ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20230630000775
Original scientific paper

Research on Transparent Access Technology of Government Big Data

Honghui LI, Yi XU*, Yunjuan PENG, Xiaorui YANG, Dalin ZHANG, Jianping FAN, Jing SUN

Abstract: E-governance is capable of driving the transition of government from an administrative power allocation of public resources to a citizen-service-oriented governance model. However, during this transformation, various governmental bodies often face the challenge of "data silos" caused by factors such as cross-regional, cross-business, and cross-departmental operations. Without altering the existing information platforms, transparent access technology serves as a key solution for data access within e-governance systems. It enables convenient access to information resources stored in different mediums and formats, facilitating the sharing and consolidation of information and data within and between governmental departments, thereby addressing the issue of "data silos" and enhancing the comprehensive service capabilities of e-governance. This paper firstly provides an overview of the concept, levels, characteristics, and application scenarios of transparent access to government big data. Secondly, it conducts a comprehensive comparative analysis of transparent access technologies in the context of cloud computing and big data. Lastly, based on the requirements of various transparent access technologies and the application of transparent access to government big data, this paper proposes a visionary framework for transparent access to government big data based on cross-domain semantics and channel coupling. This framework includes modules for cross-domain semantic interoperability, coupling of heterogeneous information channels based on knowledge graph, and tracing of multi-source heterogeneous data, aiming to provide innovative solutions for achieving transparent access to government big data.

Keywords: cloud computing; data island; distributed computing; government big data; transparent access

1 INTRODUCTION

With the profound reforms and rapid development of government digital transformation, e-government has gradually become a hot topic in the field of information science [1]. The vast amount of data accumulated by government institutions holds significant value for decision-making and management. However, challenges still exist in terms of open and transparent access to government data, such as data silos. Therefore, researching the technology for transparent access to government big data can effectively promote modernization in government governance, enhance public participation and innovation capability, and drive social progress and economic prosperity [2].

There are various sources of data for government big data, including organizational data, enterprise data, and internet data, amongst others [3]. These data types consist of structured data such as traditional documents, as well as unstructured data such as audio and video. Transparent data access technology refers to the ability to access data regardless of platform differences, allowing for convenient access to information resources stored in different media and formats, enabling flexible combinations and aggregation of data. The purpose of transparent data access technology is to provide users with an interface to access heterogeneous data without the need to understand the underlying storage details. Users can directly retrieve the required data from the heterogeneous data pool according to their needs, thus enabling personalized data support services. In the field of e-governance systems, transparent data access technology can facilitate information sharing and online aggregation among government departments without changing the existing information platforms. It effectively addresses issues such as "government data islands" and "government data divide" [18, 19].

In response to the demand for transparent access to government big data, both domestic and international academic communities, industries, and government departments have accumulated a certain amount of research and practical achievements [4]. For instance, Shanghai has introduced the "One-Net Office" reform, utilizing modern information technology to integrate the responsibilities and authorities of government levels and departments, thereby achieving "network connectivity, data connectivity, and business connectivity." Alibaba has introduced a comprehensive data protection system that promotes transparency in internal operations, making the platform's operational logs related to users visible to ensure a trusted system and application performance. Al-Mushayt, O. S. [45] proposed a government information resource management framework to achieve end-to-end digital government lifecycle management, aiming to enhance the trustworthiness, transparency, and efficiency of digital government. Qi, M. [46] developed a smart digital government platform based on the Internet of Things, leveraging data and information technology to enable real-time sensing, efficient operations, scientific decision-making, proactive services, and intelligent supervision, providing the public with higher quality, more efficient, and more transparent government services.

Thus, despite the notable achievements in the development of governance and transparent access to government data, there are still three crucial issues that require immediate attention: (1) Government data in electronic format is typically stored within specific business platforms and managed by different organizations, resulting in low data sharing levels. Additionally, the storage media, encoding formats, and data channels may vary significantly. Consequently, it is challenging to achieve transparent access to heterogeneous government big data through simple data aggregation methods [24]. (2) Each level of government data management departments has personalized requirements for data reading speed, capacity, and maintenance costs. Meeting the needs of various management users and ensuring dynamic customized data allocation with limited data resources is exceedingly difficult. (3) Current research primarily focuses on data transparency, process transparency, and data reliability, with little consideration given to decision transparency and the issue of cross-domain transparent access to heterogeneous data. It is exceptionally challenging to enhance the quality of e-government services and achieve comprehensive integration of multiple businesses within mixed and heterogeneous government platforms. These formidable challenges severely hinder further improvement of governance capability in leveraging big data. Consequently, it is of utmost significance to explore interdisciplinary and interdepartmental approaches to transparent access to government big data. Furthermore, developing technologies that facilitate cross-domain decision transparency and support the infrastructure for coupled data channels holds paramount importance [5].

In pursuit of addressing the aforementioned challenge, under the support of China's National Key Research and Development Program, the project team led by the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, has undertaken the research project "Internet Plus Government Big Data Transparent Management and Intelligent Service Platform". This project aims to tackle three major scientific issues: the aggregation and integration mechanisms of large-scale cross-disciplinary urban government data, the dynamic evolution and control mechanisms of urban elements, and the theories and methods of urban management and intelligent services. Focusing on methods for transparent access to government big data, the team is conducting research on cross-domain semantic interoperability of government big data, coupling of heterogeneous information channels, and online aggregation of heterogeneous multiple-source government big data. The ultimate goal is to achieve controllable and transparent access to government data and its online aggregation and integration.

The main contributions of this paper can be summarized as follows:

(1) This paper defines the concept of transparent access to government big data for the first time, summarizes the characteristics of government big data, levels of transparency, and application scenarios.

(2) Building upon an analysis of the application of transparent access technology in existing government systems, this paper delves into the underlying technical architecture, summarizing the access technologies in government big data and cloud computing at the levels of serverless computing, organizational distribution, and context awareness.

(3) Based on the current state of research, this paper proposes a visionary framework for transparent access to government big data based on cross-domain semantic integration and channel coupling in a hybrid architecture. This framework encompasses modules for cross-domain semantic interoperability, coupling of heterogeneous information channels, and traceability of heterogeneous data from multiple sources. By standardizing and unifying the expression of business descriptions, it enables the exchange and transformation of different business data, effectively ensuring the coordination and connectivity of business processes across multiple sources, tasks, and platforms in government systems. Additionally, it optimizes the efficiency of transparent access.

The subsequent chapters of this manuscript are outlined as follows: Chapter 2 will provide a comprehensive overview of the concept, characteristics, and tiers of government big data, and systematically review the data transparency access technologies in the context of cloud computing and big data. Chapter 3, after analyzing various transparency access technologies and the practical needs of domestic government big data, will present an envisioned transparency access framework tailored to the future needs of government big data. Finally, Chapter 4 will summarize the work conducted in this paper.

2 AN OVERVIEW OF TRANSPARENT ACCESS TECHNOLOGY FOR GOVERNMENT BIG DATA

2.1 Government Big Data and its Transparent Access Technology

2.1.1 Unique Characteristics of Government Big Data

The government serves as the leader in building digital cities and acts as the cornerstone of urban operations. Therefore, the creation of "digital governance" is particularly significant for the development of digital cities. E-governance refers to the government's use of information technology and network technology to construct its organizational structure, optimizing the public management model of government workflow and services [6]. Based on information technology, e-governance drives the government to shift from administrative power allocation and public resource management to a demand-driven governance model that serves the masses.

The development of e-governance has gradually transitioned from independent business construction to integrated business construction. Government big data refers to the massive data generated by government-related institutions during the implementation of their functions. Compared to other types of big data, government big data possesses its own characteristics, primarily reflected in the following aspects:

(1) The issue facing government big data is its heterogeneous sources and lack of consistency. The various channels through which government big data is derived, along with its diverse range of data types, often lead to data discrepancies, contradictions, and even inconsistencies between algorithm structures and logic during the process of transmission, integration, analysis, and application.

(2) Government big data is confronted with the challenge of data silos. Many government departments still suffer from "information islands" and "data chimneys". Research has shown that government-controlled data accounts for 70% to 80% of the total national data. By harnessing this portion of data, developers can create innovative applications using government big data, provide better services, and generate more value.

(3) The issue of security and sharing risks arises with government big data. Compared to other types of data, the reliability and credibility requirements for government big data are higher, as it often involves national affairs and the well-being of citizens. The internet is currently plagued by various network viruses and malicious actors such as hackers, all posing threats to the security of government big data. With personal data, sensitive information, and confidential data being included in government big data, any unauthorized exploitation by malicious actors could result in serious social harm and destabilize society.

Therefore, when faced with the challenges of a diverse range of categories, rich data types, extensive time spans, and wide geographic coverage within the government system database, it is crucial to utilize the capabilities of a big data centre for efficient data integration and processing. This requires researchers to further explore the commonalities and heterogeneities among data across departments, industries, and domains.

At the same time, it is important for government big data to strive for transparent access. Transparent access to data allows different stakeholders within the system to have complete visibility into the introduction, exchange, and utilization of data, services, and products. For government big data, the inability to achieve transparent access could lead to trust issues among users regarding independent databases, resulting in negative customer feedback and dissatisfaction. Furthermore, a lack of transparency may also compromise data traceability and trust.

Transparent access in government big data refers to the abstract access to the underlying data of government big data from multiple data sources, leveraging technologies such as big data and cloud computing, combined with transparent access techniques to provide a unified access interface for government data in different formats, storage systems, and locations. The transparency of government big data can be divided into several dimensions:

(1) Data transparency: This transparency is related to the government data itself, where users hope to understand "what data", "who is involved", "when and where they are involved", and other relevant questions during the data access process.

(2) Process transparency: This refers to the transparency of information related to various government processes, such as policy formulation.

(3) Decision/policy transparency: This transparency includes explaining government decision-making processes and the fundamental principles behind policies.

2.1.2 Research Focus on Existing Data Access Technologies

Data access technology refers to the means by which applications access databases in some form to retrieve data. It connects the application layer with the data layer, encompassing data access interfaces and data access objects, allowing applications to access different database management systems and perform operations such as connecting, adding, deleting, and querying data in a single-type database [9]. From the perspectives of users and service providers, the research on existing data access technologies primarily focuses on the following aspects: the demand for accessing data anytime and anywhere [10], the need for efficient access to large amounts of data [11], the comprehensive requirement for data and storage management [12], the demand for easy data sharing [13], the requirement for dynamic data access [14], and ensuring data security [15, 16]. Data access technologies have evolved from initially only being able to access relational databases to now being capable of accessing various data sources, including relational databases, file directories, text content, spreadsheets, and emails, among others. Currently, commonly used data access technologies include ODBC (Open Database Connectivity), DAO (Data Access Object), ADO.NET (ActiveX Data Objects Net), XML (Extensible Markup Language), and JSON (JavaScript Object Notation) [17].

Transparent data access is a crucial means to ensure effective communication among governmental big data systems. It not only provides a unified access interface for data with different formats, storage systems, and locations, but also greatly enhances the capability of collecting, transmitting, and analyzing data from various heterogeneous sources. It plays a positive role in enhancing the flexibility, efficiency, and reliability of data communication among heterogeneous governmental systems in the fields of technology and knowledge [20].

2.1.3 Application of Data Transparent Access Technology

Currently, the application of data transparent access technology has been preliminarily explored and practiced in various fields such as rail transportation, mine safety management, and e-government. In the field of rail transportation, efforts have been made to access, store, govern, and share data related to railway infrastructure detection and monitoring. This has enabled the transparency in accessing and governing underlying monitoring data for railway infrastructure, as well as providing integrated analysis services. Transparent access technology has significantly reduced data retrieval costs and facilitated quick queries and application development for various railway data needs [21]. In the field of mine safety management, the concept of "perceptive mines" has been proposed in literature [22]. Coal mining conglomerates have gradually established centralized management platforms for multi-source mine monitoring data. Following the concept of "one map" by the Ministry of Land and Resources, the heterogeneous information distributed among mine monitoring locations is transparently accessed and integrated, enhancing transparency and safety in mine construction processes. In the field of e-government, in response to the governance challenges brought by massive government data and in line with the trend of big data and cloud computing, the development of e-government data governance based on transparent access technology is still in its early stages. China is actively promoting e-government data governance and transparent access to government data in the context of advanced information technologies such as big data, cloud computing, blockchain, and artificial intelligence [10, 23].

Building on an analysis of the application of transparent access technology in the existing government big data system, and delving into the underlying technological architecture that supports its use, this section will delve into a discussion of the access technology in government big data and cloud computing from the following three perspectives:

1. Multi-cloud transparent access technology in serverless computing, with a focus on achieving transparent computation in a multi-cloud environment and resolving trust issues stemming from varying infrastructures across different cloud environments.

2. Data access technology in organization distributed environments, with an emphasis on efficient read and write operations on data stored in different clouds, addressing the

challenges of accessing geographically dispersed data in the government big data environment.

3. Context-aware global data transparent access technology, with a focus on automatically configuring data access features and determining access and storage methods based on user access context. By maximizing the utilization of existing resources, this technology addresses various issues encountered by government big data users, such as accessing and storing data on the user end.

2.2 Multi-Cloud Transparent Access Technology for Serviceless Computing

Serverless computing is a cloud-based service in which the provider dynamically allocates sufficient resources to users. It represents the next stage of IaaS (Infrastructure-as-a-Service), offering a novel approach to application hosting that eliminates the need for end-user management. By decoupling the underlying infrastructure from the developer, it effectively virtualizes the runtime through the use of a process-level virtual machine. This paradigm, often referred to as FaaS (Function-as-a-Service) serverless architecture, boasts the pivotal feature of on-demand resource allocation, departing from the traditional pre-allocation method. With serverless computing, one can focus solely on executing a given task without concerns about servers, virtual machines, or the underlying computing resources. Notably, the open source community provides various frameworks rooted in Kubernetes [26, 27].

PyWren [28] employs remotely stored stateless functions to construct a data processing framework that inherits the resilience and simplicity of the serverless paradigm, while offering flexible building blocks for more intricate abstractions that effectively alleviate users from significant staffing and administration burdens. PyWren-IBM [29] presents a simpler framework compared to PyWren and exhibits greater enrichment. However, both systems mentioned above necessitate minimal modifications to user applications and rely on proprietary application programming interfaces. Fig. 1 presents a diagrammatic depiction of the execution flow of PyWren-IBM, wherein the client initially acquires user input, including data and executable programs; subsequently, both are serialized and stored in IBM COS (IBM Cloud Object Storage), and tasks are executed using IBM Cloud Functions to fulfil user requirements; finally, the results are returned to IBM COS and outputted to the client. Crucial [30] adopts a serviceless architecture for highly concurrent application programming, wherein the model maintains the simplicity of FaaS and allows for the migration of multithreaded algorithms to this programming environment. Fiber [31] furnishes a tailored multithreaded processing module on a Kubernetes cluster to oversee the scheduling of the cluster. However, neither Crucial nor Fiber provides a portable and scalable multi-cloud serviceless architecture. Lithops [29] can seamlessly execute unaltered programs on distributed cloud resources and offers the same application programming interface as the standard multi-threaded processing library [32], enabling any program built on it to run on any serviceless computing product. Lithops not only maintains a simple and consistent interface, but also grants transparent access to disaggregated storage and memory resources in the cloud. Presently, Lithops is capable of extending support to all next-generation compute and storage backends. Moreover, by employing the ability to scale multi-threaded modules, Lithops provides the technical means to port various application processes to the cloud at little to no cost to the user, thereby enabling, to some extent, transparent access in multi-cloud serviceless computing.

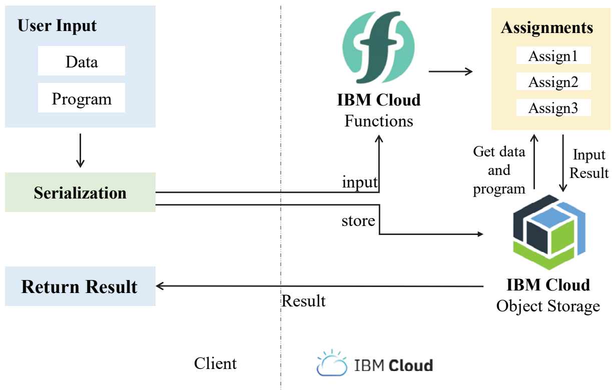

Figure 1 Schematic diagram of PyWren IBM execution process

Developers may inquire about the "most suitable serverless computing service" amidst the various options available and apprehensions regarding vendor lock-in. In such instances, adopting a multi-cloud strategy can furnish developers with the requisite agility to truly relish the advantages of serverless computing by harnessing the strengths of each cloud provider [33] and surmounting the challenges associated with provider lock-in. Although the amalgamation of services offered across diverse clouds represents a noteworthy advancement, serverless computing continues to encapsulate the intricacies of the cloud from developers and thus fails to attain complete accessibility transparency.

2.3 Organizing Data Access Technologies in Distributed Environments

The Distributed Computing Environment (DCE) is a widely adopted software technology that follows a Client/Server model. Its purpose is to construct and manage computation and data exchange in distributed computing systems, allowing users to access applications and data stored on remote servers.

IBM incorporates Active File Management (AFM) as an additional layer to its GPFS (General Parallel File System) storage. This layer enables data caching from remote sites and supports data modification. By establishing associations between data clusters, it facilitates a unified view of the data across clustered sites worldwide.

In network-based storage systems, inadequate random access performance is a common drawback of most document systems. Numerous researchers have tirelessly sought to overcome these limitations and introduce optimization mechanisms. For instance, Zhou et al. [36] proposed three approaches to optimize random queries and ensure efficient sequential access on HDFS (Hadoop Distributed File System). Gong et al. [37] introduced VarFs, a document system built on Ceph [38], which enhances random write performance through the utilization of variable-sized objects while maintaining POSIX compatibility.

Data transfer holds paramount importance in distributed environments when it comes to accessing data. In the majority of data access systems, the fundamental unit of data is a fixed-size block. When a user remotely accesses a file, the necessary data blocks must be transmitted through the network between data nodes. The stability and latency of the network link significantly affect efficiency, prompting the implementation of various optimizations. Examples of such optimizations include prefetching blocks, locally caching data, and fine-tuning network settings. Onedata [39] is able to provide access to distributed data under a single namespace and its main goal is to achieve truly transparent, efficient, scalable and cost-effective data access to data providers, but there is still an inherent lack of trust between data [40], and Fig. 2 shows the involved entity association and request processing flow in Onedata. In this case, the shared space stores the associated Users data, the resource provider provides storage resources to Onedata, and users are associated to user groups in order to simplify access control. When processing requests, File F4 is in the space provided by providers P1, P2, F4 data is stored in P2, P1 is connected to the FUSE client, the FUSE client exchanges user credentials with Global Registry (request 2) and sends a file request to P1 (request 3), P1 verifies that the FUSE client is authorized (request 4) and sends to P2 request (request 5), P2 verifies that P1 is allowed to access the data (request 6), and each provider in the system can determine the identity and authorization of the others and can proceed with the data transfer.

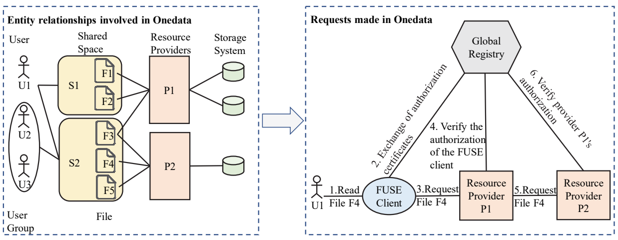

Figure 2 Entity relation and request processing flow involved in

Issues such as data provider autonomy, geographic distribution of large data sets, more complex and difficult to maintain network-based communications, data security and privacy, and decentralized authorization make current data access and sharing solutions are not good solutions for globalizing data access. Current data is often stored in large or sparse files that are read or written in arbitrary order by variable-sized blocks. While sequential access to data is often easy to handle, random reads and writes can be challenging for most network-based document systems. Especially when data is located on remote storage systems or in a distributed environment, read and write operations can trigger the transfer of entire documents or large blocks of data between storage clusters incurring unnecessary costs.

2.4 Context-Aware Global Data Transparency Access Technology

Context-awareness is the ability of a system to effectively use contextual information to make inferences about future events. In the pervasive computing environment, human and computer are constantly interacting transparently. During this interaction, the pervasive system confirms what kind of service to provide to the user by obtaining contextual information related to the user's needs, which is context-awareness, and it is an important technology in pervasive computing.

Wrzeszcz et al. [10] proposed a context-aware global data transparent access model that can automatically determine access and storage methods to adapt to user needs based on specific types of data access requests after analyzing existing solutions and the needs of data access stakeholders, such as ACID (atomicity, consistency, isolation, durability), BASE [41], reliability, and efficiency requirements. Fig. 3 illustrates the evolution of the transparent data access model. After the improvement of the original model, all clients can access the technology through context-aware transparent technology to achieve direct access to the server, which can avoid the problems of long-time storage and inaccessibility of the storage system. Specifically, Wrzeszcz et al. [42] assume a typical scenario used by independent providers distributed around the world, where data are stored in storage systems in nodes owned by the provider, each consisting of computers and storage resources, where the storage system is selected by the user according to his preferences or data application needs. In addition, there are some other issues to consider, such as the need for users to select the best solution for case-specific data access when the location of existing data, its characteristics, and the location of the storage system used to store the new data need to be determined, and the knowledge needed to make such decisions represents the context of the data access.

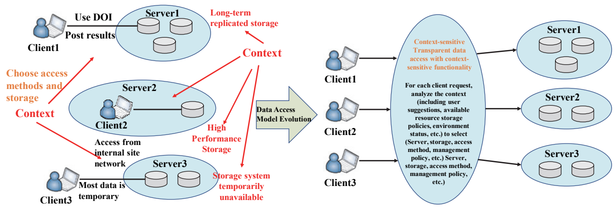

Figure 3 Evolution of the transparent data access model

Most existing solutions use metadata to provide access to data. Often, the number of metadata processed is limited in order to avoid overhead and bottlenecks, and therefore requires the support of user knowledge. For example, when users run data-intensive application processes, they need to manually migrate data to the storage where the computation is performed. If it is feasible to move all contextual information representing knowledge to an advanced data access system, when the system is logically a node accessing storage resources from multiple providers, then all data can be accessed in a unified way, using one type of client instead of using multiple tools to access different storage systems. To implement this property requires the definition of contextual information, which is represented by a defined set of extended metadata. Therefore, providing a context-aware global data access scheme based on context should enable the following attributes:

(1) Use metadata on various storage solutions to provide an abstraction layer and handle a wide range of contextual knowledge on various use cases through a unified data access interface.

(2) Different metadata consistency models can be coordinated to prevent bottlenecks when processing

metadata in parallel for efficient use of data from multiple providers.

(3) Synchronization between data organization and metadata can be achieved to support collaboration between independent providers.

Currently, sharing data often requires a lot of manual replication and migration operations, resulting in very little shared data [43]. And since different users may require different data access system attributes, such as a consistent view of distributed data, efficient data reading and writing, and avoiding or requiring data redundancy [44], it is often difficult to provide a solution that satisfies all stakeholders. When users plan to store or access data, they may still face the following problems that need to be solved: (1) Are the data access tools and storage systems localized? (2) Users have to consider not only their own needs but also the contextual environment (availability of storage space, load, type, number of storage users, etc.); (3) The above is not fully transparent to users due to the lack of appropriate domain knowledge.

In summary, none of the existing tools and solutions can simultaneously address the following issues: transparent data access, multi-source heterogeneous data access and fusion, and cross-domain semantic interoperability. Therefore, it is crucial to propose a framework that can address the above issues simultaneously, which is important for the development of transparent access to big data in government. The aforementioned research project, after fully considering the above transparent access needs of big data for government affairs, presents a visionary framework for transparent access to big data in government. The framework is based on cross-domain semantics and channel coupling technologies, in order to break the information barriers through semantic extraction and multi-channel information fusion to effectively ensure business coordination and connectivity among multi-source, multi-task and multi-platform government systems.

3 RESEARCH ON TRANSPARENT ACCESS TO GOVERNMENT BIG DATA BASED ON CROSS-DOMAIN SEMANTICS AND CHANNEL COUPLING

The research project titled "Internet + Transparent Management and Intelligent Service Platform for Government Big Data" is undertaken by a collaborative research team led by the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. The project focuses on exploring methods for transparent access to urban government big data and conducting research on technology such as cross-domain semantic interoperability of government big data, coupling of heterogeneous information channels, and online aggregation of heterogeneous government big data from multiple sources. The objective is to achieve controlled and transparent access to government data, as well as online aggregation and integration.

The research project proposes a visionary framework for transparent access to government big data based on cross-domain semantics and channel coupling in a hybrid architecture, as shown in Fig. 4. This framework enables the exchange and transformation of different business data through standardized and unified expression of business descriptions, effectively ensuring business coordination and connectivity among government systems with multiple sources, tasks, and platforms, optimizing the efficiency of transparent access. By constructing data cross-domain semantic interoperability interfaces, it achieves information integration and sharing between heterogeneous government systems, addressing the technical architecture heterogeneity and domain knowledge differences between different levels of government systems and other city systems from a technological and semantic interoperability perspective. Through the coupling technology of heterogeneous information channels, it establishes and maintains data directories to ensure accurate identification of data sources, extraction of data characteristics, ingestion and storage of heterogeneous data, and provides a unified external access interface, thus enhancing mutual verification and complementary information between channel information, and providing multi-source data for shared application platforms for government services. By employing methods such as online convergence and source tracing of heterogeneous data from multiple sources, it detects the source and knowledge creation process of government data, enabling traceability and audit of government data, thus ensuring the security of government data.

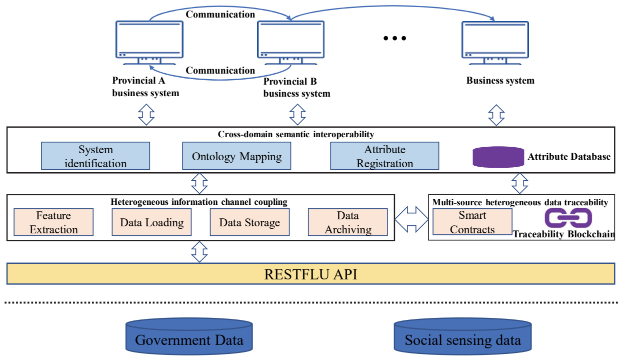

Figure 4 A transparent access framework for government big data based on cross-domain semantics and channel coupling under the mashup architecture

3.1 Cross-Domain Semantic Interoperability of Government Big Data

Interoperability of big data in government refers to the ability of multiple government systems to interconnect and operate together without restrictions. Interoperability can further facilitate the correct exchange of information and available data and ensure consistency of different business processes in key government departments [44]. Furthermore, interoperability is a prerequisite for multi-domain and sustainable digital government services, which facilitates further enhancement of public and private policy objectives, and is a key factor in increasing transparency and improving data quality in government.

The interoperability of big data in government faces the problem of heterogeneous domain knowledge between government systems and other city systems. Government departments at all levels deploy an information system that suits their own needs, resources and processes, and each government system has its own specific domain of expertise, so the vocabulary and terminology are also heterogeneous. As a result, the semantic meaning of one system is difficult to be recognized by another heterogeneous system, and the resulting semantic heterogeneity of data and the lack of a common language bring certain obstacles to data exchange and sharing. Therefore, to address the above issues, the study proposes semantic interoperability components that focus on semantic dimensions related to the ability to exchange information meanings between domains to ensure semantic understanding between cross-domain heterogeneous government systems.

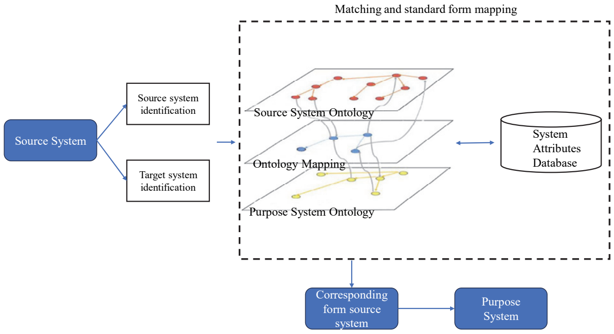

Figure 5 Cross-domain semantic interoperability model for government big data

The proposed model framework for cross-domain semantic interoperability of government big data is illustrated in Fig. 5. In this framework, a bottom-up approach is adopted for the semantic mapping between resource descriptions and attribute management in heterogeneous management domains. The core of this framework is the system attribute database.

The system attribute database is built upon a modular domain ontology database, which defines different types of participants and business service types. It is created based on a shared vocabulary among various levels of government departments, aiming to capture domain knowledge from a semantic perspective. The domain ontology database includes the following information:

(1) Definition of different types of business services, including service types, required resources, and corresponding processes.

(2) Definition of vocabulary for each participant, such as enterprises, healthcare institutions, environmental departments, transportation departments, education departments, etc., encompassing intervention processes, means, roles, etc.

Considering the diversity of shared vocabularies among different levels of departments, a modular approach is employed to construct the domain ontology database. Specifically:

(1) Various domain modules are used to represent the knowledge of information systems of government departments at different levels. These modules include enterprise module, healthcare module, environmental module, transportation module, education module, etc.

(2) A common core module is utilized to represent shared terms among different modules, ensuring more semantic interoperability among modules and facilitating their integration.

(3) A messaging module is defined to standardize communication behaviors among different systems.

The basic process of cross-domain semantic interoperability of big data for government affairs is as follows:

(1) When the component receives data from a source, it identifies the source system and the destination system, and confirms the source system ontology and the destination system ontology.

(2) The semantic layer looks up the system attribute database, matches the properties and meanings of the source system with those of the destination system through intermediate ontology mapping, and maps them to the standard form of the destination system.

(3) Finally, the component sends the form-transformed source system data to the destination system.

3.2 Coupling of Heterogeneous Information Channels in Government Big Data Based on Knowledge Graph

Information pipeline refers to the data transmission pathway that connects a specific department or system to a pool of data resources during a data retrieval process. On the other hand, the term "information channel" refers to an abstract collection of information pipelines that transmit the same type of data. In other words, an information channel can contain multiple information pipelines from different departments or systems. Channel coupling refers to the establishment of relationships between different channel data, which is a process of integrating multi-source and heterogeneous data.

The data types in typical business scenarios of government big data platforms include table data, text, image, video, sensing, location, etc. Transparent access to these data and their applications faces three main challenges: (1) data from multiple sources is heterogeneous, including various structured, semi-structured and unstructured data. (2) data contains multiple dimensions, such as real-time data, archived data, and multi-temporal data; (3) various formats data, for example, image data with .bmp, jpeg, .tif formats, etc., and video data with .avi, wmv, .mp4 formats, etc.

Knowledge graph embodies the relationships between entities in a data integration process, visualized as a graph. It exhibits characteristics of efficiency and strong scalability. However, when it comes to extracting entity relationships from multiple heterogeneous data sources, the process becomes complex due to the influence of various data types. Therefore, the proposed fusion framework in this paper primarily focuses on aggregating data of the same type into corresponding channels for processing and extracting key information. Through a standardized approach, the framework achieves data uniformity in terms of channel ID, data format, metadata, and physical storage location. The metadata consists of a set of key words. Subsequently, the knowledge graph methodology is employed for entity extraction and data integration.

The proposed model for the coupling technology of heterogeneous information channels in government big data is illustrated in Fig. 6. In order to ensure the correct identification of data sources, the system establishes a unified standard data catalogue to provide detailed descriptions of the data sources. New data sources are registered in the data catalogue upon joining, ensuring their availability, and the data catalogue is continuously monitored and maintained.

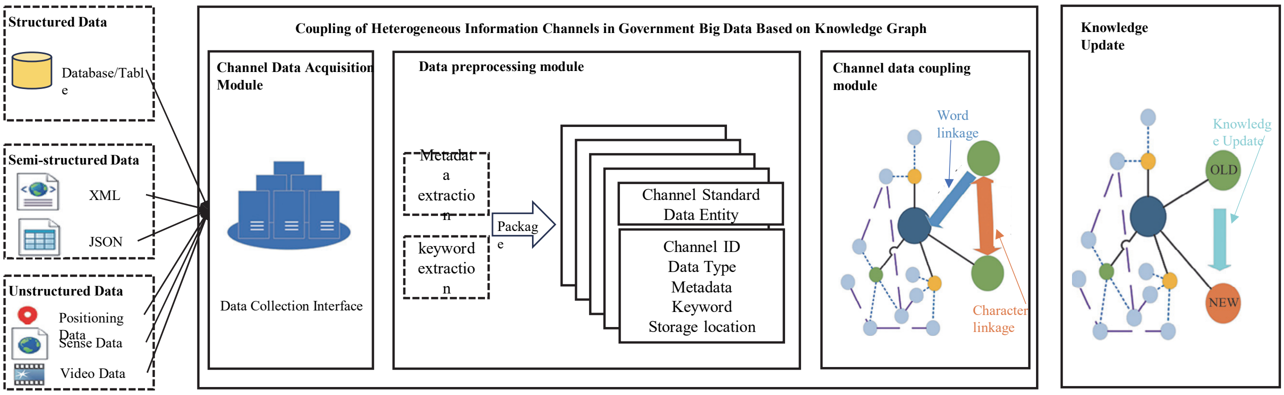

Figure 6 Heterogeneous information channel coupling technology for government big data

The basic process of the coupling technology in government big data heterogeneous information channels is as follows:

(1) Automatically or manually add keywords based on the government information resource catalogue, and establish an initial coupling knowledge graph for the information channels from top to bottom, thus achieving the ontology construction of the government knowledge graph.

(2) Divide the multiple heterogeneous data sources into different channels based on their data types. Through data preprocessing, normalize all the data within the channels into channel-standardized data, and then perform entity extraction.

(3) Based on a top-down sorting and retrieval model, integrate the extracted entity data into the extended coupling knowledge graph of the channels through data matching. By using the "word-to-word association" strategy and the "entry-to-entry correlation" strategy, knowledge fusion and knowledge processing are achieved, thus establishing relationships among entities in the coupling knowledge graph.

(4) Implement knowledge updates for the coupling knowledge graph through a top-down sorting and retrieval model.

3.3 Traceability of Multi-Source Heterogeneous Government Big Data

Data traceability refers to recording the traceability information of the origin data to trace the derivation process and the origin system of the data [47]. In the process of data sharing, exchange, and interoperation, the government data and other urban big data will generate a huge amount of derived data after many flows and other operations. The authenticity and validity of the derived data are difficult to be guaranteed, which will bring risks to the application of government affairs and related systems. Data traceability is an important means to improve the reliability and trustworthiness of government data.

Blockchain technology plays a crucial role in enhancing the security of large-scale data traceability [49]. It serves as a significant trend in the domain of data traceability, employing distributed ledger technology and consensus mechanisms to achieve decentralization, thus ensuring the integrity and tamper resistance of traceability data. Thus, the paper proposes a government data sharing solution that combines attribute-based encryption with blockchain, enabling data traceability as depicted in Fig. 7.

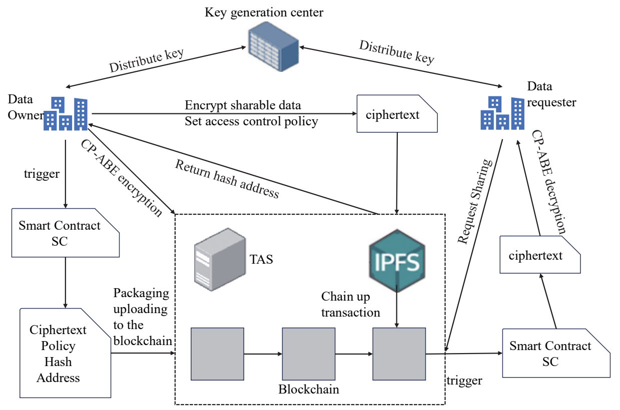

Figure 7 Model framework of multi source heterogeneous government big data traceability based on blockchain

The system architecture comprises seven components: Data Owner (DO), InterPlanetary File System (IPFS) distributed network, Data Recipient (DR), Key Generator Center (KGC), Smart Contract (SC), Blockchain (BC), and Trusted Authorization Server (TAS). The meanings and general process of each component are outlined below.

(1) DO can involve various governmental departments such as the judicial bureau, civil affairs bureau, and tax bureau as nodes in the consortium chain. When the owner wishes to share data with other departments, they can invoke a smart contract to generate an information description digest of the data. Two tasks are then performed: formulating corresponding access control policies, wherein only data requesters that meet the access policies are granted access to the data, and encrypting the data to obtain ciphertext, which is subsequently uploaded to the IPFS distributed network.

(2) The IPFS distributed network generates a unique hash identifier based on the content of the data and sends it to the DO. The DO then publishes the hash identifier to the blockchain.

(3) The DR, as a data consumer within the government department, is also a node in the consortium chain. Based on the information digest uploaded by the DO, the DR determines if the data meets its needs. The DR obtains the master secret key (MSK) from the KGC and uses its own attribute set to generate a user attribute secret key (SK). The DR then downloads the encrypted data from the IPFS distributed network and decrypts it.

(4) IPFS offers a peer-to-peer distributed storage structure that can accommodate a large amount of data files. Content addressing is utilized by IPFS to store hashes

in a distributed list, eliminating duplicate data files. DO encrypts externally shared data using attribute libraries, and then uploads the ciphertext data onto the IPFS network. The network returns a unique hash identifier based on the content of the ciphertext data, which DR uses to download the data file.

(5) KGC, as a node within the blockchain, addresses the issue of untrusted or semi-trusted traditional key generation nodes. KGC predefines all user attributes and assigns them to the respective nodes, forming a group of attribute nodes. KGC computes the private key (SK) for each node based on its attribute set. Additionally, KGC is responsible for updating and revoking user permissions, as well as revoking attribute sets. SC represents a protocol that imposes constraints between two or more blockchain nodes, where each party must fulfil their obligations according to the protocol. Hence, the smart contract can serve as a trusted third party.

4 CONCLUSION

This paper introduces the visionary framework of transparent access to government big data based on cross-domain semantics and channel coupling. The framework focuses on cross-domain semantic interoperability technology for government big data, heterogeneous information channel coupling technology for government big data, and multi-source heterogeneous government big data traceability technology. It aims to improve the current government system in terms of heterogeneous data interoperability, data fusion, and data audit tracking, laying a solid foundation for efficient sharing of government big data. In the future, we will continue to optimize the proposed framework for transparent access to government big data, providing technical support for the practical application of transparent access to government big data, and serving as a data foundation for cross-regional, cross-business, and cross-department comprehensive service innovation.

Currently, artificial intelligence is playing a significant role in various fields, driving industry transformation and upgrading. AI technology promotes the intelligence and automation of digital government services, utilizing reliable deep learning technology to advance the level of digital government services, minimizing processing time, reducing costs, enhancing transparency, and improving societal satisfaction. Looking ahead, there are several areas in which the transparent access technology of electronic government big data still needs further breakthroughs: (1) research on data aggregation and sharing standards and mechanisms for transparent access to government big data; (2) federated learning technology for transparent access to government big data towards edge computing; (3) security, privacy, and data protection technologies for transparent access to government big data; (4) transparent decision-making programmable planning technology and intelligent mining optimization technology for government big data transparent access.

In summary, the practical application and impact of the transparent access framework for government big data will play a crucial role in areas such as government decision-making, public services, data security, government innovation, and infrastructure development. It will provide strong support for achieving efficient sharing of government big data and fostering innovative cross-departmental comprehensive service applications.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Nos.2019YFB2102500).

5 REFERENCES

[1] Palmirani, M., Martoni, M., & Girardi, D. (2014). Open government data beyond transparency. In Electronic Government and the Information Systems Perspective: *Third International Conference, EGOVIS 2014*, *Munich, Germany*, 275-291. https://doi.org/10.1007/978-3-319-10178-1_22

[2] AlSukhayri, A. M., Aslam, M. A., Saeedi, K., & Malik, M. S. A. (2020). A linked open data-oriented sustainable system for transparency and open access to government data: A case study of the public's response to women's driving in Saudi Arabia. *Sustainability*, *12*(20), 8608. https://doi.org/10.3390/su12208608

[3] Fan, J., Sun, J., Li, H., Zhang, Y., Zhu, L., Qiu, A., & Duan, Y. (2023). Research Progress of Government Big Data Management Technology. *Journal of Integration Technology*, *3*, 1-18. https://doi.org/10.12146/j.issn.2095-3135.20221205001

[4] Sun, J., Fan, J., Xu, Y., Liu, Z., Zhang, Y., Qiu, A., & Wang, S. (2023). Design and application of the "Internet+" government big data intelligent service platform. *Journal of Integration Technology*, *1*, 4-16. https://doi.org/10.12146/j.issn.2095-3135.20220826001

[5] Lin, Y., Li, H., Zhang, C., Yan, J., & Wang, Z. (2023). Research on protocol conversion method based on knowledge graph. *Journal of Integration Technology*, *1*, 26-41. https://doi.org/10.12146/j.issn.2095-3135.20220730001

[6] Chen, C. (2022). Analysis on the Elements, Essence and Characteristics of E-government Service,Administration and Law. *Administration and Law*, *4*, 25-34.

[7] Meiyanti, R., Utomo, B., Sensuse, D. I., & Wahyuni, R. (2018). e-Government Challenges in Developing Countries: A Literature Review. *2018 6th International Conference on Cyber and IT Service Management (CITSM), Parapat, Indonesia*, 1-6. https://doi.org/10.1109/CITSM.2018.8674245

[8] Al-Sai, Z. A. & Abualigah, L. M. (2017). Big data and E-government: A review. *2017 8th international conference on information technology*, 580-587. https://doi.org/10.1109/ICITECH.2017.8080062

[9] Zu-kuan, W. E. I., Li-juan, J. I. A. N. G., & Jae-hong, K. I. M. (2009). Comparative Study of Database Access Technologies. *Computer and Modernization*, *1*(12), 46. https://doi.org/10.3969/j.issn.1006-2475.2009.12.013

[10] Wrzeszcz, M., Kitowski, J., & Słota, R. (2018). Towards trasparent data access with context awareness. *Computer Science*, *19*(2), 201-201. https://doi.org/10.7494/csci.2018.19.2.2844

[11] He, D. (2015). Distributed Database Access Technology Research Based on the. Net. *2015 2nd International Conference on Electrical, Computer Engineering and Electronics*, 1697-1701. https://doi.org/10.2991/icecee-15.2015.321

[12] Hünich, D. & Müller-Pfefferkorn, R. (2010). Managing large datasets with iRODS - A performance analysis. *Proceedings of the International Multiconference on Computer Science and Information Technology*, 647-654.

https://doi.org/10.1109/IMCSIT.2010.5679862

[13] Mazurek, M. L., Arsenault, J. P., Bresee, J., Gupta, N., Ion, I., Johns, C., ... & Reiter, M. K. (2010). Access control for home data sharing: Attitudes, needs and practices. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 645-654. https://doi.org/10.1145/1753326.1753421

[14] Kent, A. D. (2016). Cyber security data sources for dynamic network research. *Dynamic Networks and Cyber-Security*, 37-65. https://doi.org/10.1142/9781786340757_0002

[15] Jia, J. (2021). Verifiable Access Technology of Hybrid Database in Distributed System Under Big Data. *Journal of Physics: Conference Series*, *2010*(1), 012071. https://doi.org/10.1088/1742-6596/2010/1/012071

[16] He, K. (2021). Research on Verifiable Access Technology of Mixed Database in Distributed System under Big Data. *2021 International Conference on Aviation Safety and Information Technology*, 636-640. https://doi.org/10.1145/3510858.3511346

[17] Jin, W., Xu, R., Lim, S., Park, D. H., Park, C., & Kim, D. (2021). Integrated Service Composition Approach Based on Transparent Access to Heterogeneous IoT Networks Using Multiple Service Providers. *Mobile Information Systems*, *2021*, 1-19. https://doi.org/10.1155/2021/5590605

[18] Gupta, M. P. & Jana, D. (2003). E-government evaluation: A framework and case study. *Government information quarterly*, *20*(4), 365-387. https://doi.org/10.1016/j.giq.2003.08.002

[19] Bertot, J. C. & Choi, H. (2013). Big data and e-government: issues, policies, and recommendations. *Proceedings of the 14th annual international conference on digital government research*, 1-10. https://doi.org/10.3233/IP-140328

[20] Li, H., Zhao, A., Zhang, D., & Zhang, J. (2018). Research on building software usage model based on UML model. *International Journal of System Assurance Engineering and Management*, *9*, 675-683. https://doi.org/10.1007/s13198-017-0619-3

[21] Tao, K. et al. (2021). Design and Implementation of Railway Infrastructure Inspection and Monitoring Data Management and Analysis Service Platform Based on Big Data. *Railway Engineering*, *61*(9), 147-150. https://doi.org/10.3969/j.issn.1003-1995.2021.09.31

[22] Jing, H. C., Zhang, J., Meng, F. W., Yang, Y. L., & Ma, C. H. (2013). Research of data tree model in coal mine heterogeneous database integration. *Applied Mechanics and Materials*, *263*, 312-315. https://doi.org/10.4028/www.scientific.net/AMM.263-266.312

[23] Zhang, D. et al. (2014). Optimizing static analysis based on defect correlations. *Journal of Software*, *25*(2), 386-399. https://doi.org/10.13328/j.cnki.jos.004538

[24] Blakeley, J. A. & Pizzo, M. J. (1998). Microsoft universal data access platform. *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, 502-503. https://doi.org/10.1145/276304.276354

[25] Wang, H., Li, H., & Zhang, J.(2017).Cloud Task Scheduling Algorithm Based on Modified K-means Clustering. *Computer and Modernization*, *2*, 1. https://doi.org/10.3969/j.issn.1006-2475.2017.02.001

[26] Shamim, M. S. I., Bhuiyan, F. A., & Rahman, A. (2020). Xi commandments of kubernetes security: A systematization of knowledge related to kubernetes security practices. *2020 IEEE Secure Development (SecDev)*, 58-64. https://doi.org/10.1109/SecDev45635.2020.00025

[27] Sampe, J., Garcia-Lopez, P., Sanchez-Artigas, M., Vernik, G., Roca-Llaberia, P., & Arjona, A. (2020). Toward multicloud access transparency in serverless computing. *IEEE Software*, *38*(1), 68-74. https://doi.org/10.1109/MS.2020.3029994

[28] Jonas, E., Pu, Q., Venkataraman, S., Stoica, I., & Recht, B. (2017). Occupy the cloud: Distributed computing for the 99%. *Proceedings of the 2017 symposium on cloud computing*, 445-451. https://doi.org/10.1145/3127479.3128601

[29] Sampé, J., Vernik, G., Sánchez-Artigas, M., & García-López, P. (2018). Serverless data analytics in the IBM cloud. *Proceedings of the 19th International Middleware Conference Industry*, 1-8. https://doi.org/10.1145/3284028.3284029

[30] Barcelona-Pons, D., Sánchez-Artigas, M., París, G., Sutra, P., & García-López, P. (2019). On the faas track: Building stateful distributed applications with serverless architectures. *Proceedings of the 20th international middleware conference*, 41-54. https://doi.org/10.1145/3361525.3361535

[31] Zhi, J., Wang, R., Clune, J., & Stanley, K. O. (2020). Fiber: A platform for efficient development and distributed training for reinforcement learning and population-based methods. *arXiv preprint arXiv:2003.11164*. https://doi.org/10.48550/arXiv.2003.11164

[32] Aziz, Z. A., Abdulqader, D. N., Sallow, A. B., & Omer, H. K. (2021). Python parallel processing and multiprocessing: A rivew. *Academic Journal of Nawroz University*, *10*(3), 345-354. https://doi.org/10.25007/ajnu.v10n3a1145

[33] Li, H. et al. (2018). Research and Application of Cloud Computing Platform Status Monitoring Technology. *Software*, *39*(01), 9-13. https://doi.org/10.3969/j.issn.1003-6970.2018.01.003

[34] Bauerdick, L., Benjamin, D., Bloom, K., Bockelman, B., Bradley, D., Dasu, S., ... & Yang, W. (2012). Using xrootd to federate regional storage. *Journal of Physics: Conference Series*, *396*(4), 042009. https://doi.org/10.1088/1742-6596/396/4/042009

[35] Fajardo, E., Tadel, A., Tadel, M., Steer, B., Martin, T., & Würthwein, F. (2018). A federated Xrootd cache. *Journal of Physics: Conference Series*, *1085*(3)032025. https://doi.org/10.1088/1742-6596/1085/3/032025

[36] Zhou, W., Han, J., Zhang, Z., & Dai, J. (2012). Dynamic random access for hadoop distributed file system. *2012 32nd International Conference on Distributed Computing Systems Workshops*, 17-22. https://doi.org/10.1109/ICDCSW.2012.74

[37] Gong, Y., Hu, C., Xu, Y., & Wang, W. (2016). A distributed file system with variable sized objects for enhanced random writes. *The Computer Journal*, *59*(10), 1536-1550. https://doi.org/10.1093/comjnl/bxw057

[38] Weil, S. A., Brandt, S. A., Miller, E. L., Long, D. D., & Maltzahn, C. (2006). Ceph: A scalable, high-performance distributed file system. *Proceedings of the 7th symposium on Operating systems design and implementation*, 307-320.

[39] Wrzeszcz, M., Trzepla, K., Słota, R., Zemek, K., Lichoń, T., Opioła, Ł., ... & Kitowski, J. (2016). Metadata organization and management for globalization of data access with onedata. *Parallel Processing and Applied Mathematics: 11th International Conference*, 312-321.

[40] Opioła, Ł., Dutka, Ł., Słota, R. G., & Kitowski, J. (2018). Trust-driven, decentralized data access control for open network of autonomous data providers. *2018 16th Annual Conference on Privacy, Security and Trust (PST)*, 1-10. https://doi.org/10.1109/PST.2018.8514209

[41] Tudorica, B. G. & Bucur, C. (2011). A comparison between several NoSQL databases with comments and notes. *2011 RoEduNet international conference 10th edition: Networking in education and research*, 1-5. https://doi.org/10.1109/RoEduNet.2011.5993686

[42] Wrzeszcz, M., Dutka, Ł., Słota, R. G., & Kitowski, J. (2021). New approach to global data access in computational infrastructures. *Future Generation Computer Systems*, *125*, 575-589. https://doi.org/10.1016/j.future.2021.06.054

[43] Borgman, C. L. (2012). The conundrum of sharing research

data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059-1078. https://doi.org/10.1002/asi.22634

[44] Fan, J., Sun, J., & Zhang, Y. (2021). Design and Application of "Internet+" Government Big Data Intelligent Service Platform. *Journal of Integration Technology*, *12*(1), 4-16. https://doi.org/10.12146/j.issn.2095-3135.20220826001

[45] Al-Mushayt, O. S. (2019). Automating E-government services with artificial intelligence. *IEEE Access*, *7*, 146821-146829. https://doi.org/10.1109/ACCESS.2019.2946204

[46] Qi, M. & Wang, J. (2021). Using the internet of things E-government platform to optimize the administrative management mode. *Wireless Communications and Mobile Computing*, *2021*, 1-11. https://doi.org/10.1155/2021/2224957

[47] Wang, F., Zhao, H., & Ma, J.(2019).Research and Practice Progress of Data Provenance from the Perspective of Data Science. *Journal of Library Science in China*, *45*(5), 79-100.

[48] Han, C., Li, H., Yan, J., Lin, Y., Qu, J., & Jia, Z. (2023). Research on coupling technology of multi-source heterogeneous information channels based on knowledge graph. *Journal of Integration Technology*, (03), 48-60. https://doi.org/10.12146/j.issn.2095-3135.20221026001

[49] Zhang, H. & Sakurai, K. (2020). Blockchain for iot-based digital supply chain: A survey. *Advances in Internet, Data and Web Technologies: The 8th International Conference on Emerging Internet, Data and Web Technologies (EIDWT-2020)*, 564-573. https://doi.org/10.1007/978-3-030-39746-3_57

Contact information:

Honghui LI, MS, Professor
School of Computer and Information Technology College of Beijing Jiaotong University,
Beijing Jiaotong University,
No. 3 Shang Yuan Village, Xizhimen Wai, Haidian District, Beijing
E-mail: hhli@bjtu.edu.cn

Yi XU, MS
(Corresponding author)
School of Computer and Information Technology College of Beijing Jiaotong University,
Beijing Jiaotong University,
No. 3 Shang Yuan Village, Xizhimen Wai, Haidian District, Beijing
E-mail: 23111101@bjtu.edu.cn

Yunjuan PENG, MS
School of Software Engineering College of Beijing Jiaotong University,
Beijing Jiaotong University,
No. 3 Shang Yuan Village, Xizhimen Wai, Haidian District, Beijing
E-mail: 21126333@bjtu.edu.cn

Xiaorui YANG, MS
School of Computer and Information Technology College of Beijing Jiaotong University,
Beijing Jiaotong University,
No. 3 Shang Yuan Village, Xizhimen Wai, Haidian District, Beijing
E-mail: 22110124@bjtu.edu.cn

Dalin ZHANG, PhD, Associate Professor
School of Software Engineering College of Beijing Jiaotong University,
Beijing Jiaotong University,
No. 3 Shang Yuan Village, Xizhimen Wai, Haidian District, Beijing
E-mail: dalin@bjtu.edu.cn

Jianping FAN, PhD, Professor
Shenzhen Institute of Advanced Technology Chinese Academy of Sciences,
No. 1068 Xueyuan Avenue, Shenzhen University City,
Xili, Nanshan District, Shenzhen
E-mail: jp.fan@siat.ac.cn

Jing SUN, Senior Engineer
Shenzhen Institute of Advanced Technology Chinese Academy of Sciences,
No. 1068 Xueyuan Avenue, Shenzhen University City,
Xili, Nanshan District, Shenzhen
E-mail: jing.sun1@siat.ac.cn