Journal of Electronic Research and Application, 2024, Volume 8, Issue 3
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Designing and Implementing an Advanced Big Data Governance Platform

Yekun Chen[1], Tianqi Xu[1], Yongjiang Xue[2]*

[1]CETC Academy of Electronics and Information Technology Group Co., Ltd., Beijing 100040, China
[2]School of Computer Science and Technology, Tiangong University, Tianjin 300387, China

Corresponding author: Yongjiang Xue, xcydj83@126.com

Abstract: Contemporary mainstream big data governance platforms are built atop the big data ecosystem components, offering a one-stop development and analysis governance platform for the collection, transmission, storage, cleansing, transformation, querying and analysis, data development, publishing, and subscription, sharing and exchange, management, and services of massive data. These platforms serve various role members who have internal and external data needs. However, in the era of big data, the rapid update and iteration of big data technologies, the diversification of data businesses, and the exponential growth of data present more challenges and uncertainties to the construction of big data governance platforms. This paper discusses how to effectively build a data governance platform under the big data system from the perspectives of functional architecture, logical architecture, data architecture, and functional design.

Keywords: Big data; Data governance; Cleansing and transformation; Data development; Sharing and exchange

Online publication: June 14, 2024

1. Introduction

The burgeoning fields of big data, machine learning, and deep learning have led enterprises to amass vast quantities of data, recognizing its value as a strategic asset crucial for business intelligence [1-3]. As data‘s role as a foundational strategic resource grows, effective data governance becomes imperative to harness its potential fully. Data governance integrates methodologies to manage and utilize both internal and external organizational data assets effectively [4]. However, deficiencies in data governance architecture or capabilities can impede leveraging big data, leading to issues like data obscurity, demand fulfillment challenges, and difficulties in data sharing. Thus, establishing a robust data governance framework is vital for mitigating these challenges and enhancing data utilization.

Data governance involves defining clear organizational roles and responsibilities, forming dedicated teams to develop and implement data governance strategies, and ensuring compliance across departments. It encompasses setting standards, managing data practices, and establishing frameworks for data ownership, permissions, usage, security, quality, and lifecycle management. This framework supports precise data collection, storage, management, sharing, and utilization, facilitating informed decision-making and meeting business needs. Effective data governance also requires regular assessments of its efficiency and outcomes, with mechanisms like data auditing and reporting ensuring compliance and data integrity. By prioritizing data security training and adopting rigorous data management protocols, organizations can mitigate risks of data leaks and misuse, thereby bolstering data reliability and security, and supporting strategic business initiatives [5,6].

2. Functional architecture

Modern data governance platforms are increasingly integrating with data middleware technologies, aligning functions and domains for enhanced data management. Data governance sets the quality control standards, infusing rigor, and discipline into data management, utilization, optimization, and protection processes. Data middleware provides the infrastructure that transforms data into business assets, bridging the gap between front-end and back-end systems, and facilitating coordination between application and data development. Hence, data governance defines the standards, while data middleware serves as the practical tool. The technical architecture of data governance platforms leverages data middleware technologies for design and implementation, focusing on tasks such as formulating governance strategies, integrating data assets, creating resource directories, and mapping data lineages. These efforts ensure continuous monitoring, quality measurement, risk management, and the establishment of system standards for data management, resource cataloging, metadata management, and data security, aiming to optimize data systems and guarantee data availability, security, and quality. In practice, a data governance platform should support or enhance the functionalities of big data platforms, data asset management systems, and data service platforms. The overall architecture of a data governance platform is illustrated in **Figure 1**.

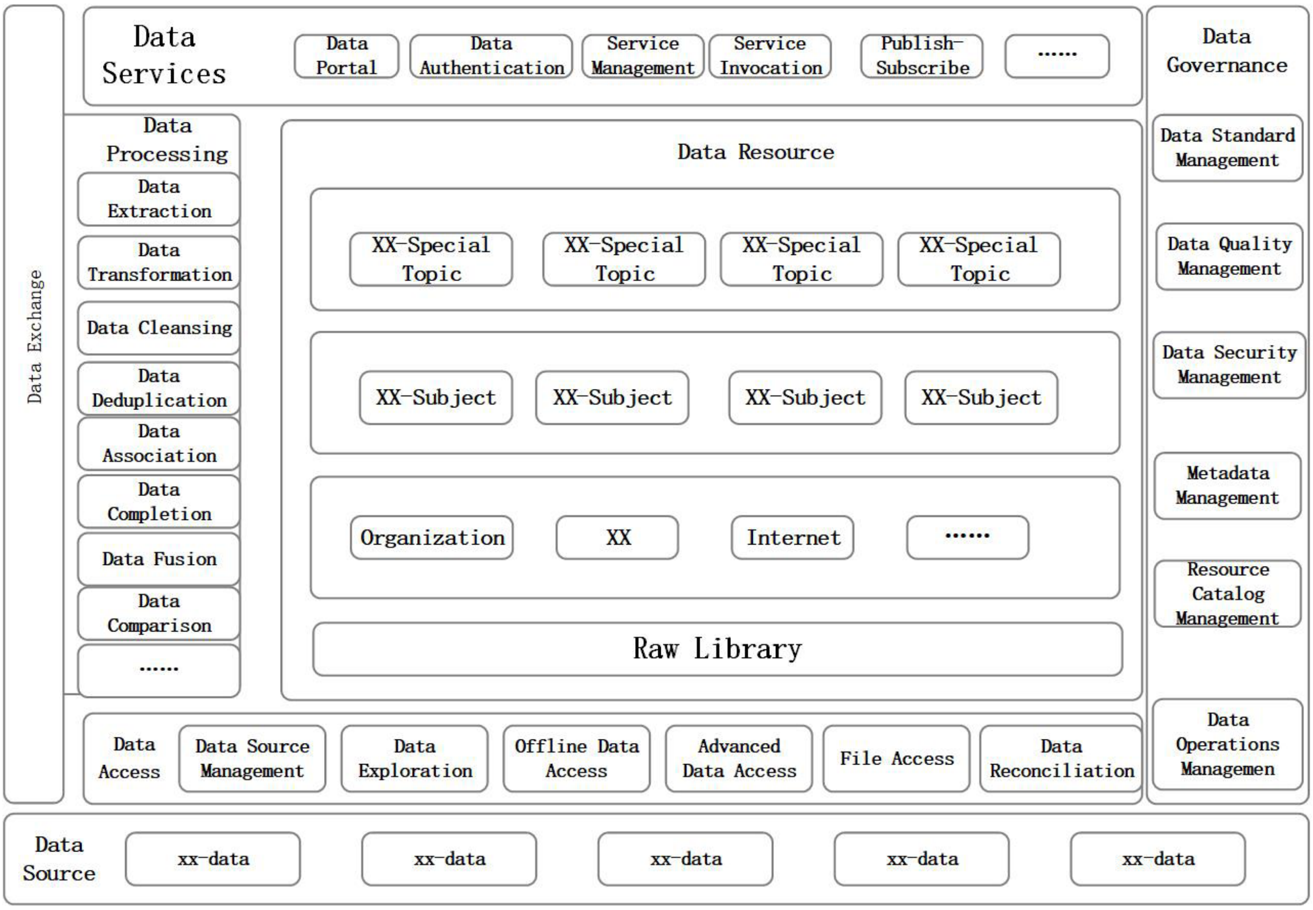

Figure 1. Functional architecture of a data governance platform

3. Logical architecture

Data governance necessitates the creation of four pivotal systems: Data Standards System, Big Data Collection and Access System, Data Resource Governance System, and Data Security System.

Data Standards System: Established prior to implementing data governance, it sets crucial data standards for consistency and accuracy across an organization's diverse data sources. This includes defining data formats, quality criteria, and metadata standards to support business, technology, and management [7].

Big Data Collection and Access System: Acting as the gateway for data integration, this system utilizes various methods to collect and input data into the big data platform, ensuring accurate data capture and availability for analysis and mining.

Data Security System: This system implements a range of security measures to protect data confidentiality [8], integrity, and availability, preventing unauthorized access, modification, or destruction through encryption, access controls, and data masking.

Data Resource Governance System: At the heart of data governance, this system establishes a rigorous framework for data resource management, focusing on security, integrity, and efficiency. It manages the data lifecycle and enhances data quality, metadata, lineage, and cataloging to improve data asset usability.

Collectively, these systems enable robust data access, organization, processing, governance, services, and exchange, aligning with organizational goals and ensuring effective data utilization and protection under stringent governance standards.

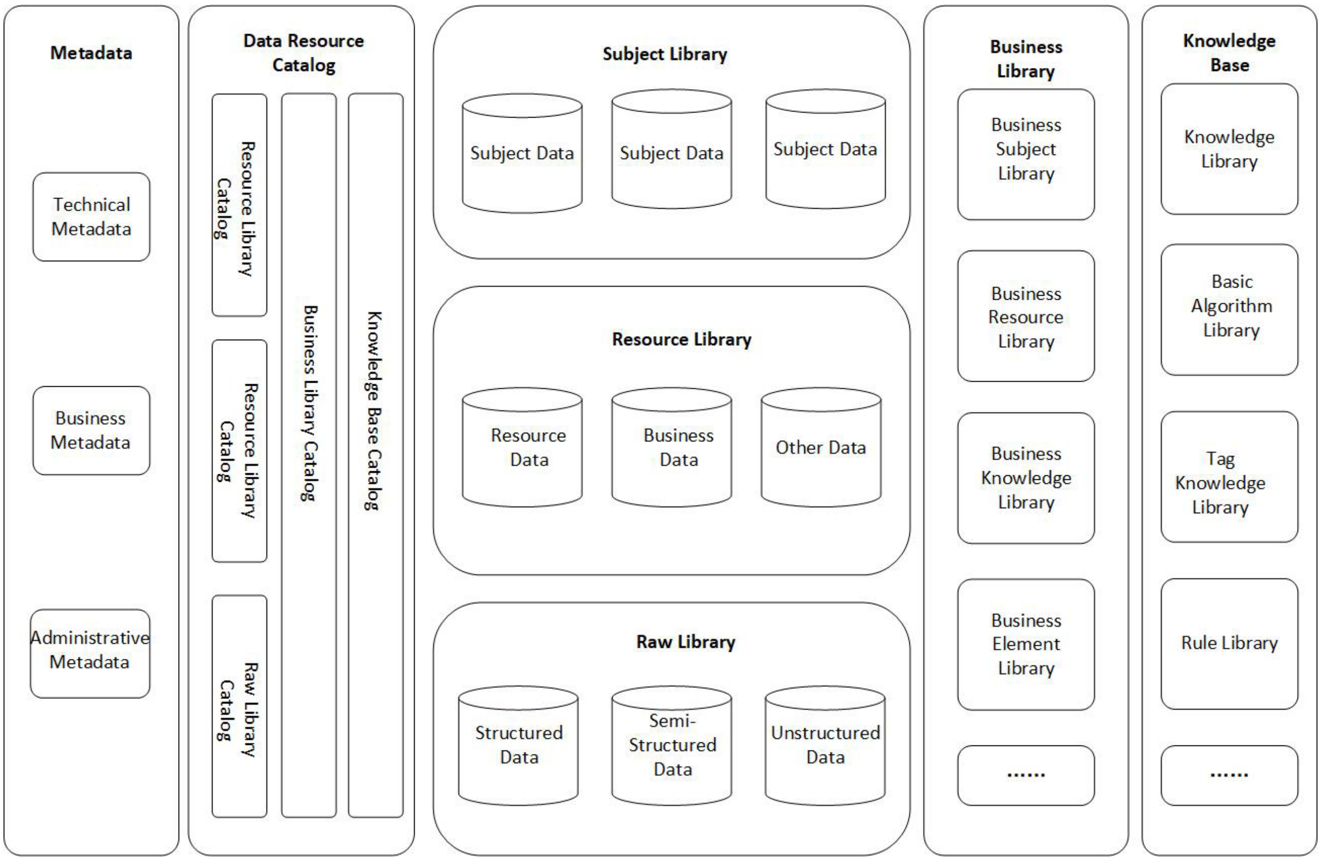

Figure 2. Data architecture

4. Data architecture

Over tIn managing multi-source heterogeneous data, a unified data architecture is crucial, integrating data access modes, storage formats, processing technologies, and service applications. This architecture ensures efficient data access, governance, and service deployment.

Raw Data Repository: Captures data directly from business systems, serving as the base for further data management processes and maintaining data in its original, unprocessed state.

Resource Repository: Builds upon the Raw Data Repository, normalizing and governing data around specific themes, thus organizing and standardizing data for enhanced accessibility and utility.

Subject Repository: Utilizes data from raw and resource repositories to focus on subject-specific collections, facilitating detailed analysis and insights across various dimensions.

Business Repository: Supports different business scenarios with structured data tailored to operational and strategic needs.

Knowledge Repository: Stores domain-specific knowledge, data, and methodologies, supporting advanced data-driven decision-making. Additional components can be established based on data construction needs, including:

Unified Index Repository: Consolidates index information for all data resources, enhancing data retrieval and access efficiency.

Data Resource Directory Repository: Acts as a comprehensive catalog, documenting data attributes, locations, volumes, and permissions.

Metadata Database: Contains essential metadata providing context for data resource management and utilization.

Management Information Repository: Holds information for data experiments and management configurations, aiding in operational governance.

This streamlined data architecture framework allows for effective data integration and governance, maximizing the value derived from diverse data sources. The data architecture diagram is shown in **Figure 2**.

5. Technical architecture

Big data governance platforms categorize data structures into three main types: structured, unstructured, and semi-structured data.

Structured data is stored in relational databases, characterized by a rigid schema of tables and columns that facilitate efficient storage, querying, and analysis.

Unstructured data includes diverse formats like audio, video, images, and documents, managed through distributed file systems without a pre-defined data model, making them richer in content but harder to process.

Semi-structured data [9], such as log files, XML, and JSON documents, combines elements of structured data with flexible markers for semantic separation and hierarchical organization.

Big data platforms leverage open-source technologies such as Hadoop to provide foundational storage and computing resources. These platforms support diverse data types through components like graph databases for complex data structures, object storage for scalability, and document storage for semi-structured data. Data handling is optimized with solutions like HBase for hot data and Hive for cold data storage, while data lake technology manages both real-time and offline data efficiently.

This streamlined architecture enables the platforms to handle various data types effectively, ensuring robust data management across modern, complex data environments. The technical architecture diagram is shown in **Figure 3**.

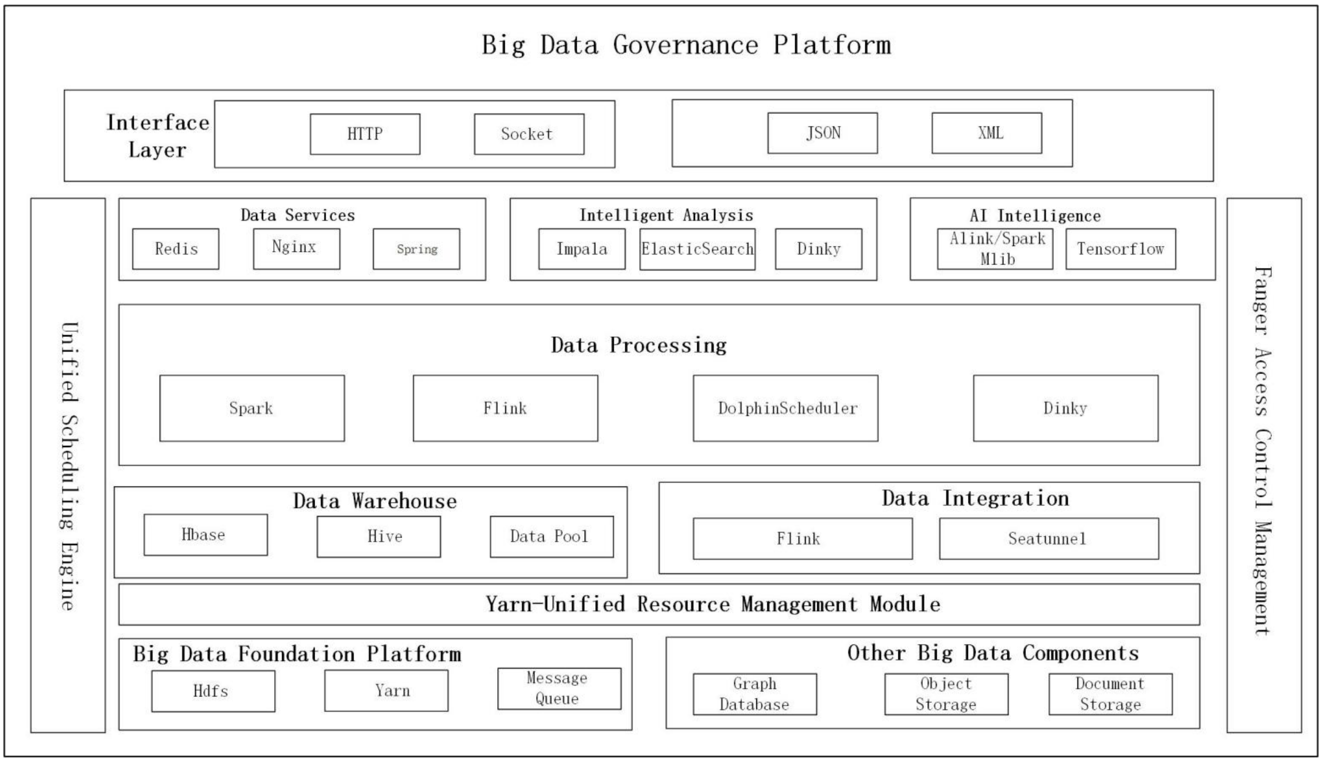

Figure 3. Technical architecture

6. Functional design

6.1. Data access

Data access utilizes a unified mode to standardize and modularize the integration of various data sources. This system probes data sources to understand their structure, quality, and scale, facilitating dynamic configuration of data access tasks such as strategy, task scheduling, and breakpoint continuation. It supports diverse sources including relational and non-relational databases, APIs, and file systems (FTP, SFTP, HDFS, S3), with methods ranging from real-time streaming to incremental batch access.

6.2. Data organization

Data is organized into multiple layers: raw, resource, subject, business, and knowledge repositories, each tailored to specific business needs. The raw repository holds unprocessed data, the resource repository cleanses and integrates data for analysis, the subject repository consolidates thematic data for deeper insights, the business repository supports specialized business analyses, and the knowledge repository facilitates queries on specialized data.

6.3. Data processing

Data processing standardizes data through extraction, cleansing, association, comparison, identification, and objectification, supporting both real-time and offline computation and batch processing. It incorporates AI for structured and unstructured data processing, using graph and in-memory computing to enhance data value. Model systems, tag engineering, and knowledge graphs further increase data value density, preparing and abstracting data for intelligent applications.

6.4. Data governance

Data governance includes standards, metadata management, data asset management, data security, data quality management, and operational management, ensuring high-quality data lifecycle management and operationalizing data value. It manages data standards, supports metadata for lineage and asset management, and ensures data security and quality, facilitating compliance and operational efficiency.

6.5. Data services

Data services offer internal data and API services externally, including publishing, subscription, approval, service management, and credential management. Publishing makes data available as APIs in the resource directory for subscription through the data portal. Subscription management approves and manages API services, while credential management controls access through AK/SK verification, ensuring data security and appropriate access.

7. Conclusion

The development and implementation of an advanced big data governance platform necessitate a nuanced understanding of data architectures, governance models, and the dynamic nature of big data. This paper has outlined a structured approach, integrating various data types through a unified system architecture that facilitates efficient data management and utilization. By layering data organization and employing cutting-edge technologies for data processing and standardization, the proposed governance platform addresses the multifaceted challenges of managing vast data landscapes. Central to our framework is a comprehensive data governance strategy, incorporating standards, metadata management, security, and quality controls to ensure the integrity and accessibility of data across its lifecycle. Additionally, the platform‘s data services broaden its applicability, offering scalable solutions for external data integration and API management.

Author contributions

Conceptualization: Yekun Chen
System platform building: Tianqi Xu
Data analysis: Yongjiang Xue

Disclosure statement

The authors declare no conflict of interest

References

[1] Chen XW, Lin X, 2014, Big Data Deep Learning: Challenges and Perspectives. IEEE Access, 2: 514–525.

[2] Zhang Q, Yang LT, Chen Z, et al., 2018, A Survey on Deep Learning for Big Data. Information Fusion, 42: 146–157.

[3] Li Y, Huang C, Ding L, et al., 2019, Deep Learning in Bioinformatics: Introduction, Application, and Perspective in the Big Data Era. Methods, 166: 4–21.

[4] Pansara R, 2023, Unraveling the Complexities of Data Governance with Strategies, Challenges, and Future Directions. Transactions on Latest Trends in IoT, 6(6): 46–56.

[5] Ridzuan F, Zainon WMNW, 2019, A Review on Data Cleansing Methods for Big Data. Procedia Computer Science, 161: 731–738.

[6] Chu X, Ilyas IF, Krishnan S, et al., 2016, Data Cleaning: Overview and Emerging Challenges. Proceedings of the 2016 International Conference on Management of Data, 2201–2206.

[7] Espinoza J, Xu NY, Nguyen KT, et al., 2023, The Need for Data Standards and Implementation Policies to Integrate CGM Data into the Electronic Health Record. Journal of Diabetes Science and Technology, 17(2): 495–502.

[8] Mohammed S, Nanthini S, Krishna NB, et al., 2023, A New Lightweight Data Security System for Data Security in the Cloud Computing. Measurement: Sensors, 29: 100856.

[9] Adeoye-Olatunde OA, Olenik NL, 2021, Research and Scholarly Methods: Semi-Structured Interviews. Journal of the American College of Clinical Pharmacy, 4(10): 1358–1367.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.