisms (point systems, online computing service models). Researchers such as Wang Qing [14] and Li Chengzan et al. [15] have conducted in-depth analyses of data sharing mechanisms from policy, technical, and evaluation perspectives. Among privacy protection technologies for data sharing, blockchain is most representative. Ding Wei et al. [16] and Weng Jian et al. [17] have proposed blockchain-based data sharing methods that store data on blockchains through asymmetric encryption algorithms (public-private key pairs), providing greater privacy protection validated in healthcare and genomics domains.

4. Scientific Big Data Management System Project Progress

Supported by the National Key R&D Program "Scientific Big Data Management System" and the CAS 13th Five-Year Informatization Plan "Scientific Big Data Engineering" project, we have collaborated with over 20 research institutions in computer science, astronomy, high-energy physics, and microbiology to explore scientific big data management. We have developed the Big Scientific Data Management System (BigSDMS), with core components including scientific big data management engines, system integration, and application demonstrations [Figure 3: see original paper].

4.1 Scientific Big Data Management Engines

BigSDMS includes three types of management engines: large-scale graph data management, large-scale semi-structured data management, and large-scale relational data management. The large-scale graph database Gstore supports 10 billion triples, graph data management, and second-level query response times. The large-scale semi-structured database Eventdb supports trillion-level high-

energy physics events and EB-scale data management. The large-scale relational database AstroServer supports management of 100 billion rows of astronomical catalog data, with query optimization for typical operations at large, medium, and small scales while meeting data processing precision and real-time requirements. These three databases fundamentally satisfy current large-scale data management needs in common scientific experiments.

4.2 Scientific Big Data System Integration

BigSDMS integration comprises four parts: Elastic Deployment (EMR), Pipeline (Piflow), Fusion Query (Simba), and Data Sharing (Pishare). EMR's elastic scaling combines progressive and quantitative approaches: when workload model confidence is below threshold, it uses progressive scaling and refines the model based on post-expansion resource contention; when confidence reaches threshold, it switches to quantitative scaling. Piflow, based on Petri nets, manages processing units that transition between unknown, active, and hibernated states to execute and monitor workflows. Simba, built on Spark SQL, enables fusion query analysis of multiple data sources through SQL in the Zeppelin visualization interface. Pishare, based on the Hyperledger blockchain platform, encrypts and certifies data ownership on-chain, rewards data providers through a point mechanism ("Science Coin"), and facilitates data marketplace transactions.

4.3 Scientific Big Data Application Demonstrations

Based on BigSDMS, we have constructed three application demonstrations:

1. **Astronomy:** Using 10 billion star catalog entries, we defined five light curve processing workflows, achieving insertion of 6.8 million catalog records in under 3 seconds and "anomaly discovery" in less than 1 second [Figure 4a: see original paper].
2. **High-Energy Physics:** Using 94.29 billion event records from BESIII, query efficiency improved by over 10 times compared to the industry-standard Boss system [Figure 4b: see original paper].
3. **Microbiology:** Integrating information on 200 microbial species, we constructed a 500 million-triple RDF knowledge graph [Figure 4c: see original paper].

5. Conclusion and Future Directions

As humanity's understanding of the objective world deepens, an increasing number of social and natural phenomena can be quantified through observation devices, leading to continuous growth in scientific data volume and variety. Under the data-driven scientific discovery model, addressing the SPUS challenges in scientific big data management has become an urgent task. The National Key R&D Program "Scientific Big Data Management System," led

by the CAS Computer Network Information Center, has deeply explored these issues and developed the BigSDMS system. Future research will further investigate four aspects: elastic deployment, pipeline management, data fusion, and data publication/sharing—including quantification and prediction of resource contention, intermediate data model design for pipelines, polystore integration of multiple query engines, and optimization of data sharing mechanisms. As research on scientific big data management technologies and systems continues to advance, the contribution of scientific big data to scientific discovery will grow ever greater!

References

1. Hey T, Tansley S, Tolle K M. The Fourth Paradigm: Data-intensive Scientific Discovery. Redmond, WA: Microsoft Research, 2009.
2. *(Reference incomplete in source)*
3. *(Reference incomplete in source)*
4. Yang C, Weng Z, Meng X, et al. Astronomy big data challenges and real-time processing technology. Journal of Computer Research and Development, 2017, 54(2): 248-257.
5. Cheng Y, Zhang X, Wang P, et al. High-energy physics big data challenges and massive event feature indexing technology research. Journal of Computer Research and Development, 2017, 54(2): 258-266.
6. Ward J S, Barker A. Undefined By Data: A Survey of Big Data Definitions. arXiv, 2013: 1309.5821v1.
7. Li H, Shi M. Workflow models and their formal description. Chinese Journal of Computers, 2003, 26(11): 1456-1463.
8. Wang H, Li J, Shen Z, et al. Approximations and Bounds for (n, k) Fork-Join Queues: A Linear Transformation Approach. 2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID). arXiv, 2017: 1707.08860v7.
9. Lu J, Holubova I. Multi-model Data Management: What's New and What's Next? Extending Database Technology, 2017: 602-605.
10. Sheth A P, Larson J A. Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys (CSUR), 1990, 22(3): 183-236.
11. *(Reference incomplete in source)*
12. Davoudian A, Chen L, Liu M. A Survey on NoSQL Stores. ACM Computing Surveys (CSUR), 2018, 51(2): 40.
13. Gadepally V, Chen P, Duggan J, et al. The BigDAWG polystore system and architecture. IEEE High Performance Extreme Computing Conference, 2016. arXiv: 1609.07548.
14. *(Reference incomplete in source)*
15. Li C, Zhang L, Hou Y, et al. Scientific big data open sharing: models and mechanisms. Information Studies: Theory & Application, 2017, 40(11):

45-51.
16. Ding W, Wang G, Xu A, et al. Research on key technologies and information security issues of energy blockchain. Proceedings of the CSEE, 2018, 38(4): 1026-1034.
17. *(Reference incomplete in source)*

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