Hot Issues in Big Data: Where Is Data Infrastructure Headed?

Authors: Gu Liping, Gu Liping

Date: 2023-09-28T00:00:00+00:00

Abstract

Big data analytics and applications encompass five challenges, including: computational infrastructure, data management practices, researcher preferences, various collaboration opportunities, and technology-concealed costs. This paper proposes several practical and feasible recommendations to address these issues: (1) establish open scientific platforms and data repositories to promote data sharing and exchange; (2) strengthen interdisciplinary collaboration mechanisms to encourage experts from different fields to participate in data analysis and research; (3) formulate clear codes of conduct and standards to ensure data quality and privacy protection; (4) utilize cloud computing technologies and automation tools to improve the efficiency of data processing and analysis; and (5) invest in education in the field of big data, cultivate more talent, and enhance the technical level of the entire industry.

Full Text

The Hot Issue of Big Data: Where Should Data Infrastructure Go?

Gu Liping
1. National Science Library, Chinese Academy of Sciences
2. Department of Information Resource Management, School of Economics and Management, University of Chinese Academy of Sciences

Abstract

Big data analysis and application face five major challenges: computing infrastructure limitations, data management practices, multi-stakeholder collaboration complexities, researcher preference variations, and cost structures obscured by technological complexity. This article proposes practical recommendations to address these issues: (1) establishing open scientific platforms and

data warehouses to facilitate data sharing and communication; (2) strengthening cross-disciplinary collaboration mechanisms and encouraging experts from diverse fields to participate in data analysis and research; (3) developing clear codes of conduct and technical specifications to ensure data quality and privacy protection; (4) leveraging cloud computing technologies and automation tools to improve data processing and analysis efficiency; and (5) investing in big data education to cultivate more talent and elevate the technical capabilities of the entire field.

Keywords: Open Science; Open Data; Data Sharing; Big Data Analysis; Big Data Processing; Big Data Storage

1. Current Status and Challenges of Big Data Infrastructure

The rapid development of big data technologies over the past two decades has transformed scientific research and data management practices. However, significant challenges persist in computing infrastructure and data management. Current infrastructure struggles to accommodate the exponential growth in data volume, with many institutions facing capacity constraints and inefficient resource utilization. The "20-90" phenomenon—where 20% of data resources consume 90% of infrastructure capacity—highlights the uneven distribution and management inefficiencies plaguing the field.

Data management practices remain fragmented across institutions, lacking standardized protocols for data collection, storage, and sharing. This fragmentation creates barriers to collaboration and limits the potential for large-scale, multi-institutional research projects. Furthermore, researcher preferences for specific tools and platforms, combined with rapidly evolving technologies, have obscured the true costs of data infrastructure, making long-term planning and investment decisions increasingly difficult.

2. Data Management and Collaboration Mechanisms

Effective data management requires robust collaboration mechanisms that bridge disciplinary boundaries. Current approaches often suffer from isolated data silos and inconsistent management standards. Establishing cross-disciplinary collaboration frameworks is essential for integrating diverse expertise into data analysis workflows.

Standardization efforts must address both technical specifications and governance models. Technical standards should encompass data formats, metadata schemas, and interoperability protocols, while governance models need to define clear responsibilities for data quality assurance, access control, and privacy protection. The development of comprehensive codes of conduct will help ensure that data sharing practices align with ethical guidelines and regulatory requirements.

3. Data Sharing and Open Science

Data sharing represents a cornerstone of open science initiatives, yet implementation remains inconsistent. The FAIR principles—Findable, Accessible, Interoperable, and Reusable—provide a valuable framework for guiding data management practices. However, translating these principles into operational reality requires substantial investment in both infrastructure and cultural change within research communities.

Open scientific platforms and data warehouses can serve as central hubs for data sharing and communication. These platforms must support diverse data types and provide tools for data discovery, analysis, and visualization. Equally important is the development of community norms that recognize and reward data sharing contributions, addressing the current incentive misalignment that often discourages researchers from sharing their data.

4. Computing Infrastructure and Cloud Technologies

Modern big data infrastructure must leverage cloud computing and automation to achieve scalability and efficiency. Cloud-based solutions offer elastic resource allocation, enabling institutions to handle variable workloads without maintaining expensive on-premises infrastructure. Automation tools can streamline data processing pipelines, reducing manual intervention and improving reproducibility.

The integration of cloud technologies requires careful consideration of data security, privacy, and compliance requirements. Institutions must develop clear policies for data residency, access control, and encryption while ensuring that cloud-based workflows maintain the same rigorous standards as traditional infrastructure. Investment in automation should focus on reducing repetitive tasks and enabling researchers to concentrate on higher-level analytical work.

5. Education and Talent Development

The shortage of skilled big data professionals represents a critical bottleneck for the field. Educational institutions must expand curricula to include data science, computational methods, and domain-specific analytics. This requires not only technical training but also education in data ethics, privacy protection, and responsible research practices.

Professional development programs should target both early-career researchers and established professionals seeking to update their skills. Industry-academia partnerships can provide practical training opportunities and ensure that educational programs align with real-world needs. Building a robust talent pipeline is essential for sustaining long-term innovation in big data infrastructure and applications.

6. Policy and Governance

Effective governance frameworks must balance innovation with responsible data stewardship. Policies should address data ownership, intellectual property rights, and liability issues while promoting open access where appropriate. Governance structures need to be flexible enough to accommodate emerging technologies and evolving research practices.

Institutional policies should clarify roles and responsibilities for data management across the research lifecycle, from initial collection through long-term preservation. This includes establishing clear procedures for data quality assessment, version control, and archival. International coordination is also necessary to harmonize standards and facilitate cross-border data sharing.

7. Future Development Directions

Looking forward, big data infrastructure must evolve toward more integrated, intelligent, and sustainable models. This includes developing adaptive systems that can automatically optimize resource allocation based on workload patterns and implementing advanced analytics capabilities that support real-time decision-making.

Investment priorities should focus on foundational technologies that enable interoperability, such as standardized APIs and data exchange protocols. Equally important is fostering a culture of collaboration and continuous improvement, where infrastructure development is driven by community needs and guided by shared principles of openness, fairness, and sustainability.

References

[1] Anonymous. Big Data Infrastructure and Management Challenges[J]. *Information Science Research*, 2023, 1(35): 39-55.

[2] Dylan Ruediger. Big Data Infrastructure at the Crossroads: Support Needs and University Challenges[EB/OL]. [2021-12-01]. https://doi.org/10.18665/sr.316121

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