Reference Architecture for Big Data in the Public Sector

Victoria T. Oliveira[1], Rossana M. C. Andrade[1], Miguel Fraklin de Castro[1], Pedro Almir M. Oliveira[1,2], Maria Inês Vale Silva[1,4], Davyson S. Ribeiro[1,4], Ismayle S. Santos[1,3]

[1]Group of Computer Networks, Software Engineering, and Systems (GREat)
Federal University of Ceará (UFC) – Fortaleza, CE – Brazil

[2]Federal Institute of Maranhão (IFMA) – Pedreiras, MA – Brazil

[3]Ceará State University (UECE) – Fortaleza, CE – Brazil

[4]Finance Secretariat of Ceará (SEFAZ-CE) – Fortaleza, CE – Brazil

victoriat.oliveira@alu.ufc.br, rossana@ufc.br, miguel@ufc.br,
pedro.oliveira@ifma.edu.br, ines.vale@sefaz.ce.gov.br
davyson.ribeiro@sefaz.ce.gov.br ismayle.santos@uece.br

Abstract. *This article presents the proposal and evaluation of a Reference Architecture for Big Data Systems in the public sector, designed based on modern data engineering principles and aligned with approaches such as Data Mesh and the Medallion architecture. The architecture organizes data flow into well-defined layers, promoting scalability, governance, and data reuse throughout the analytical lifecycle. Additionally, it incorporates guidelines that foster team autonomy, process standardization, and improved communication among stakeholders. The evaluation involved 18 professionals with experience in Data Science and Big Data. The results indicate a positive perception regarding the clarity, organization, and applicability of the proposed architecture. Participants highlighted the clear definition of components and their support for task execution and team integration. Furthermore, the architecture demonstrated adaptability to different contexts and demands. It is concluded that the proposed architecture represents a viable alternative to support Big Data initiatives in the public sector, contributing to greater efficiency, organization, and quality in data usage.*

Resumo. *Este artigo apresenta a proposta e a avaliação de uma Arquitetura de Referência para Sistemas de Big Data no Setor Público, concebida com base em princípios modernos de engenharia de dados e alinhada a abordagens como Data Mesh e arquitetura Medallion. A arquitetura organiza o fluxo de dados em camadas bem definidas, promovendo escalabilidade, governança e reutilização de dados ao longo do ciclo analítico. Além disso, incorpora diretrizes que favorecem a autonomia das equipes, a padronização de processos e a melhoria da comunicação entre os atores envolvidos. A avaliação contou com a participação de 18 profissionais com experiência em Ciência de Dados e Big Data. Os resultados indicam uma percepção positiva quanto à clareza, organização e aplicabilidade da arquitetura proposta. Os participantes destacaram a definição clara*

dos componentes e seu suporte à execução das atividades e à integração entre equipes. Adicionalmente, a arquitetura demonstrou adaptabilidade a diferentes contextos e demandas. Conclui-se que a arquitetura proposta se apresenta como uma alternativa viável para apoiar iniciativas de Big Data no setor público, contribuindo para maior eficiência, organização e qualidade no uso de dados.

1. Introduction

In recent years, emerging technologies have significantly transformed Software Engineering, particularly with the rise of Big Data, which shifts the focus from functionality-oriented systems to data-centric approaches based on distributed processing and knowledge generation [Shahnawaz and Kumar 2025]. The exponential growth in data volume, velocity, and variety, combined with the increasing demand for data-driven decision-making, has driven the adoption of Big Data systems across multiple domains [Chamikara et al. 2020, Davoudian and Liu 2020]. These systems encompass the collection, management, analysis, and dissemination of large volumes of heterogeneous data, requiring scalable, resilient solutions capable of operating in distributed environments [Oliveira et al. 2025]. In this context, technologies such as large-scale processing, distributed storage, and automated pipelines become essential components to support the data lifecycle, from ingestion to value generation. Governments have increasingly adopted smart city initiatives to improve quality of life by leveraging data to support decision-making in areas such as healthcare and education. However, there remains a gap in the literature regarding structured processes to guide the use of Big Data in the public sector. In practice, decision-makers face challenges such as data integration, quality assurance, and the lack of clear guidelines for conducting analytical projects.

Nevertheless, applying traditional Software Engineering principles in this context significantly increases project complexity. Data-driven initiatives are often conducted in an ad hoc manner, with low levels of standardization, strong dependence on expert knowledge, and limited use of formal artifacts, which compromises reproducibility, institutional transferability, and the auditability of solutions [Warren and Marz 2015]. Furthermore, intrinsic characteristics of Big Data, Volume, Velocity, Variety, Veracity, and Value, amplify the impact of poor decisions made in early stages, particularly during data ingestion, integration, and modeling. Another critical challenge concerns data governance and data quality, which are essential to ensure reliable analyses and effective decision-making. The absence of clear guidelines may lead to inconsistencies, redundancies, and integration difficulties across different data sources and systems, ultimately hindering the strategic potential of data [Santos et al. 2023].

Given this scenario, there is a clear need for more structured and systematic approaches to the development of Big Data systems in the public sector. In this regard, the definition of reference architectures becomes particularly important, as they provide guidelines, standards, and best practices to support the design, development, and evolution of such systems. These architectures contribute to reducing complexity, improving interoperability, and fostering more robust, sustainable solutions aligned with Software Engineering principles.

2. Software Engineering, Architecture, and Big Data

Software Engineering provides a systematic body of principles, practices, and artifacts aimed at reducing uncertainty, standardizing decisions, and transforming system development into a repeatable and controllable process. It is grounded in three essential elements: processes, methods, and tools. Among the artifacts produced throughout this process, software architecture stands out as a key element. Architecture represents the high-level organizational structure of a system, encompassing its main components, the relationships among them, and the principles that guide its design and evolution [PRESSMAN 2021].

In this sense, when Software Engineering principles, particularly processes and their associated activities, are applied to the Big Data context, the complexity of the challenge increases significantly. Projects driven by advanced analytics and data science are often conducted in an ad hoc manner, with strong reliance on expert intuition and poorly formalized artifacts, resulting in short-lived solutions, low institutional transferability, and limited auditability. Furthermore, intrinsic characteristics of Big Data systems, such as volume, variety, velocity, and the volatility of data sources, amplify the cost of poorly grounded decisions made in early stages, especially during data ingestion and canonical data modeling [Warren and Marz 2015].

Big Data is not characterized solely by its large volume, but also by its complexity and heterogeneity, encompassing multiple data types and formats that challenge the capabilities of systems in terms of storage, retrieval, sharing, visualization, and analysis. Fundamental Software Engineering principles, such as modularity, abstraction, reuse, and maintainability, must be reconsidered in the context of Big Data [Wu et al. 2016]. Systems dealing with large-scale data must rapidly adapt to new technologies and continuously evolving organizational demands, while maintaining efficiency and reliability.

In this context, the application of Software Engineering foundations to Big Data systems is not merely a superficial adaptation, but a structural necessity. In summary, Software Engineering for Big Data goes beyond the use of data tools; it involves the design of sociotechnical systems aimed at producing evidence in a disciplined, predictable, and auditable manner. This becomes particularly relevant when such systems support the formulation of data-driven public policies, where verifiability, explainability, and reproducibility are no longer desirable attributes but institutional requirements [Kleppmann 2017].

3. Context

The research was conducted in a real-world public sector context, focusing on the use of data to support decision-making and governance. The proposal is grounded and validated through two initiatives: the BigDataFortaleza Platform, aimed at integrating and analyzing municipal data for monitoring indicators and supporting strategic planning, and the Sophia-Q Platform, focused on software quality governance at the state level.

While BigDataFortaleza supports public policies, particularly those related to Early Childhood, Sophia-Q transforms data from the software development lifecycle into managerial indicators. Together, these initiatives demonstrate the practical applicability, versatility, and generalization potential of the proposed approach across different domains within the public sector.

3.1. Platform Big Data Fortaleza

IPPLAN plays a strategic role in urban planning and public policy evaluation in Fortaleza, highlighting the need to strengthen the use of data, especially in areas such as Early Childhood. In this context, the Big Data Fortaleza Platform was created through a partnership between IPPLAN and GREat, aiming to integrate data from multiple sectors (health, education, social assistance, and planning). This integration enables intersectoral analysis and supports the monitoring of strategic indicators aligned with the Fortaleza 2040 plan.

More than a technological solution, this initiative consolidated a structured model for data-driven analysis, incorporating governance, standardization, and evidence generation to support decision-making. As a result, it enabled concrete actions, such as strategies for childhood vaccination. Throughout its evolution, the reference architecture was progressively developed and refined based on practical challenges, culminating in a multi-layered model.

3.2. Platform SophiaQ

In the context of software quality governance in the public sector, the Sophia-Q Platform was conceived as a project developed by SEFAZ-CE in partnership with GREat, within the scope of the Chief Scientist Project. The Sophia-Q Platform was designed to enhance managerial visibility over the quality and performance of software development processes. The initiative integrates data from multiple tools (such as task management, version control, and testing), transforming dispersed operational data into structured, traceable, and auditable information to support decision-making.

Through managerial dashboards and indicators, the platform enables continuous and evolutionary analysis, contributing to the standardization, measurement, and improvement of development processes. In addition, Sophia-Q reuses the reference architecture, the data analysis process, and the requirements elicitation model developed in Big Data Fortaleza, demonstrating the flexibility and adaptability of the approach across different domains, in this case, software quality governance.

4. Related Work

The analyzed studies highlight the evolution of modern data architectures, with emphasis on the Data Mesh and Data Lake paradigm and scalable Big Data approaches applied to the public sector. [Ramos et al. 2022] focuses on data use in the governmental context, emphasizing the presentation and consumption layer. Architecturally, it does not propose a complete reference architecture or detail core data layers, concentrating instead on organizing information for dashboards with an emphasis on clarity and decision support.

In the field of integration with artificial intelligence, [Mishra et al. 2024] presents a Data Mesh architecture enriched with large language models (LLMs), knowledge graphs, and multi-agent systems, aimed at analyzing large volumes of data in scenarios such as public works and government procurement. Similarly, [Ismail et al. 2025] proposes a layered Big Data framework encompassing data ingestion, transformation, storage, analysis, machine learning, and governance, with particular attention to compliance with regulations such as the LGPD.

In light of this landscape, the architecture proposed in this work directly aligns with the literature by combining a layered approach with the Medallion pipeline, struc-

turing the system around data flows and analytical capabilities. Furthermore, it incorporates Data Mesh principles, particularly regarding distributed governance among different stakeholders, promoting domain autonomy while ensuring control and compliance. This combination strengthens governance practices, auditability, and scalability in Big Data solutions within the public sector. A comparative summary is presented in Table 1.

Table 1. Comparison between the analyzed studies and the proposed architecture

Paper	**Main Approach**	**Key Contributions**	**Governance Focus**
Ramos et al. (2022)	Data Lakes	Provides guidelines for organizing information visually, emphasizing clarity.	Limited
Mishra et al. (2024)	Data Mesh	Integration of LLMs, knowledge graphs, and multi-agent systems	Moderate
Ismail et al. (2025)	Layered Big Data Architecture	Comprehensive framework including ingestion, processing, analysis, ML, governance, and LGPD compliance	Strong (regulatory compliance)
Proposed Architecture	Layers + Medallion + Data Mesh	Integration of the Medallion pipeline, data flow-oriented organization, distributed governance, and analytical support	Strong (distributed and auditable)

5. Reference Architecture for Big Data in the Public Sector

The proposed architecture, illustrated in Figure 1, is grounded in a top-down approach, in which the architectural definition begins with the identification of functional and non-functional requirements, as well as user profiles, prior to the selection of technologies. This process takes into account the heterogeneity of stakeholders involved, such as public managers, administrators, data scientists, and external systems, enabling the solution to be structured across multiple levels of abstraction and different points of value generation.

Based on these profiles, the architecture is organized in a modular and multi-layered manner, combining architectural domains and functional layers. On one hand, four main domains are defined, IT infrastructure, support services, Big Data services, and web application (backend and frontend), aiming to reduce coupling and enable the independent evolution of components. On the other hand, the architecture is detailed into operational layers that structure the system's functioning in a secure, scalable, and efficient way.

At the outermost layer are the users and systems that interact with the platform, responsible for providing and consuming data through interfaces and APIs. Next, the network layer acts as a protection and traffic distribution mechanism, incorporating ele-

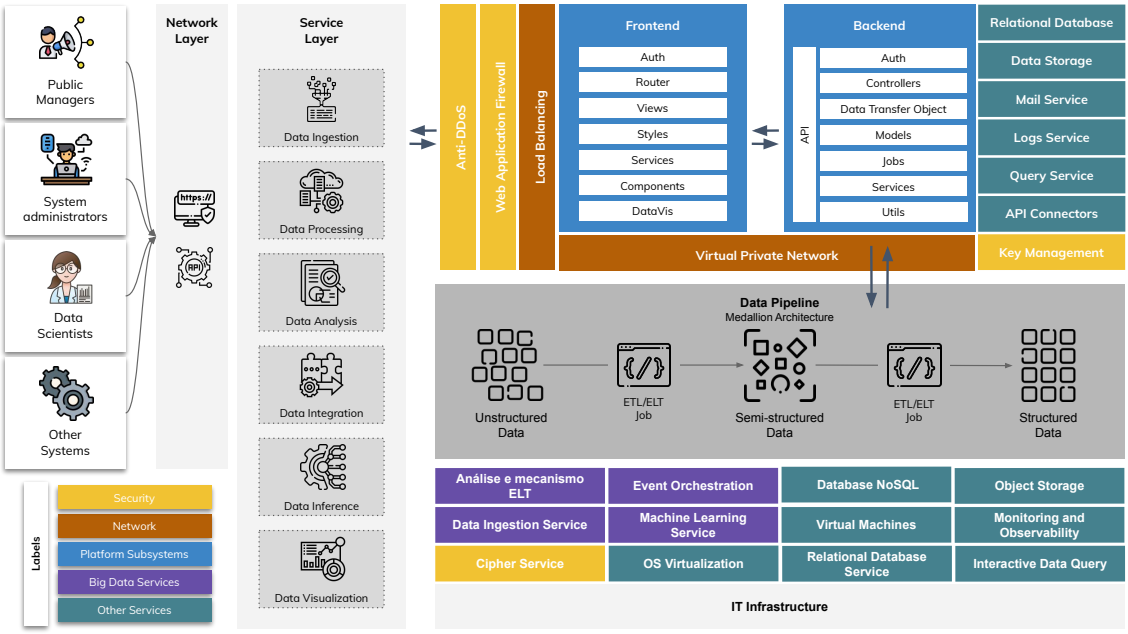


Figure 1. AReference Architecture.

ments such as firewalls, load balancing, and attack mitigation, highlighting that security is treated as a cross-cutting capability rather than an isolated component.

The core of the architecture lies in the services layer, which is responsible for the data lifecycle, including ingestion, processing, integration, analysis, inference, and visualization. This organization reinforces a data flow-oriented design, in which value is progressively generated throughout the pipeline. Additionally, the frontend and backend modules are structured in a decoupled manner, the frontend focuses on user interaction and data visualization, while the backend manages business logic, request processing, security, and the execution of asynchronous tasks.

The data and storage layer ensures data persistence and integrity by using both relational and non-relational databases, as well as services for logging, messaging, integration, and cryptographic key management. This layer supports the data pipeline, which can operate under ETL or ELT models, maintaining flexibility for different organizational contexts.

Regarding data processing, the Medallion architecture (Bronze, Silver, and Gold) is adopted, organizing data evolution into progressive levels of refinement. Unlike traditional approaches, this model emphasizes governance, traceability, and explainability of transformations, making intermediate data states explicit throughout the pipeline. This characteristic is particularly relevant in the public sector, where transparency and auditability are fundamental requirements.

Furthermore, the architecture includes a set of specialized services, such as analytical engines, event orchestration, machine learning, monitoring, and object storage, which operate independently of the physical infrastructure. Finally, the IT infrastructure forms the foundation that supports all layers, including servers, networks, and storage resources, ensuring performance, scalability, and reliability.

The proposed architecture is not driven by specific tools, but by capabilities and the data value flow, integrating modularity, governance, security, and flexibility. This approach ensures architectural consistency and enables adaptation to the technological, organizational, and analytical needs of different contexts, particularly in the public sector.

6. Evaluation of the Reference Architecture for Big Data in the Public Sector

The evaluation of the reference architecture was conducted based on the Goal-Question-Metric (GQM) approach, enabling the analysis of its quality according to criteria such as clarity, usability, completeness, standardization, reproducibility, and agility. In the evaluation instrument, the architecture was primarily assessed from a structural perspective, considering:

- data organization through a layered architecture;
- support for communication among teams;
- adaptability of the architecture to different contexts;
- support for traceability in the analytical data flow.

The architecture was applied and evaluated in different real-world contexts (Big Data Fortaleza and SEFAZ-CE projects), which enabled the verification of its practical feasibility and adaptability across distinct organizational environments. The analysis of the results considered both quantitative and qualitative data, allowing the assessment of the architecture's consistency, applicability, and adoption potential. Additionally, the comparison across different participant profiles contributed to identifying perceptions regarding its ease of understanding and practical usefulness.

6.1. Research Questions

The evaluation instrument was structured based on the GQM planning and incorporated constructs related to clarity, usability, completeness, standardization, reproducibility, and agility. Additionally, principles from the Technology Acceptance Model (TAM) were considered, particularly regarding perceived usefulness and ease of use, without addressing behavioral aspects of adoption.

Data collection was carried out through a questionnaire composed of closed-ended questions using a five-point Likert scale (1 – Strongly disagree to 5 – Strongly agree) and open-ended questions, enabling both quantitative and qualitative analyses. The instrument covered aspects such as the definition of components and layers, data organization, support for communication among teams, adaptability, and traceability in the analytical data flow.

6.2. Case Study and Participant Selection

The case study was conducted within the Big Data Fortaleza project. The team, organized under a hybrid model based on Scrum, was involved in the development of the platform, encompassing activities such as requirements engineering, user experience, data science, development, and testing. To enhance the validity of the evaluation, the reference architecture was subsequently applied in the Data and Software Quality project at SEFAZ-CE, within a similar context, also involving hybrid and multidisciplinary teams.

Additionally, participants with experience in Big Data and Data Science projects, but without prior exposure to the architecture, were included. Participants were divided

into two groups, with and without prior experience, enabling the analysis of both practical application and perceptions of usefulness and ease of understanding. This strategy contributed to a more comprehensive evaluation of the architecture's adoption feasibility.

6.3. Data Collection and Analysis Procedure

Data collection was conducted after presenting the architecture to the group without prior experience and, for the experienced group, after its application in a real-world context. Data analysis involved descriptive statistics for the closed-ended questions and qualitative analysis through thematic categorization of open-ended responses. Additionally, a comparison between the groups was performed to identify differences in perceptions regarding usefulness, clarity, and applicability. The integration of these results enabled a more robust evaluation of the architecture's adoption feasibility.

The evaluation of the reference architecture was conducted based on the following closed-ended questions using a five-point Likert scale (1 – Strongly disagree to 5 – Strongly agree):

- (Q1) The reference architecture is easy to understand?
- (Q2) The components and layers of the architecture are well defined?
- (Q3) The organization of architectural components contributes to greater agility in task execution?
- (Q4) The architecture facilitates communication among different teams?
- (Q5) The architecture demonstrates adaptability to new scenarios?
- (Q6) The layered structure, based on the Medallion Architecture, contributes to data organization?
- (Q7) The architecture based on the Medallion Architecture contributes to reducing the execution time of tasks related to data processing and analysis?
- (Q8) The architecture supports data traceability throughout the analytical flow?

The following open-ended questions were also included:

- Which characteristics of the architecture do you consider most relevant?
- Did you identify any limitations or points for improvement in the architecture?
- What improvements or adjustments would you suggest to make the architecture more suitable for organizational practice?

7. Results and Discussion

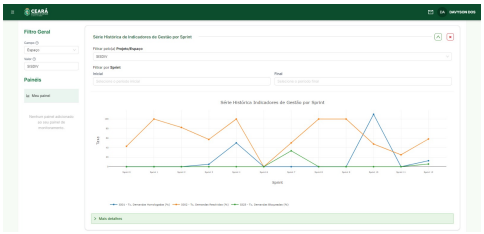
In this section, we present the implementation results of the Big Data Fortaleza and Sophia-Q Platforms (Subsection 7.1) and the evaluation results of the Reference Architecture for Big Data in the Public Sector (Subsection 7.2).

7.1. Implementation in the Big Data Fortaleza and Sophia-Q Platforms

The Big Data Fortaleza and Sophia-Q platforms were developed based on the proposed reference architecture. The architectural principles, layers, and components defined in this work guided the design and implementation of both platforms, enabling the structuring of data flows, the integration of heterogeneous data sources, and the support for analytical processes. This adoption demonstrates the practical applicability of the architecture in real-world scenarios, contributing to improved data organization, governance, and decision-making support in the public sector.

Figure 2. Big Data Fortaleza and Sophia-Q Platforms.

(a) Big Data Fortaleza Platforms



(b) Sophia-Q Platforms

7.2. Evaluation Results

The evaluation of the Reference Architecture for Big Data Systems involved 18 professionals, predominantly with technical backgrounds in Data Science and software development. There was a balanced distribution between Data Scientists/Researchers (55.6%) and Developers (44.4%), indicating a sample directly engaged in data-driven solutions. Regarding academic background, a high level of qualification was observed among the participants. Specifically, 33.3% were undergraduate students, 22.2% held a bachelor's degree, 16.7% were pursuing a master's degree, 16.7% had completed a master's degree, and 11.1% held a doctoral degree. These results indicate that a significant portion of the participants had graduate-level education (completed or in progress), reinforcing the suitability of the sample to evaluate conceptual and technical aspects related to a reference architecture for Big Data environments.

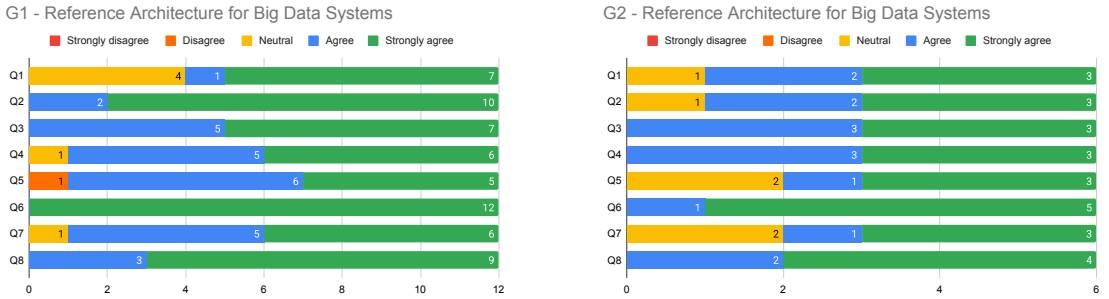
Regarding experience, most participants reported between 3 to 5 years of experience in Big Data, demonstrating relevant practical expertise in the field. In terms of context, 66.7% of the participants were involved in institutional projects such as Big Data Fortaleza and SophiaQ, while 33.3% participated in other academic and institutional initiatives, ensuring diversity of perspectives. Finally, participants were organized into two groups: (G1) professionals with prior experience using the process in the Big Data Fortaleza and SophiaQ projects, and (G2) professionals with no prior exposure to the process.

The comparative analysis between groups G1 (participants with prior knowledge of the architecture) and G2 (participants without prior knowledge) provides a deeper understanding of the perception of the proposed reference architecture for Big Data systems, considering different levels of familiarity.

Overall, the results, which can be observed in Figure 3, indicate a broadly positive evaluation across both groups, with a predominance of responses in the categories partially "agree" and "strongly agree". However, a consistent difference in the level of agreement can be observed: G1 shows responses more concentrated in "strongly agree", whereas G2 presents a more balanced distribution among "partially agree", "strongly agree", and, in some cases, "neutral". This behavior suggests that prior familiarity with the architecture contributes to a more consolidated perception of its benefits.

In G1, Q1 (ease of understanding) already presents a predominantly positive scenario, although with some neutral responses, indicating that even for experienced users, the initial clarity of the architecture may require some cognitive effort. Nevertheless, for Q2 (definition of components and layers) and Q3 (organization contributing to agility),

Figure 3. Evaluation of the Reference Architecture for Big Data Systems.



(a) Group 1 Professionals

(b) Group 2 Professionals

there is a significant shift toward higher levels of agreement, showing that once understood, the architecture is perceived as well-structured and functional.

Q4 (facilitation of communication among teams) and Q5 (adaptability) also present positive evaluations, though with a higher incidence of "partially agree", suggesting that these aspects, while recognized, may depend on contextual factors such as organizational maturity or data governance practices. Still, there is no significant disagreement, reinforcing the adequacy of the architecture in these aspects.

One of the strongest points observed in G1 is Q6, related to the adoption of the Medallion Architecture, which achieved unanimous "strongly agree" responses. This result is particularly relevant, as it highlights that the layered organization (Bronze, Silver, and Gold) is perceived as highly effective for structuring and organizing data. Q7 (reduction in execution time) and Q8 (data traceability) also show strongly positive evaluations, indicating that the operational and analytical benefits of the architecture are clearly recognized by experienced users.

In G2, although the general trend is also positive, there is a greater presence of neutral responses, especially in Q1 and Q2. This suggests that the initial understanding of the architecture and the clear identification of its components may not be immediate for users without prior experience. This result points to a potential need for improvement in how the architecture is presented, documented, or taught.

On the other hand, as the questions address more practical and tangible aspects, such as organization (Q3), communication (Q4), and especially the layered structure of the Medallion Architecture (Q6), there is an increase in "strongly agree" responses even in G2. This demonstrates that, although initial conceptual understanding may be challenging, the benefits of the architecture become evident when analyzed from a practical application perspective.

Furthermore, the questions related to performance (Q7) and traceability (Q8) also show positive evaluations in G2, albeit with slightly lower intensity compared to G1. This result reinforces that the architecture is able to convey value even to users who do not yet fully master its concepts.

Overall, the results indicate that the reference architecture is well evaluated in terms of organization, usefulness, and support for data analysis activities, regardless of

prior knowledge level. However, the observed differences between groups show that familiarity with the architecture enhances the perception of its benefits, particularly in more abstract aspects such as structural clarity and adaptability.

Finally, the presence of neutral responses in G2 highlights opportunities for improvement, especially regarding the communication of the architecture, the clarity of its documentation, and the training of new users. Nevertheless, the predominance of positive evaluations in both groups reinforces the robustness and applicability of the proposed architecture in the context of Big Data systems.

To complement the evaluation, a qualitative analysis was conducted based on three open-ended questions, aiming to understand participants' perceptions of the proposed architecture. The results indicate that the main strengths of the proposed architecture lie in its layered structure based on the Medallion model (Bronze, Silver, Gold), which facilitates data organization, understanding of processing stages, and traceability. Participants also highlighted modularity, clear separation of responsibilities, scalability, and embedded security mechanisms as key advantages, along with alignment with best practices. Regarding limitations, the main gaps identified relate to data governance and real-time processing, including the need for better metadata management, data lineage, data catalogs, and data quality monitoring. As for improvements, suggestions focused on strengthening governance, improving architectural clarity, and incorporating mechanisms such as quality gates, automated data validation, and alert systems. Overall, the recommendations emphasize enhancing governance, data quality, and automation to improve the architecture's applicability across different contexts.

8. Conclusion

The exponential growth in the volume, velocity, variety, veracity, and value of data has intensified the relevance of Big Data solutions in the public sector, while also highlighting ongoing challenges related to the definition of architectures capable of addressing scalability, governance, interoperability, and efficient data flow organization. In this context, this work presented the proposal and evaluation of a Reference Architecture for Big Data Systems, grounded in contemporary data engineering principles and aligned with modern approaches such as Data Mesh and the Medallion architecture.

The results obtained from the evaluation with domain professionals indicate that the proposed architecture is understandable, well-structured, and applicable to different contexts, contributing to the organization of data flow and improving communication among teams. The clear definition of layers and components proved to be a relevant factor in supporting task execution and promoting greater efficiency in data usage. Additionally, the architecture demonstrated adaptability to different scenarios, reinforcing its flexibility in complex, data-driven environments.

As future work, it is suggested to apply the architecture in real-world case studies, as well as to evolve it through the incorporation of new patterns and emerging technologies, further expanding its practical validation and contribution to the advancement of Big Data solutions in the public sector.

9. Acknowledgment

We would like to thank FUNCAP for the financial support provided to this project and CNPq for the productivity grants awarded to Rossana M. C. Andrade (306362/2021-0) and Ismayle Santos (312173/2025-3).

References

Chamikara, M., Bertok, P., Liu, D., Camtepe, S., and Khalil, I. (2020). Efficient privacy preservation of big data for accurate data mining. *Information Sciences*, 527:420–443.

Davoudian, A. and Liu, M. (2020). Big data systems: A software engineering perspective. 53(5).

Ismail, F. N., Sengupta, A., and Amarasoma, S. (2025). Big data architecture for large organizations. *arXiv preprint arXiv:2505.04717*.

Kleppmann, M. (2017). *Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems*. " O'Reilly Media, Inc.".

Mishra, S., Shinde, M., Yadav, A., Ayyub, B., and Rao, A. (2024). An ai-driven data mesh architecture enhancing decision-making in infrastructure construction and public procurement. *arXiv preprint arXiv:2412.00224*.

Oliveira, V., Andrade, R., Castro, M., and Santos, I. (2025). From acquisition to interpretation: A model for creating data storytelling in big data. In *Anais do XIII Latin American Symposium on Digital Government*, pages 227–236, Porto Alegre, RS, Brasil. SBC.

PRESSMAN, Roger S.; MAXIM, B. R. (2021). *Engenharia de Software*. McGraw-Hill Brasil, [S. l.], 9 edition.

Ramos, G., Fernandes, D., Coelho, J., and Aquino, A. (2022). Data lakes lógicos como plataformas para dados governamentais em sociedades e cidades inteligentes. In *Anais do X Workshop de Computação Aplicada em Governo Eletrônico*, pages 215–226, Porto Alegre, RS, Brasil. SBC.

Santos, I., Oliveira, P., Oliveira, V., Nogueira, T., Dantas, A., Menescal, L., Élcio Batista, and Andrade, R. (2023). Big data fortaleza: Plataforma inteligente para políticas públicas baseadas em evidências. In *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*, pages 200–211, Porto Alegre, RS, Brasil. SBC.

Shahnawaz, M. and Kumar, M. (2025). A comprehensive survey on big data analytics: Characteristics, tools and techniques. *ACM Comput. Surv.*, 57(8).

Warren, J. and Marz, N. (2015). *Big Data: Principles and best practices of scalable realtime data systems*. Simon and Schuster.

Wu, J., Guo, S., Li, J., and Zeng, D. (2016). Big data meet green challenges: Greening big data. *IEEE Systems Journal*, 10(3):873–887.