Postprint: Research on the Design of Big Data Integration Architecture for Public Cultural Services

Authors: Hua Bolin, Zhao Dong is at, Shen Yongguo

Date: 2023-04-01T16:15:54+00:00

Abstract

[Purpose/Significance] To address the multi-source heterogeneous data from current public cultural service institutions such as libraries and cultural centers, an effective integration architecture is designed. [Method/Process] Based on a thorough analysis of public cultural big data resources, the types and distribution of public cultural service big data are analyzed, and the integration architecture for public cultural big data is designed in conjunction with its application scenarios. [Results/Conclusion] An integration architecture for public cultural service big data composed of five layers—data source layer, system integration layer, data fusion layer, storage layer, and application layer—is proposed, and key technologies such as collection and storage are studied.

Full Text

Abstract

[Purpose/Significance] This study aims to design an effective integration architecture for the multi-source heterogeneous data currently held by public cultural service institutions such as libraries and cultural centers. **[Method/Process]** Based on a comprehensive analysis of public cultural big data resources, we examine the types and distribution patterns of public cultural service big data and propose an integration architecture tailored to its application scenarios. **[Results/Conclusion]** We present a five-layer public cultural service big data integration architecture comprising a data source layer, system integration layer, data fusion layer, storage layer, and application layer, and investigate key technologies for data acquisition and storage.

Keywords: library; cultural center; public culture; big data; data integration; integrated architecture

1. Introduction

China's library system comprises three main sectors: public libraries, university libraries, and specialized libraries. In the specialized library domain, the National Science and Technology Library (NSTL) was established in 2000 to integrate resources from scientific and technical information institutions. Operating under a mechanism of "unified procurement, standardized processing, joint online access, and resource sharing," NSTL collects, preserves, and develops scientific and technical literature across various disciplines, providing public welfare and universal access to scientific information services nationwide [1]. In the university library sector, the China Academic Library & Information System (CALIS) was launched in 1998, creating a distributed "China Higher Education Digital Library" with nearly 2,000 member institutions. This system is structured around CALIS's online cataloging system, resource discovery and delivery system, and collaborative service platform, with provincial digital library platforms and university library systems serving as branches [2]. In contrast to university and specialized libraries, public libraries serve the general populace and have developed more prominent local characteristics, which has historically prevented the formation of a unified national platform in this sector.

In recent years, growing national emphasis on public culture has gradually spurred the development of public cultural service systems, with increasing demand for regional coordination and cross-institutional services. Many cities have implemented municipal-level master-branch library mechanisms, integrating resources and services. Shanghai and other cities have introduced all-in-one card services that, beyond transportation and medical appointment functions, enable unified authentication and access across libraries, cultural centers, and other public cultural institutions. The National Public Culture Cloud platform already hosts over a hundred cultural service institutions providing resources, activities, and services. Platforms such as Zhejiang Jiaxing's "Cultural Appointment" and Shanghai's Cultural Cloud have developed rapidly. The "Cultural Jiading Cloud" intensively aggregates cultural resources and service information from district-level libraries, cultural centers, museums, art galleries, and street-level community service centers, offering the public comprehensive, one-stop, and equitable remote digital reading, virtual venue experiences, access to specialized resources, cultural event announcements, public facility reservations, and online exchange and exhibition services through websites, mobile apps, WeChat, and Weibo service clusters [3].

As online platforms for public cultural services continue to emerge, the volume of digital resources is growing constantly, and the data generated on these platforms is showing increasing diversification. Integrating and analyzing this data offers broad application prospects, from reader recommendations to "you-select-books-we-purchase" programs, from circulation rankings and visitation statistics to big data dashboards, from self-service borrowing to robotic inventory management, from online streaming to cultural clouds, and from interlibrary loans to culture-tourism integration. These applications effectively connect multi-source,

heterogeneous, and even cross-regional resource data and user data, providing better precision services for users, real-time business monitoring and management for cultural institutions, and more comprehensive situational awareness and decision support for administrative departments. Achieving these applications requires integrating cross-regional, cross-institutional, and cross-platform data, which represents both a prerequisite and a key challenge for the development of big data in public cultural services.

2. Literature Review

Big data has enabled user profiling and precision recommendation, dynamic real-time monitoring and remote surveillance, risk warning and trend forecasting across many fields. Realizing these application scenarios depends on data integration from various domain-specific sources. Big data integration differs from traditional data integration in that, from a data structure perspective, the integration objects are not limited to structured data in databases but also include semi-structured and unstructured data such as log data, image data, video data, and voice data. As information systems proliferate and data collection diversifies, data integration has become a key constraint on big data mining and utilization across industries, serving as the foundation for big data organization, analysis, and mining.

2.1 Domain-Specific Big Data Integration Research

2.1.1 Natural Science Domain Big data integration originated in fields such as signal processing, remote sensing monitoring, and industrial automation. Multi-source geospatial big data provides unprecedented social sensing capabilities for understanding the distribution patterns, interactions, and dynamic evolution of geographic phenomena [4]. Wang Juanle et al. proposed an integration and standardization framework for earth big data that leverages network big data, remote sensing big data, and socio-economic big data, analyzing key technologies for network data acquisition and analysis, intelligent extraction and processing of remote sensing data, and spatialization of socio-economic data [5]. Zhao Fen et al. elaborated on four modules for ecological environment big data technology platforms: data acquisition, storage, computing patterns/systems, and analysis [6]. Lv Youlong et al. proposed a smart factory technical architecture comprising five layers: object interconnection, object perception, data analysis, business application, and cloud services, plus a big data center [7]. Li Shaobo et al. identified five key technologies for manufacturing under big data: data integration, storage, processing, analysis, and visualization [8]. Wang Song et al. suggested that future data integration research will focus on algorithm acceleration, integration of complex data sources, and crowdsourcing-based methods [9]. Notably, data sources in natural sciences primarily comprise “hard data” transmitted from hardware devices such as sensors, remote sensing, telemetry, and satellites. Data integration serves as the prerequisite and foundation for big data analysis and visualization.

2.1.2 Smart City Big Data Integration Beyond hard data from sensors, smart cities also generate management and social data. In the government information domain, expanding dataset scales and divergent departmental informatization processes have created information silos that hinder full integration and sharing. Scholars have extensively explored this issue. Ye Xin et al. argued that "Internet Plus Government Services" cloud platforms based on big data and knowledge can help eliminate information, knowledge, and business silos [10]. Yang Xingkai et al. reviewed integration methods in government information, noting that research on standardization for e-government information resource integration remains limited, slowing development in standardized data and business model construction [11]. Hong Zhixu et al. proposed a distributed data integration and visualization method that connects, extracts, and integrates data from databases across different network routes based on big data processing patterns, enhancing dynamic description and Web visualization capabilities to provide service-oriented intelligent social governance decision analysis [12]. Liu Yan et al. designed a heterogeneous data source integration system architecture centered on Hadoop by building a big data center [13].

2.1.3 Intelligence Big Data Integration Intelligence agencies share strong similarities with public cultural institutions in data resources, business processes, and service functions, making their big data integration experiences highly relevant. Tang Mingwei et al. categorized data integration theories into heterogeneous data theory and system integration theory, proposing a big-data-oriented intelligence analysis framework where the big data cluster layer—comprising intelligence resources, computing clusters, and application pools—forms the core for handling big data applications [14]. Ba Zhichao et al. described ubiquitous collaborative perception and acquisition of elements or data from the physical world and human society, mapping them into information space for ordered organization, information fusion, and integrated analysis to guide decision-making in human society and the physical world [15]. Lu Xiaobin et al. proposed a general big data analysis system architecture for bank risk management, aiming to integrate different data types into a unified, standardized, and user-friendly system [16]. Chen Wei et al. reviewed advances in massive heterogeneous data integration and data management/analysis methods and tools, proposing an overall architecture for building a data-driven scientific intelligence research model [17].

2.2 Public Culture Big Data Research

2.2.1 Theoretical Discussion Early discussions on combining big data with public culture began in library research. In 2012, Han Cuifeng recognized big data's impact on libraries, noting that it would demand greater resource storage capacity and user needs mining capability, requiring changes in technology development, data integration, and talent management [18]. Ji Ting et al. categorized public culture big data into business, network, and management data, exploring its acquisition, storage, and analysis methods [19]. Su Xinning en-

visioned future digital library development from three perspectives: resource construction, technology application, and services [20]. Liu Wei et al. examined top-level design for public cultural service big data development, addressing policy and macro-management, industrial chains, industry ecology, and technical standards [21]. These studies demonstrate the necessity of combining big data with public cultural services, exploring public culture big data from different perspectives and establishing preliminary theories that provide theoretical support for application development.

2.2.2 System Research Building on theoretical guidance, researchers have designed public culture big data systems. J. Li et al. discussed big data applications in libraries from five aspects: human resources, literature resources, technical support, service innovation, and infrastructure [22]. Cao Shujin et al. proposed a library big data system for precision services, comprising four bottom-up layers: multi-source data acquisition, data preprocessing and storage, precision data analysis modeling, and precision management and service application [23]. Guo Lusheng et al. conducted top-level design for public culture big data application systems based on Enterprise Architecture (EA), aligning application architecture, IT architecture, and governance architecture with strategic objectives [24]. Zhang Chunjing categorized public culture service big data application models into three driver types: data-driven, cloud platform-driven, and holistic-driven [25].

2.2.3 Integration Research Beyond system frameworks, some scholars have specifically addressed data integration or technical platform implementation. Li Guangjian et al. argued that public culture service big data research should focus on conceptual boundaries, methodologies, data integration, user profiling, precision services, and development strategies [26]. Liu Shuang et al. proposed that integrated library information systems should interconnect library operating information systems (LOIS), library management information systems (LMIS), and library service information systems (LSIS) [27]. Cao Jian et al. introduced a Hadoop-based big data analysis system for university library digital resources, including five functional modules: basic data integration, user tagging, resource analysis, business analysis, and comprehensive system management [28]. Libraries' data-intensive nature, widespread unstructured data distribution, and demand for service precision make big data integration increasingly urgent.

2.3 Research Review

Comparing domains horizontally, big data integration research in remote sensing monitoring, industrial manufacturing, agricultural ecology, and smart cities is already well-developed, offering valuable references for public culture big data integration. In contrast, public culture big data integration research remains in its infancy.

Within the public culture domain itself, awareness of big data is relatively ma-

ture. From business development, national mandates, and user demand perspectives, public culture big data development faces favorable opportunities and challenges, with growing research activity. Overall, theoretical studies dominate while practical implementation research remains insufficient. Although multiple studies mention data integration issues, specialized discussion on how to achieve multi-source heterogeneous data integration is lacking. Therefore, this paper analyzes public culture big data resources comprehensively and designs an integration architecture tailored to its application scenarios, examining key implementation technologies.

3. Public Culture Service Big Data Resource Analysis

Different domains possess distinct data resources whose distribution patterns, structures, and types determine integration approaches. We must identify whether data is structured, semi-structured, or unstructured to determine appropriate acquisition and storage technologies.

3.1 Integration Objects

3.1.1 Data Sources Public culture service big data integration targets data generated by libraries, cultural centers, museums, art galleries, memorial halls, and mass art centers. The core includes resource data, user data, staff data, management data, service data, business data, and their relationships [29]. From the Ministry of Culture and Tourism Key Laboratory perspective, data comprises open data, system data, base-processed data, and public culture cloud data. Open data refers to service data obtained via web crawling from libraries and cultural centers, plus business data extracted from annual reports. System data resides in relational database systems (SQL Server, Oracle, MySQL, Sybase) of cultural service institutions. Base-processed data includes statistical data from libraries, cultural centers, and stations, transmitted through reporting or file transfers. Public culture cloud data encompasses basic data, resource catalog data, resource content data, user data, and activity data.

3.1.2 Data Classification Since data structures and processing models vary across sources, we must clearly distinguish data types. Structured data has explicit, unified structures, primarily from relational databases. Semi-structured data carries certain markers and forms specific structures, such as XML or JSON formats. Unstructured data lacks explicit structures, mainly stored as documents, images, audio, and video files.

System data comprises structured data from relational databases of service institutions' portals, management systems, and business systems. Cultural cloud data contains both structured and unstructured data: structured data includes basic data, resource catalog data, user profile data, and activity data; semi-structured data primarily refers to XML or JSON log data; unstructured data includes activity notification texts, user comment texts, and resource con-

tent videos on cultural clouds, plus self-media data (posts, Weibo, WeChat), PDF/Word documents (papers, annual reports, research reports) from institutional websites, and other free-text web data. Table 1 shows the classification structure of public culture service integration objects:

Table 1. Classification of Public Culture Service Big Data

Data Type	Subcategory	Examples
Structured Data	System data	Relational databases from portals, management systems, business systems
Semi-structured Data	Log data	XML/JSON logs from cultural cloud platforms
	Network data	Marked data like MARC records, metadata-tagged bibliographic data, base-reported data
Unstructured Data	Document data	PDF/Word papers, annual reports, research reports
	Self-media data	Posts, Weibo, WeChat content
	Website data	Activity notifications, user comments, video content from cultural clouds

3.2 Integration Challenges

To effectively address data integration, we must first understand its root causes. Li Kang et al. identified three main difficulties: heterogeneity, distribution, and autonomy [30]. Heterogeneity refers to differences in management environments, data models, expression methods, and semantics across sources. Challenges include system integration issues and data integration issues.

3.2.1 System Integration Issues Building a public culture service big data integration platform requires seamless communication between different source systems. Even when all systems run on the same hardware platform and use ODBC/JDBC-compliant relational databases supporting SQL standards, problems persist. Although SQL is a standard query language for relational

databases, different public culture service platforms implement it differently, requiring reconciliation during integration. Data sources are not limited to relational databases but also include web, text, CSV, and JSON document sources that resist complex querying [31].

3.2.2 Data Integration Semantic Issues Multi-source data heterogeneity poses a significant challenge, and effectively addressing it is key to ensuring integration quality. Issues include semantic ambiguity, instance representation ambiguity, data inconsistency, redundancy, and missing data.

(1) **Semantic Ambiguity**: This includes cases where different names represent identical content, or the same name represents different meanings. Field names for identical data may differ across institutional databases—for example, "visits" in a library database versus "station visits" in a cultural center database. Statistical applications require appropriate field name unification. Conversely, different sources may use identical field names for different meanings. When integrating an online reference system with a "One Person, One Art Cloud Platform," both have a "title" field, but one refers to the user's query title while the other refers to activity names. Such issues can be resolved through metadata mapping.

(2) **Instance Representation Ambiguity**: Different base systems represent the same entity differently—for instance, user click-through rates expressed as percentages in one system and as counts in another. Both are numeric but create representation ambiguity. Inconsistent time/date formats also fall into this category. These require conversion rules to standardize formats.

(3) **Data Inconsistency**: Causes include synchronization issues, multiple classifications, statistical 口径 (caliber) differences, calculation errors, input errors, and outdated information. Different institutions may describe the same event's time or venue inconsistently, with one instance likely being inaccurate. Annual report descriptions may contradict webpage content. Institutions may also categorize the same lecture differently (e.g., as "cultural" versus "lifestyle"). Unlike instance representation ambiguity, inconsistency stems from different data values, often due to synchronization problems. In distributed big data integration, semantic conflicts in attribute features include character-type, numeric-type, Boolean-type, and interval-value data [32]. Conflict resolution can be achieved through syntactic fusion, logic tree fusion, and frequency fusion methods [33].

(4) **Data Redundancy**: Public culture service big data comprises digital collection resources and non-collection resources (lectures, exhibitions, activities, user participation data). With numerous institutions having varying data understanding and technical capabilities, integration easily creates redundancy, duplication, and errors. Redundancy includes complete duplication (identical fields), inclusion relationships (one dataset containing

another), and partial duplication (some identical fields). Redundancy issues are generally resolved by keeping the larger dataset.

(5) **Data Missing**: Missing data may result not from multiple sources but from human error, data loss, or collection difficulties, leading to incomplete data or instances lacking certain attributes. During data cleaning, several methods can address this: manual filling, global constant filling, attribute central tendency filling, most probable value filling (regression, Bayesian methods, or decision tree induction), or tuple ignoring [34].

4. Public Culture Service Big Data Integration Design

Mainstream data integration patterns include federated databases, data warehouses (replication architecture), middleware, and ontology-based integration. For multi-source heterogeneous data characteristics and the public culture service domain, we designed an integration architecture covering overall processes and key technologies.

4.1 Integration Architecture

Through comprehensive analysis of public culture institution data and investigation of integration methods, we developed a solution and architecture comprising five layers: data source, system integration, data fusion, storage, and application. The process includes source acquisition, transmission/collection, problem domain analysis, processing, storage, and application, as shown in Figure 1 [Figure 1: see original paper].

The **data source layer** includes four categories: open data, system data, base-processed data, and public culture cloud data. By type, these comprise real-time data, internet data, business data, and log data.

The **system integration layer** handles data transmission, using different import tools for various metadata and structures. Real-time data flows through distributed message queues via Kafka; relational databases use Sqoop or ETL tools to import directly into HDFS; high-security user data and offline data use hardware replication or FTP; log text data uses Flume; internet data is crawled and imported. Integration encounters issues like inconsistency, field name mismatches, redundancy, missing data, and noise.

The **data fusion layer** employs methods including similarity coefficient calculation, metadata processing, correlation analysis, regression, Bayesian discrimination, binning, clustering, and human-computer hybrid approaches. Similarity coefficients detect and unify inconsistent data; metadata processing standardizes field names; redundancy is addressed through Pearson correlation or deduplication (for fields) and compression (for images); missing data is handled via regression and Bayesian methods; noise data is managed through binning, regression, clustering, or hybrid methods.

The **storage layer** stores classified and preprocessed data in distributed file system HDFS, distributed database HBase, or relational database MySQL according to specific needs. HBase suits massive semi-structured data like logs; Oracle and MySQL store structured data; HDFS stores semi-structured data, FTP-transferred documents, and crawled web data.

The **application layer** provides unified authentication, data management, personalized recommendations, venue/activity reservations, digital resource retrieval, cultural resource services, and information management, enabling real-time monitoring, dynamic management, precision services, and decision support.

4.2 Key Integration Technologies

Integration prerequisite is collecting diverse-source, diverse-structure data for deep analysis and mining to enhance value. The Hadoop-based framework involves key technologies for data acquisition, storage, and analysis, as shown in Figure 2 [Figure 2: see original paper].

Data Acquisition: Different tools extract data based on structure and timing. Structured data from base systems and cultural clouds (e.g., "One Person, One Art Cloud Platform," OPAC, OAuth platforms) uses Sqoop for batch migration between structured data and Hadoop. Log data from various institutions (Nanjing Library, Fujian Library, Chongqing Art Museum, Ningbo Cultural Center, Shanghai Library) uses Flume for serialized semi-structured data collection. For high real-time requirements, Kafka and Flume are combined: business data is first stored in Kafka clusters in real-time, then Flume's Source component processes Kafka Topics and Flume Sink sends consumed data to HDFS or HBase [35]. This leverages Kafka's real-time collection and Flume's efficient HDFS writing. Document data from base reporting uses FTP (via FlashFXP, FileZila, Cuteftp). Institutional, service, dynamic, and self-media data uses Scrapy crawlers.

Storage: The repository includes HDFS, real-time HBase, and relational Oracle/MySQL databases. HBase stores massive semi-structured data like Flume-collected logs. Oracle/MySQL stores structured data. HDFS stores semi-structured data, FTP-transferred documents, and crawled web data.

Data Processing: Beyond basic extraction and statistical algorithms, processing includes algorithms for converting semi-structured/unstructured to structured data and deep content understanding, involving natural language processing, video/image comprehension, and text mining. Processing quality directly determines statistical analysis accuracy and user experience [13]. For real-time requirements, use stream processing like Storm; for large-scale offline analysis, use batch processing like MapReduce; for integrated real-time analysis and deep mining, use hybrid frameworks like Spark or Impala for high-performance analysis.

Data Sources: Include relational databases (MySQL, Oracle, SQL Server) and

log/text sources (XML, Excel).

Applications: Results include personalized recommendations, log analysis, data management, and unified authentication, enabling real-time monitoring and decision support.

Conclusion

This paper identified key issues in public culture service big data integration and designed an architecture addressing them. Since public culture big data includes not only structured data but also unstructured images, videos, and comments, traditional integration methods are inadequate. We designed a Hadoop-based architecture tailored to diverse sources and structures. Different acquisition and storage technologies are needed for different data models, and the data fusion layer faces logical issues like inconsistency, field mismatches, redundancy, missing data, and noise. Solutions include similarity coefficients, metadata processing, correlation/regression analysis, Bayesian methods, binning, clustering, and hybrid approaches.

While this architecture addresses technical challenges, multi-source heterogeneous data integration also involves personnel, management, and interests—institutions are often reluctant to share data, a problem evident in smart city development and equally relevant to public culture. Internet companies excel at data integration, having reaped substantial benefits and built data middle platforms that enable rapid response to environmental changes and market demands, offering valuable lessons for the public culture sector. With growing public demand for cultural services and continued institutional efforts, public culture service big data integration will gradually gain attention and achieve systematic implementation.

References

[1] National Science and Technology Library [EB/OL]. [2020-02-17]. https://www.nstl.gov.cn/Portal/zzjg_{jgjj}.html. [2] China Academic Library & Information System [EB/OL]. [2020-02-17]. http://www.calis.edu.cn/pages/list.html?id=6e1b4169-ddf5-4c3a-841f-e74cea0579a0. [3] Shanghai Jiading Government Website [EB/OL]. [2020-02-17]. http://www.jiading.gov.cn/zwpd/zwdt/content_{34571}. [4] Liu Yu, Zhan Zhaohui, Zhu Di, et al. Integrating multi-source geographic big data to perceive urban spatial differentiation patterns [J]. Geomatics and Information Science of Wuhan University, 2018, 43(3): 327-335. [5] Wang Juanle, Cheng Kai, Bian Lingling, et al. Earth big data integration framework and key technologies for SDGs and Beautiful China evaluation [J]. Remote Sensing Technology and Application, 2018, 33(5): 775-783. [6] Zhao Fen, Zhang Liyun, Zhao Miaomiao, et al. Preliminary study on ecological environment big data platform architecture and technologies [J]. Chinese Journal of Ecology, 2017, 36(3): 824-832. [7] Lv Youlong, Zhang Jie. Smart factory technical framework based on big data [J]. Computer Integrated Manufacturing Systems,

2016, 22(11): 2691-2697. [8] Li Shaobo, Chen Yongqian. Analysis of key manufacturing technologies under big data environment [J]. Application of Electronic Technique, 2017, 43(2): 18-21, 25. [9] Wang Song, Peng Yuwei, Lan Hai, et al. Development and prospect of data integration methods [J]. Journal of Software, 2020, 31(3): 893-908. [10] Ye Xin, Dong Lu'an, Song Yu. Research on construction and service strategies of "Internet Plus Government Services" cloud platform based on big data and knowledge [J]. Journal of Intelligence, 2018, 37(2): 154-160, 153. [11] Yang Xingkai, Liu Chang. Review of government information resource integration methods [J]. E-Government, 2013(5): 81-87. [12] Hong Zhixu, Chen Hao, Cheng Liang. Social governance data integration and decision analysis methods based on big data [J]. Journal of Tsinghua University (Science and Technology), 2017, 57(3): 264-270. [13] Liu Yan, Wang Hua, Qin Yeyang, et al. Multi-source heterogeneous big data processing framework for smart cities [J]. Big Data Research, 2017, 3(1): 51-60. [14] Tang Mingwei, Su Xinning, Xiao Lianjie. Big-data-oriented intelligence analysis framework [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(5): 467-476. [15] Ba Zhichao, Li Gang, An Lu, et al. Comprehensive information integration of national security big data: Application architecture and implementation path [J]. China Soft Science, 2018(7): 9-20. [16] Lu Xiaobin, Xu Chao. Research on bank big data analysis system architecture for risk management [J]. Journal of Information Resources Management, 2018, 8(2): 4-12. [17] Chen Wei, Yang Rui, He Tao, et al. New models of scientific and technical intelligence research under big data environment [J]. Science & Technology Review, 2018, 36(16): 78-85. [18] Han Cuifeng. Impact and challenges of big data for libraries [J]. Library and Information Service, 2012(5): 37-40. [19] Ji Ting, Wu Zheng. Research on sources, acquisition, and analysis of public cultural service big data [J]. Library Development, 2015(11): 21-24. [20] Su Xinning. Opportunities and challenges for digital libraries in the big data era [J]. Journal of Library Science in China, 2015, 41(6): 4-12. [21] Liu Wei, Zhang Qi, Zhang Yu. Research on big data innovation in public cultural services [J]. Library Development, 2016(3): 4-8. [22] Li J, Lu M, Dou G, et al. Big data application framework and feasibility analysis in library [J]. Information Discovery and Delivery, 2017, 45(4): 161-168. [23] Cao Shujin, Liu Huiyun, Wang Lianxi. Research on library precision services driven by big data [J]. Journal of Academic Libraries, 2019, 37(4): 54-60. [24] Guo Lusheng, Liu Chunnian. Top-level design of public cultural service big data application system based on EA [J]. Library Science Research, 2019(5): 31-37. [25] Zhang Chunjing, Cao Lei, Qu Yun. Research on application models and trends of public cultural service big data [J]. Library Journal, 2015, 34(12): 4-8. [26] Li Guangjian, Hua Bolin. System and content of public cultural service big data research [J]. Library Tribune, 2018, 38(7): 62-71. [27] Liu Shuang, Qian Chengcheng. Research on framework and classification standards for big-data-integrated library information systems [J]. Library Science Research, 2017(5): 31-37. [28] Cao Jian, Qin Ronghuan, Sun Huiqing, et al. Hadoop-based big data analysis system for university library digital resources [J]. Library Work and Study, 2018(3): 74-78, 83. [29] Wen

Tingxiao. Thoughts on library innovation and development in the big data era [J]. Library, 2019(5): 15-22, 27. [30] Li Kang, Li Xinming, Liu Dong. Review of multi-source heterogeneous equipment data integration [J]. Journal of China Academy of Electronics and Information Technology, 2015, 10(2): 162-168. [31] Doan A, Halevy A, Ives Z. Principles of Data Integration [M]. Translated by Meng Xiaofeng, Ma Ruxia, Ma Youzhong, et al. Beijing: Mechanical Industry Press, 2014. [32] Liang Yong. Simulation of distributed big data integration conflict resolution in relational databases [J]. Computer Simulation, 2019, 36(5): 399-402. [33] Wang Yue. Distributed big data integration conflict resolution algorithm in relational databases [J]. Science Technology and Engineering, 2018, 18(3): 63-67. [34] CSDN. Data preprocessing [EB/OL]. [2020-02-17]. https://blog.csdn.net/weixin_{42144636}/article/details/81584372. [35] CSDN. Using Flume to consume Kafka data to HDFS [EB/OL]. [2020-02-17]. https://www.cnblogs.com/smartloli/p/9984140.html.

Author Contributions

Hua Bolin: Research design, literature review, manuscript revision;
Zhao Dongzai: Data collection, partial writing, figure/table preparation;
Shen Yongguo: Data collection, manuscript revision.

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