Proceedings of SEML 2025 Workshop: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/146/2025.TJ21928

Research on Key Technologies, System Construction and Application Outlook of Big Data Governance

Zifeng Zhao

School of Software, Shandong University, Jinan, China
202200300159@mail.sdu.edu.cn

Abstract: In the era of digital transformation, big data has emerged as a pivotal resource driving innovation across various sectors. The effective utilization of big data, however,is fundamentally contingent upon robust governance frameworks. This paper conducts a comprehensive exploration of the multifaceted domain of big data governance, with a focus on three critical dimensions: key technological implementations, policy-system architecture, and application prospects. By employing a combination of literature review, case analysis, and comparative research methods, it addresses several key research questions, such as how to improve the efficiency and quality of data management in big data governance, how to construct a scientific and reasonable policy system, and how to better apply big data governance in different scenarios. The research findings demonstrate that advanced techniques, particularly data wrangling, play an indispensable role in data preprocessing and quality enhancement. Furthermore, the study reveals that a sophisticated, multi-tiered policy system involving multiple stakeholders is progressively evolving in the realm of policy-system construction. In terms of application, big data governance has achieved remarkable results in areas such as government decision-making and business operations, but challenges remain. Overall, this research provides valuable insights for promoting the development and application of big data governance.

Keywords: Big data governance, Data wrangling, Policy system, Application scenarios

1. Introduction

In the digital age, the exponential growth in data volume, variety, and velocity has ushered in the era of data, which has grown exponentially, transforming it into a strategic asset for governments, enterprises, and institutions. This transformation holds immense potential to drive innovation, enhance decision-making processes, and bolster competitiveness. For example, in the government sector, big data can be used to optimize urban planning, improve public service efficiency, and enhance social governance capabilities [1,2]. In the business field, it can help enterprises better understand market trends and customer needs and improve product development and marketing strategies. Despite its transformative potential, the effective utilization of big data faces significant challenges. The complexity of data sources, the heterogeneity of data formats, and the low quality of data often make it difficult to extract valuable information from big data. Moreover, issues such as data privacy, security, and governance mechanisms also need to be addressed urgently. Current research on big data governance predominantly revolves around three core areas: technological advancements, policy formulation, and application scenarios. From a technological perspective, key

DOI: 10.54254/2755-2721/146/2025.TJ21928

focus areas include data wrangling, data integration, and data quality management [3,4]. In the policy-making field, the construction of a policy system for big data governance, especially in the context of emerging technologies like artificial intelligence, is attracting increasing attention. In application scenarios, big data governance has been widely applied in e - commerce, finance, healthcare, and other industries, but there is still room for improvement in terms of effectiveness and efficiency.

This paper adopts a comprehensive research method, combining a literature review to understand the current research status, case analysis to explore practical applications, and comparative research to analyze different governance models. It seeks to address critical research questions, including strategies to improve data management efficiency and quality, frameworks for establishing scientifically robust policy systems, and methodologies to optimize big data governance across diverse scenarios. The significance of this research lies in providing theoretical support and practical guidance for promoting the development of big data governance. By exploring key techniques, policy - system construction, and application prospects, it can help relevant stakeholders better understand big data governance, make more informed decisions, and promote the wide - spread and in - depth application of big data in various fields.

2. The Key Techniques of Big Data Governance

2.1. Data Structuring Processing

2.1.1. Information Extraction from Unstructured Data

As data sources continue to diversify, a significant portion of data exists in unstructured data exists in forms such as text, images, and audio. For instance, in the realm of news media big data, a vast number of news articles are in text format. o process and analyze such data, named-entity recognition (NER) techniques are employed, which can be categorized into three primary approaches: regular expression-based methods, dictionary-based methods, and machine learning models. These techniques enable the identification of key entities such as individuals, organizations, and events within textual data [4]. For example, in a news article about a corporate merger, the names of the two merging companies, the date of the merger, and the key figures involved can be extracted through named - entity recognition.

2.1.2. Conversion of Semi - structured Data

Semi - structured data, such as JSON and XML, has a certain degree of structure but is more flexible than traditional relational data. When dealing with JSON - formatted data from an e - commerce platform that records product information, including product names, prices, descriptions, and customer reviews, specific rules need to be determined to convert it into a more structured form, such as a two - dimensional table, to facilitate further data analysis.

2.2. Data Quality Assessment and Cleaning

2.2.1. Quality Problem Detection through Visualization

Data quality issues are prevalent in big data, including problems such as missing values, inconsistent data, and outliers. Visualization techniques play a crucial role in detecting these problems. For example, in a sales data set of a retail enterprise, using bar charts or scatter plots to visualize sales volume by region and time can easily reveal abnormal sales values in certain regions or time periods, which may be due to data entry errors or other reasons [5].

DOI: 10.54254/2755-2721/146/2025.TJ21928

2.2.2. Data Cleaning Rules and Iterative Processing

Upon identifying data quality issues, it is essential to establish appropriate cleaning rules to address them. For example, if there are missing values in a customer age field in a customer relationship management system, rules can be set to fill in the missing values based on the average age of customers in the same demographic group or through more complex machine - learning - based interpolation methods. This process often requires multiple rounds of human - machine interaction to continuously improve the quality of the data [6-12]. Such an approach ensures that the data remains accurate, consistent, and reliable for subsequent analysis and decision-making.

2.3. Data Normalization

2.3.1. Handling of Low - level Data Normalization

Simple data normalization tasks include data type conversion, unit transformation, and format conversion. For example, in a scientific research data set, temperature data may be recorded in different units such as Celsius and Fahrenheit. To ensure data consistency, all temperature data needs to be converted to a unified unit.

2.3.2. Entity Linking and Granularity Selection in Complex Data Normalization

In more complex data normalization tasks, such as address normalization, the challenge of entity linking arises. This involves mapping different representations of the same entity to a unified form. At the same time, the selection of granularity is also crucial. For example, an address can be expressed at different granularities, such as country - province - city - district - street. The appropriate granularity needs to be determined according to the specific application requirements.

2.4. Data Fusion and Extraction

2.4.1. Challenges and Solutions in Data Fusion

Data fusion aims to integrate multiple data sets from different sources to obtain more comprehensive information. However, this process is fraught with challenges stemming from the inherent heterogeneity and autonomy of data sources. For example, in a cross - departmental data initiative within a government agency, different departments may use different data formats and naming conventions for the same type of data. Addressing this issue necessitates the implementation of entity-linking operations. By using entity - linking technology, the same entities in different data sources can be identified and linked, such as identifying the same citizen's information in the public security department's population data and the social security department's data [13].

2.4.2. Data Extraction for Specific Analysis Tasks

In some cases, not all integrated data is needed for analysis tasks. For example, in a market research project on a specific product, only data related to the target product's sales, customer feedback, and competitor information needs to be extracted from a large - scale e - commerce data set. This data extraction process needs to be based on the characteristics of the analysis task to ensure that the extracted data can effectively support the analysis.

2.5. Summary

These key techniques in big data governance are closely interconnected, working in concert to achieve the overarching objectives of big data governance. Data structuring processing forms the foundation,

DOI: 10.54254/2755-2721/146/2025.TJ21928

transforming various raw data into a format that is convenient for subsequent processing, and providing an orderly data basis for data quality assessment and cleaning. Data quality assessment and cleaning, in turn, are carried out on the basis of structured data. It detects and corrects data issues to enhance data usability, providing high-quality data for data normalization as well as data fusion and extraction. Data normalization further refines the cleaned data by standardizing and unifying it, thereby ensuring consistency and universality across different scenarios. This standardization is crucial for enabling seamless data integration and analysis. Building on the achievements of the previous techniques, integrate multi-source data and extract key information in a targeted manner, realizing the maximized utilization of data value. This series of techniques progress in a step-by-step manner and support one another, jointly ensuring the efficient operation of big data governance. This framework not only enhances the reliability and utility of data but also facilitates its effective application across diverse fields, driving innovation and informed decision-making,

3. The Construction of the Big Data Governance Policy System

3.1. Policy - making Process and Current Situation

The policy-making process for big data governance constitutes a complex system engineering endeavor, driven by national development strategies, technological innovation, and social needs. Globally, nations worldwide have formulated policies to promote the development of artificial intelligence-related big data governance. In 2017, China issued the "New Generation Artificial Intelligence Development Plan," which mandates the enhancement of data infrastructure and the integration and sharing of data resources. Policy types include normative documents, regulations, and standards. Normative documents, such as development plans and action plans, account for a large proportion of big data governance policies. Regional distribution is more abundant in developed regions, with states like California and New York having more detailed and comprehensive policies covering aspects such as data privacy protection, open-sharing, and security.

3.2. Policy Content and Key Points

The policy underscores the development of multi-level and multi-subject collaborative governance mechanisms for AI development, encompassing international cooperation in critical areas such as data standard formulation and cross-border data flow management. A prominent example is the European Union's General Data Protection Regulation (GDPR) regulates data protection within the EU and impacts international data governance cooperation. At the national level, the government plays a leading role in formulating policies, regulations, and standards, while local governments implement national policies and formulate corresponding implementation rules. The governance process also involves active participation from various stakeholders, including enterprises, social organizations, and the public. The policy explicitly outlines data requirements for high-quality AI development are clearly defined. A wide range of data sources are needed, including data from various industries like healthcare, finance, and transportation. For example, medical AI requires a large amount of patient medical records, imaging data, and genetic data. Data quality is prioritized, emphasizing authenticity, integrity, timeliness, accuracy, and diversity. For example, in a financial risk prediction AI model, accurate and timely input data is essential for accurate risk prediction results.

DOI: 10.54254/2755-2721/146/2025.TJ21928

4. The Application of Big Data Governance in Different Scenarios

4.1. In the Government Sector

Enhancing Decision-making and Optimizing Public Services through Big Data Analysis: Governments can leverage big data governance to significantly improve decision-making processes and optimize public services. In urban planning, for example, the analysis of comprehensive datasets on population distribution, traffic patterns, and land utilization enables the formulation of more scientifically grounded and informed decisions. This is particularly useful in the planning of new city districts, where insights from population growth trends, employment distribution, and transportation demand data can lead to more strategic placement of residential areas, commercial zones, and transportation facilities. Furthermore, big data governance plays a crucial role in enhancing public services. In the realm of public transportation, the examination of passenger travel data—such as travel times, origin-destination patterns, and transfer information—allows for the optimization of service operations. This includes adjusting bus routes to better match peak-hour travel demands and dynamically altering subway train frequencies to boost both the efficiency and convenience of public transportation systems. Together, these applications of big data not only streamline urban development but also significantly enhance the quality of life for citizens by making public services more responsive to their needs.

4.2. In the Business Sector

In the contemporary business environment, the integration of big data governance has become a cornerstone for driving customer-centric strategies and optimizing operational efficiency. Companies can leverage big data to transform their marketing and product development processes, ensuring they are closely aligned with customer preferences and market dynamics. For instance, e-commerce platforms can analyze vast amounts of customer data, including browsing patterns, purchase histories, and product reviews, to gain deep insights into customer preferences and needs. This data-driven approach enables businesses to offer personalized product recommendations, enhancing the customer experience. A fashion e-commerce platform, for instance, can recommend styles that resonate with a customer's past purchases and browsing habits, while also utilizing trend analysis to design new products that cater to the shifting preferences of diverse customer segments. This not only boosts customer satisfaction but also drives innovation and competitiveness in the market. Beyond customer-focused initiatives, big data governance plays a pivotal role in optimizing supply chain management, a critical component of business operations. By analyzing data from suppliers, production processes, logistics, and inventory, enterprises can streamline their supply chains to achieve greater efficiency and cost-effectiveness. For example, predictive analytics can be used to forecast raw material supply levels based on historical data and market trends, allowing businesses to adjust production plans proactively. Additionally, logistics routes can be optimized to reduce transportation costs and delivery times, further enhancing supply chain performance. In a globalized economy where supply chain disruptions can have significant repercussions, the ability to leverage big data for real-time decision-making and optimization is invaluable. Collectively, these applications of big data governance—ranging from customer-oriented marketing to supply chain optimization—empower businesses to stay agile, responsive, and competitive in an increasingly data-driven world. By integrating these strategies, enterprises can not only meet customer expectations more effectively but also achieve operational excellence and sustainable growth. This holistic approach ensures that businesses are well-equipped to navigate the complexities of the modern marketplace, driving both customer satisfaction and long-term success.

DOI: 10.54254/2755-2721/146/2025.TJ21928

5. Conclusion

This research comprehensively explores big data governance from the aspects of key techniques, policy - system construction, and application scenarios. In terms of key techniques, data wrangling techniques such as data structuring processing, quality assessment and cleaning, normalization, and fusion and extraction are crucial for improving data quality and usability in big data governance. These techniques help to transform raw and complex data into valuable information resources that can support decision – making. Regarding policy-system construction, a multi-level, multi-subject collaborative governance framework is gradually emerging. This system promotes the coordinated development of big data governance at different levels and among different stakeholders, and provides a solid institutional foundation for the healthy development of big data governance. It also clearly defines data requirements for high - quality AI development, which is conducive to promoting the innovation and application of artificial intelligence technology. In application scenarios, big data governance has achieved certain results in the government and business sectors. In the government, it has enhanced decision - making and optimized public services; in the business field, it has promoted customer - oriented marketing, product development, and supply chain management optimization.

However, this study is not without limitations. For instance, in the examination of key techniques, the analysis of emerging technologies such as blockchain-based data governance remains insufficient. Similarly, in the exploration of policy-system construction, the adaptability of policies across diverse cultural and economic contexts has not been fully addressed. In future research, more attention can be paid to the integration of emerging technologies in big data governance, such as the combination of artificial intelligence and data governance techniques to improve governance efficiency. At the same time, cross - cultural and cross - regional research on big data governance policies can be carried out to provide more comprehensive policy recommendations. In application scenarios, further in-depth case studies are needed to identify more effective models and solutions, ensuring the continued evolution and optimization of big data governance practices.

References

[1] *Jian Y, Huaijie Z, Jianxing Y, et al. A panoramic framework for big data governance [J]. Big Data.,2020, 6 (1): 19 - 26.*

[2] *Wenhong Z, Yahan Y, and Xiaofang X. Progress and Prospects of Building a Data Governance Policy System for Artificial Intelligence in China [J]. Library Forum,2025.*

[3] *HELLERSTEIN J M, HEER J, KANDEL S. Self - service data preparation: research to practice [J]. IEEE Data Engineering Bulletin, 2018, 41 (2): 23 - 34.*

[4] *HEER J, HELLERSTEIN J M, KANDEL S. Data wrangling [M]//Encyclopedia of big data technologies 2019. [S.l.:s.n.], 2019.*

[5] *LI G L, ZHENG Y D, FAN J, et al. Crowdsourced data management: over view and challenges [C]//The 2017 ACM International Conference on Management of Data, May 14 - 19, 2017, Chicago, USA. New York: ACM Press,2017: 1711 - 1716.*

[6] *DOAN A H, ARDALAN A, BALLARD J R, et al. Toward a system building agenda for data integration [J]. IEEE Data Engineering Bulletin, 2018, 41 (2): 35 - 46.*

[7] *SONG X Y, WANG Y H. Data integration and application integration [M]. Beijing: China Water and Power Press, 2008.*

[8] *ABEDJAN Z, CHU X, DENG D, et al. Detecting data errors: where are we and what needs to be done [J]. Proceedings of the VLDB Endowment, 2016, 9 (12): 993 - 1004.*

[9] *BOHANNON P, FAN W F, GEERTS F, et al. Conditional functional dependencies for data cleaning [C]//2007 IEEE 23rd International Conference on Data Engineering, April 15 - 20, 2007, Istanbul, Turkey. Piscataway: IEEE Press, 2007: 746 - 755.*

[10] *CHU X, ILYAS I F, PAPOTTI P. Holistic data cleaning: putting violations into context [C]//2013 IEEE 29th International Conference on Data Engineering (ICDE), April 8 - 12, 2013, Brisbane, Australia. Piscataway: IEEE Press, 2013: 458 - 469.*

DOI: 10.54254/2755-2721/146/2025.TJ21928

[11] CHU X, MORCOS J, ILYAS I F, et al. KATARA: a data cleaning system powered by knowledge bases and crowdsourcing [C]//The 2015 ACM SIGMOD International Conference on Management of Data, May 31 - June 4, 2015, Melbourne, Australia. New York: ACM Press, 2015: 1247 - 1261.

[12] YAKOUT M, BERTI - ÉQUILLE L, ELMAGARMID A K. Don't be scared: use scalable automatic repairing with maximal likelihood and bounded changes [C]//The 2013 ACM SIGMOD International Conference on Management of Data, June 22 - 27, 2013, New York, USA. New York: ACM Press, 2013: 553 - 564.

[13] REKATSINAS T, CHU X, ILYAS I F, et al. HoloClean: holistic data repairs with probabilistic inference [J]. Proceedings of the VLDB Endowment, 2017, 10 (11): 1190 - 1201.