Reference Process for Integrating Data Science Workflows and Governance in Big Data Systems

Victória T. Oliveira
Computer Networks, Software and Systems Engineering Group (GREat),
Federal University of Ceará (UFC)
Fortaleza, CE, Brazil
victoria.oliveira@great.ufc.br

Rossana M. de Castro Andrade
Computer Networks, Software and Systems Engineering Group (GREat),
Federal University of Ceará (UFC)
Fortaleza, CE, Brazil
rossana@ufc.br

Pedro Almir Martins Oliveira
Computer Networks, Software and Systems Engineering Group (GREat),
Federal Institute of Education, Science and Technology of Maranhão (IFMA)
Pedreiras, MA, Brazil
pedromartins@great.ufc.br

Ismayle Santos
Computer Networks, Software and Systems Engineering Group (GREat),
State University of Ceará (UECE)
Fortaleza, CE, Brazil
ismayle.santos@uece.br

Miguel Franklin de Castro
Computer Networks, Software and Systems Engineering Group (GREat),
Federal University of Ceará (UFC)
Fortaleza, CE, Brazil
miguel@ufc.br

ABSTRACT

This article presents a structured workflow for the development of data-based analysis in the public sector, from the identification of the institutional challenges to evidence-based decision-making. The proposed process organizes the steps around three main profiles - public manager, administrator and data scientist - and integrates data governance practices, LGPD compliance and reproducible analytical processes. The technical method for adding new analytics to the institutional platform is also detailed, involving the preparation of scripts, the use of machine learning models and the publication of analytical products with automated visualizations. The proposal contributes to standardization, transparency and efficiency in the adoption of analytical intelligence by public bodies, promoting evidence-based decisions.

KEYWORDS

Big Data Systems, Software Engineering, Data Sciense, Workflows, Governance, Public Sector

1 Introduction

The huge increase in the volume and speed of data, coupled with the growing demand for automation, has made Big Data technology an efficient solution to many of today's challenges, including public management. This technological transformation has contributed to the popularization of Big Data systems. The development of software systems incorporating Big Data components is on the rise and is being exploited in various sectors to aid decision-making. Big Data systems comprise scalable software technologies in which large amounts of heterogeneous data are collected from multiple sources, managed, analyzed and delivered to end users and/or external applications [8]. These systems have brought several challenges to software development, for example, challenges related to governance.

The growing demand for evidence-based decision-making in public agencies has driven the adoption of Big Data systems and intelligent analytical solutions. The strategic use of data has become a key differentiator for modern public management. The ability to transform large volumes of data into useful analytical products allows public managers to make more precise, efficient, and transparent decisions [5]. However, this process requires a well-structured workflow of interaction between different institutional and technical actors, respecting governance and data protection standards, such as the General Data Protection Act (LGPD).

Data analysis is used to extract hidden insights and information from academics, practitioners, and managers to encourage decision-making in smart cities. In this context, Big Data analytics stands out as a crucial enabler for data-driven decision-making. Studies [1, 7, 16] have demonstrated the benefits of data-driven decision-making and its positive impact on organizational productivity.

Governments and municipalities are implementing smart city projects to address challenges in education, healthcare, social services, and more. These projects generally aim to provide citizens with a better quality of life. Given the challenges faced by the city of Fortaleza, data collection and analysis are essential to diagnose problems and support the decision-making of municipal managers. Through intelligent analysis of these data, it is possible, for example, to improve the quality of services provided to citizens and predict the impacts of changes in urban infrastructure. Therefore, it is necessary to create a reference process for data analysis in public agencies, describing the entire workflow, from identifying challenges to decision-making.

This article presents a governance process for developing Big Data systems in public agencies. This process presents a workflow for data analysis in public agencies, from problem identification to decision-making. It also details the technical method for adding new analyses to the institutional platform, led by data scientists. The goal is to demonstrate how structured processes enable secure, action-oriented data with data governance.

2 Big Data for Evidence-based Public Policies

Big data and data analytics are new paradigms in public administration practices. When implemented correctly, they produce positive results in public administration, including effectiveness, efficiency, and citizen satisfaction [3]. These advantages result from a considerable increase in the accuracy of decision-making, which means managers can make more accurate decisions based on facts rather than assumptions. Evidence-based decision-making, in a management context across all sectors, is described by [6] as "...a dynamic process through which evidence is obtained, interpreted, and used as a basis for decision-making". Evidence must be generated through a rigorous process and must be relevant to the context in which it is invoked.

Currently, various aspects of human life can be stored in data form. As people's lives become increasingly digital, the amount of electronic data created globally is increasing. By combining this information with Big Data methods, this vast amount of data can be stored, processed, and transformed into information that helps public administration implement more efficient policies [11].

New methods of collecting, storing, and processing data to improve its accessibility, maintainability, and visualization are being developed simultaneously [13]. New analytical and logical techniques are also being actively developed. The ability to analyze data has reached unprecedented heights due to advances in information technology, both in terms of hardware and software [14]. The availability of data and innovative approaches are increasingly understood in the public and corporate sectors [18]. Several public administrations have implemented big data strategies or policies. In terms of efficiency and effectiveness, big data and strategies for using it are emerging phenomena in the management landscape that yield excellent results [10].

New big data analysis techniques can help governments better understand their population's behavior and improve public services [9]. According to contemporary research on the topic, big data can be used in a variety of ways to improve public sector outcomes in areas such as health, education, social welfare, etc. This includes increasing government efficiency and effectiveness, making more informed policy decisions, and providing better services based on data and scientific evidence [15].

3 Related Work

Although the focus is conceptual, [4] presents clear guidelines for implementing a Data Mesh. The article emphasizes that a successful Data Mesh implementation requires clearly defined roles, with decentralized responsibilities, but aligned by organizational standards and policies. The article also emphasizes the Data Product Lifecycle, each data product in the Data Mesh is treated as a software asset, with a different structured lifecycle. Furthermore, the author recommends, interoperability between domains and systems is critical in Data Mesh, especially in public environments where there are legacy systems and multiple sources.

In [17], data governance is treated as one of the fundamental pillars for the success of a self-service data platform in a Data Mesh-oriented environment. The proposal goes beyond technical controls and is structured on three main fronts: roles, policies and auditing, with specific architectural decisions for each. The article highlights the importance of clearly assigning roles within the data fabric, aligning them with the decentralization proposed by Data Mesh.

The work [12] is based on the modern concept of Data Mesh, which breaks with traditional centralized data management models (such as data lakes or single data warehouses) and adopts a decentralized approach. This means that instead of having all the data centralized in a single technical team, each business domain (such as transport, public works, government procurement, energy, etc.) is responsible for its own data products. Each domain acts as a producer of data (e.g. tender contracts, construction schedules) and also as a consumer of data from other domains (e.g. the construction sector can use climate risk data from the environmental domain).

The cited articles share common ground with this work, particularly their emphasis on decentralized governance, data products, and autonomous domains within the Data Mesh framework. However, this work differs in its procedural approach, the way it defines the challenge, the formalization of operational roles, and its concrete applicability to the Brazilian public sector.

4 Context

At the City Hall of Fortaleza, the Planning Institute of Fortaleza serves as one of the main municipal agencies, with the mission of producing knowledge, monitoring and evaluating public policies, coordinating strategic planning, and fostering innovative initiatives.

To guide the capital's urban development, the City Hall developed a strategic plan with short, medium, and long-term goals, structured around a vision that extends to 2040. This effort resulted in the creation of Fortaleza 2040[1], a participatory strategic plan that seeks to integrate physical and territorial development with social and economic progress. The initiative promotes discussions about the city from various perspectives, involving different sectors, territories, and levels of government. Among the topics addressed by the plan are urgent issues such as childhood vaccination, infant mortality, and prenatal care.

Among the challenges identified, early childhood care stands out, with a special focus on children aged from zero to six. Because it is a cross-cutting area, it requires integrated actions across health, education, social assistance, and other sectors. In this context, considering the challenges faced by the city and the strategic role of Fortaleza, the use of data collection and analysis as diagnostic and decision-making tools becomes essential. Intelligent data analysis enables the improvement of public service offerings, contributing to improving the population's quality of life. Therefore, it is essential to have a robust infrastructure to collect, store, and process data from various sources.

In this scenario, the Big Data Fortaleza platform[2] was developed, focused on analyzing data related to early childhood public policies. The platform allows for the continuous monitoring of key indicators, ideally aligned with international standards, providing a solid basis for data- and evidence-driven decisions.

In the area of education, ensuring access to daycare centers and schools specialized in early childhood education became a priority. In the health field, comprehensive care for pregnant women—from

[1] https://fortaleza2040.fortaleza.ce.gov.br/site/

[2] https://bigdata.fortaleza.ce.gov.br/

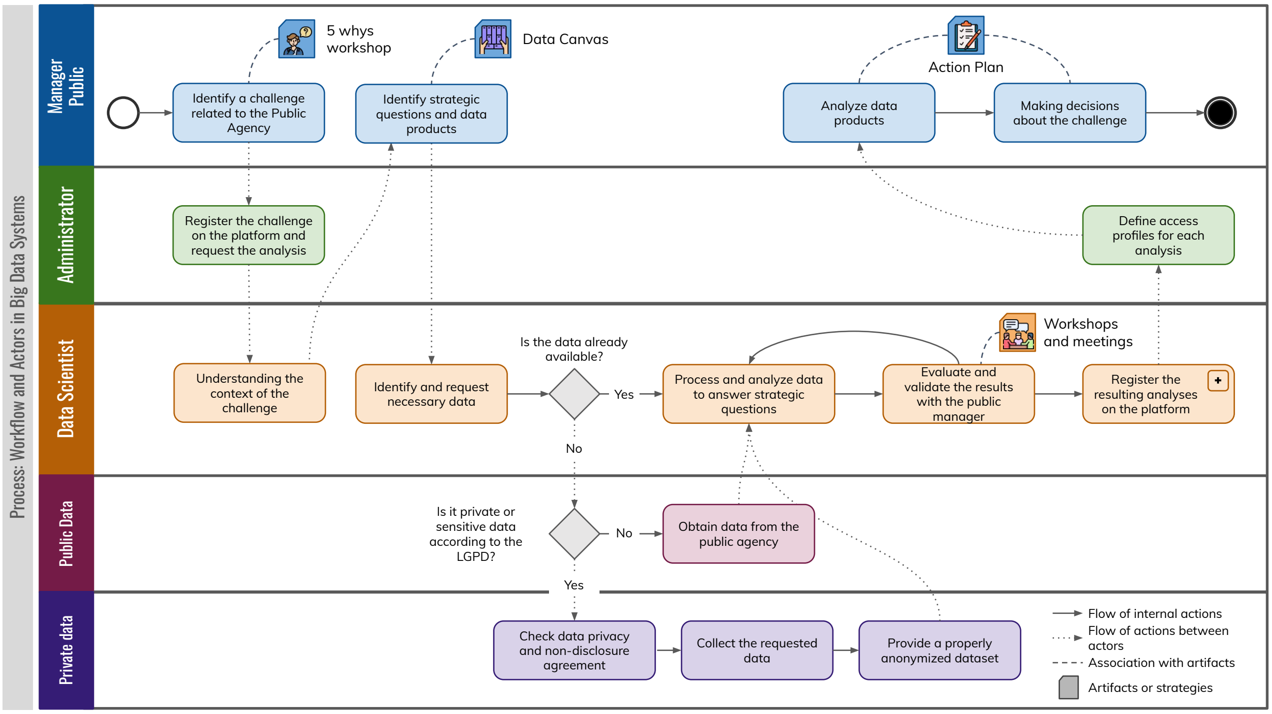
Process: Workflow and Actors in Big Data Systems
Public Manager
Administrator
Data Scientist
Public Data
Private data
5 whys workshop
Data Canvas
Action Plan
Identify a challenge related to the Public Agency
Identify strategic questions and data products
Analyze data products
Making decisions about the challenge
Register the challenge on the platform and request the analysis
Define access profiles for each analysis
Understanding the context of the challenge
Identify and request necessary data
Is the data already available?
Yes
No
Process and analyze data to answer strategic questions
Evaluate and validate the results with the public manager
Workshops and meetings
Register the resulting analyses on the platform
+
Is it private or sensitive data according to the LGPD?
No
Yes
Obtain data from the public agency
Check data privacy and non-disclosure agreement
Collect the requested data
Provide a properly anonymized dataset
Flow of internal actions
Flow of actions between actors
Association with artifacts
Artifacts or strategies

Figure 1: Process: Workflow and Actors for Big Data Systems in the Public Sector.

prenatal to postpartum—and newborns, with a focus on vaccination, was identified as essential. In the context of human rights and social development, it was necessary to identify families in vulnerable situations, including those experiencing homelessness, so they could be supported through social benefits. These actions aim to reduce inequalities and ensure access to services and fundamental rights, especially for children and pregnant women.

5 Process: Workflow and Actors for Big Data Systems in the Public Sector

Figure 1 details the workflow for developing analyses in Big Data systems, aimed at public sectors, highlighting the main players involved, the actions and the interaction between public and private data, in accordance with the rules of the LGPD (General Personal Data Protection Law). The workflow is organized into five horizontal bands, representing the different roles (actors), such as Public Manager, Administrator and Data Scientist, as well as types of data (e.g., Public Data and Private Data). It also follows a process, from identifying a challenge to making a decision based on data.

The process begins with the Public Manager, who is responsible for identifying a relevant institutional challenge. This challenge must be aligned with the body's strategic needs. Two workshops are used to structure this phase [2]:

- 5 Whys Workshop: an investigative method that seeks to identify the root causes of the problem presented;
- Data Canvas: structures the problem into strategic questions and defines the expected data products (indicators, analyses, reports, etc.).

These inputs are essential to ensure that the analysis has focus, value and concrete applicability. Once the scope has been defined, the Administrator is responsible for registering the challenge on the institutional analysis platform and formalizing the analysis request to the data science team.

After formalization, the Data Scientist plays a central role in converting strategic questions into visualizations (analytics). Their activities include: Understanding the context and technical depth of the proposed challenge; Identifying and requesting the necessary data, based on the requirements of the problem; and Checking the availability of the data. If already available, the data scientist proceeds with processing and analysis. If not available, he/she formally requests the data from the responsible public agency.

Generally, in public sectors, we have public data and private or sensitive data, so the process adopts two strategies to deal with each case. If the data identified is public or non-sensitive, the flow is simplified, so the data scientist obtains the data directly from a public agency. In accordance with good practice, the data scientist processes and anonymizes this data. He/she then make it available in open or tabular format, ensuring reuse and integration with analytical tools. In the case of data protected by LGPD or considered to be sensitive, additional measures are taken, such as the verification of confidentiality agreements and privacy terms, in order to have the controlled release of data only after compliance with legal requirements. The data scientist carries out secure processing and rigorous anonymization in accordance with data protection standards.

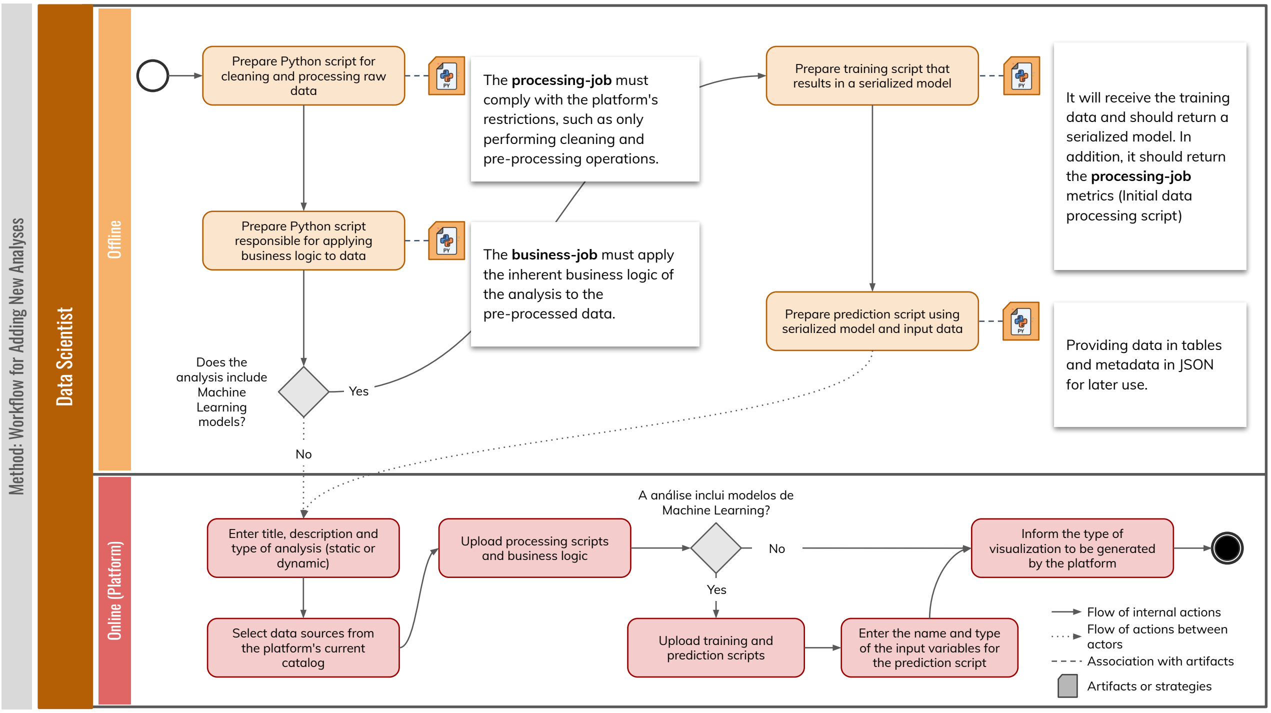
Method: Workflow for Adding New Analyses
Data Scientist
Offline
Online (Platform)
Prepare Python script for cleaning and processing raw data
The **processing-job** must comply with the platform's restrictions, such as only performing cleaning and pre-processing operations.
Prepare Python script responsible for applying business logic to data
The **business-job** must apply the inherent business logic of the analysis to the pre-processed data.
Does the analysis include Machine Learning models?
Yes
No
Prepare training script that results in a serialized model
It will receive the training data and should return a serialized model. In addition, it should return the **processing-job** metrics (Initial data processing script)
Prepare prediction script using serialized model and input data
Providing data in tables and metadata in JSON for later use.
Enter title, description and type of analysis (static or dynamic)
Select data sources from the platform's current catalog
Upload processing scripts and business logic
A análise inclui modelos de Machine Learning?
No
Yes
Upload training and prediction scripts
Enter the name and type of the input variables for the prediction script
Inform the type of visualization to be generated by the platform
Flow of internal actions
Flow of actions between actors
Association with artifacts
Artifacts or strategies

Figure 2: Method: Workflow for Adding New Analyses for Big Data Systems in the Public Sector.

Once the data is available, processed and analyzed, the Data Scientist participates in workshops and meetings with the public manager to ensure alignment and validation of the findings. Once the results have been validated, the analytical products are registered on the institutional platform. The Administrator defines the access profiles to the resulting products, guaranteeing governance, information security and transparency.

With the analyses validated and the products registered, the Public Manager accesses the results, based on the access profiles defined. The Public Manager makes decisions based on the evidence generated, enabling more effective public policies based on concrete data.

This process promotes integration between technical and institutional areas, with a strong emphasis on governance, transparency and accountability in the use of data. By structuring all the stages - from identifying the problem to delivering the analysis - the process ensures that the data products are relevant, secure and applicable, strengthening institutional intelligence and evidence-based public management.

6 Method: Workflow for Adding New Analyses for Big Data Systems in the Public Sector

Once the analytical product has been requested on the platform, the data scientist follows a standard technical method that guarantees the quality, security and reusability of the scripts and models. This process is divided into two stages: offline and online, as shown in the Figure 2.

First, the data scientist develops the processing script (processing-job). This script is developed in Python and aims to clean and pre-process the raw data. This script must comply with the platform's restrictions, but no business logic or modeling is applied at this stage.

The data scientist then develops the business-job script, which is developed in Python and aims to apply the analytical logic associated with the institutional challenge. The data scientist uses the pre-processed data to produce the expected results. If Machine Learning (ML) is used, the data scientist develops the Training script which receives the training data and returns a serialized model and the metrics used. If prediction is required, the data scientist develops the Prediction script that applies the model to the input data and generates predictions.

Once the scripts are ready, the data scientist registers the metadata (title, description, type of analysis - static or dynamic). He/she then uploads the processing-job and business-job scripts. If applicable, he/she uploads the ML scripts (training and prediction). Next, the data scientist indicates the data sources used from the institutional catalog, defines the type of visualization to be generated automatically by the platform (e.g. line graph, bar graph, etc.) and maps the input and output variables, making it easier to reuse the model and automate new analyses.

This method guarantees standardization, traceability and alignment with good analytical development practices in institutional environments.

7 Results and Discussion

From this reference process, the Big Data Fortaleza platform was developed, focused on analyzing data on evidence-based public policies. The solution enables continuous monitoring of key indicators, aligned with national and international standards, offering analysis for data- and evidence-driven decision-making.

With its launch, the platform enabled the integrated analysis of data from sectors such as early childhood, education, healthcare, social assistance, and distribution of medication. More than 27 dashboards and three automatic notification alerts were made available, providing public officials with valuable insights for formulating new policies. A concrete example was the creation of vaccination programs in daycare centers. One month after the platform's launching, more than 2,000 children had their vaccination schedules updated.

Figure 3 and Figure 4 illustrate the analytical potential of the Big Data platform implemented to support public policies focused on early childhood in the city of Fortaleza. The maps were used to cross-reference territorial data and support evidence-based decision-making.

Figure 3 presents a georeferenced map of municipal and partner daycare centers throughout the city. Each point corresponds to a service unit, and the colors indicate the vaccination dropout rate among the enrolled children, classified as green (low dropout rate), yellow (medium), and red (high). This visualization enabled us to identify territorial concentrations with greater vaccination fragility, allowing us to direct specific corrective actions, such as local vaccination campaigns and intersectoral interventions.

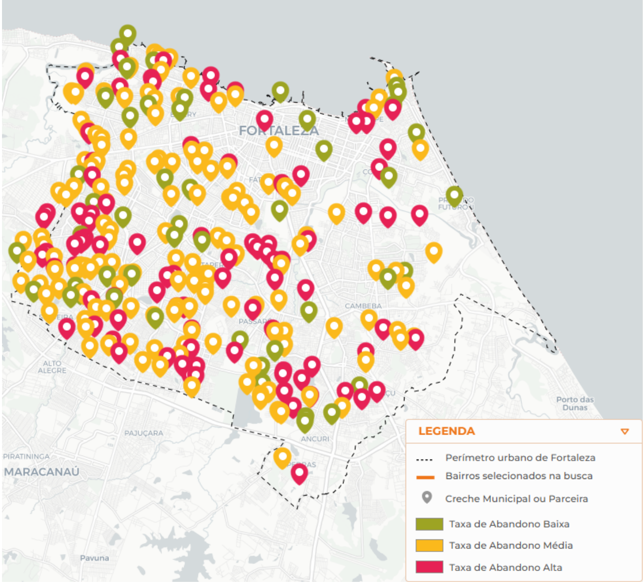

Figure 3: Vaccines per Daycare in the City of Fortaleza

Figure 4 is a choropleth map that displays the number of pregnant women by neighborhood, with color gradations indicating the density of pregnant women in different regions. By overlaying this data with vaccination dropout rates and daycare center locations, it was possible to anticipate high-risk areas and plan more effective prevention strategies, such as strengthening prenatal communication about the vaccination schedule and prioritizing neighborhoods with greater vulnerability for health and social assistance initiatives.

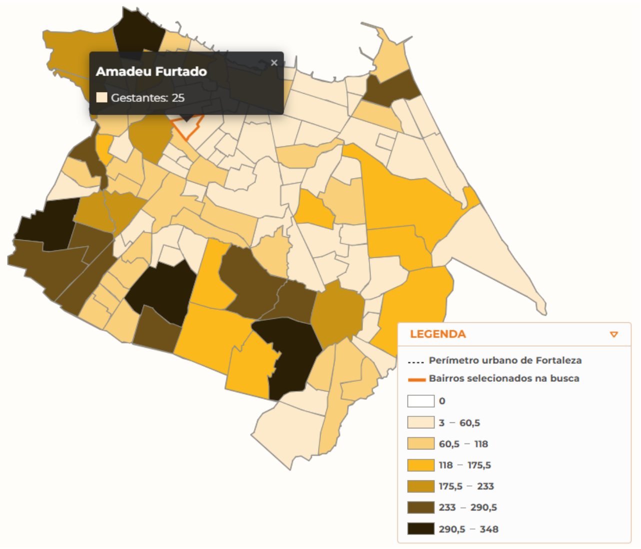

Figure 4: Pregnant Women by Neighborhood in the City of Fortaleza

The integrated use of this information exemplifies how the platform supports the development of more effective and targeted public policies, directly contributing to reducing inequalities in access to health and education in early childhood.

The strategic results presented were achieved through the process and method outlined in this article. The Process that contains the workflow and the actors for big data systems in the pblic sector strengthens co-accountability and legitimacy of analytical products, in addition to ensuring compliance with the LGPD in cases of sensitive data, through anonymization and usage agreements. The method, which is composed of a workflow for adding new analyses, standardizes the structure and ensures that the analytical products generated are aligned with strategic objectives and are securely integrated into the platform. The processed data is made available in structured formats (tables and metadata in JSON), enabling the generation of automated analytics and alerts.

In the specific case of analyzing vaccination abandonment and pregnant women, the process and the method presented in this paper enabled the following:

- The integration of multiple data sources (education, health, social assistance);
- The creation of visual products with geographic and semantic breakdowns (such as rates and absolute volume).

In short, the combination of a collaborative, intersectoral process and a well-defined technical method was essential for transforming raw data into actionable intelligence, strengthening institutional capacity to respond to complex social challenges and making public policies evidence-based.

8 Final Remarks

The experience report demonstrates that adopting a standardized technical process for developing analyses, combined with a structured collaborative process among different institutional stakeholders, is essential for transforming raw data into intelligence applied to public management. The clear methodology, with well-defined steps for preparation, processing, modeling, and dissemination of analyses, ensured the technical consistency of the products developed, while the intersectoral process included analyses of the strategic priorities of public agencies, respecting legal and ethical aspects, such as the LGPD.

This operational arrangement enabled the production of analytical visualizations, such as maps of vaccination abandonment and concentration of pregnant women, which served as the basis for agile, evidence-driven decisions. The collaborative work of managers, platform administrators, data scientists, and data-driven departments was crucial in generating a real impact on the population.

We can concluded that, more than a technology per se, it is the process and the alignment between strategy, technology, and governance that enables the effective use of Big Data for public policy. This model can be replicated in other contexts, especially for public agencies that work with massive data and require evidence-based policies.

ACKNOWLEDGMENTS

We would like to thank FUNCAP for the project financial support and CNPq for the productivity grant awarded to Rossana M. C. Andrade (306362/2021-0).

REFERENCES

[1] Alaa Alsaig, Vangalur Alagar, Zaki Chammaa, and Nematollaah Shiri. 2019. Characterization and efficient management of big data in iot-driven smart city development. *Sensors* 19, 11 (2019), 2430.

[2] Carol Andrade, Fernanda Campagnucci, José Borbolla Neto, José Macedo, Marianna Gonçalves, Mariana Zonari, Silvana Paula Martins de Melo, and Ticiana Linhares. 2021. *A Era dos Dados para o Setor Público: Uma Nova Cultura Organizacional Analítica.* ÍRIS; AWS Institute; Social Good Brasil; Open Knowledge Brasil; Branded Brain; Programa Cientista Chefe do Governo do Ceará, Brasil. Tema: Alfabetização em Dados.

[3] Michela Arnaboldi and Giovanni Azzone. 2020. Data science in the design of public policies: dispelling the obscurity in matching policy demand and data offer. *Heliyon* 6, 6 (2020).

[4] Otmane Azeroual and Radka Nacheva. 2023. Data Mesh for Managing Complex Big Data Landscapes and Enhancing Decision Making in Organizations.. In *KMIS*. 202–212.

[5] Élcio Batista, Rossana MC Andrade, Ismayle S Santos, Tales P Nogueira, Pedro AM Oliveira, Valeria Lelli, and Victória T Oliveira. 2024. Fortaleza city hall strategic planning based on data analysis and forecasting. In *Congresso Ibero-Americano em Engenharia de Software (CIbSE)*. SBC, 433–436.

[6] Manuel Pedro Rodríguez Bolívar and Albert J Meijer. 2016. Smart governance: Using a literature review and empirical analysis to build a research model. *Social Science Computer Review* 34, 6 (2016), 673–692.

[7] Bin Cheng, Salvatore Longo, Flavio Cirillo, Martin Bauer, and Ernoe Kovacs. 2015. Building a big data platform for smart cities: Experience and lessons from santander. In *2015 IEEE International Congress on Big Data*. IEEE, 592–599.

[8] Ali Davoudian and Mengchi Liu. 2020. Big Data Systems: A Software Engineering Perspective. 53, 5, Article 110 (sep 2020), 39 pages. doi:10.1145/3408314

[9] Nihit Goyal, Ola G El-Taliawi, and Michael Howlett. 2022. The prevalence of big data analytics in public policy: is there a research-pedagogy gap? In *Emerging Pedagogies for Policy Education: Insights from Asia*. Springer, 99–123.

[10] Jens Kandt and Michael Batty. 2021. Smart cities, big data and urban policy: Towards urban analytics for the long run. *Cities* 109 (2021), 102992.

[11] Bram Klievink, Bart-Jan Romijn, Scott Cunningham, and Hans de Bruijn. 2017. Big data in the public sector: Uncertainties and readiness. *Information systems frontiers* 19, 2 (2017), 267–283.

[12] Saurabh Mishra, Mahendra Shinde, Aniket Yadav, Bilal Ayyub, and Anand Rao. 2024. An AI-Driven Data Mesh Architecture Enhancing Decision-Making in Infrastructure Construction and Public Procurement. *arXiv preprint arXiv:2412.00224* (2024).

[13] Martijn Poel, Eric T Meyer, and Ralph Schroeder. 2018. Big data for policymaking: Great expectations, but with limited progress? *Policy & Internet* 10, 3 (2018), 347–367.

[14] Fajar Rahmanto, Ulung Pribadi, and Agus Priyanto. 2021. Big data: What are the implications for public sector Policy in society 5.0 era?. In *IOP Conference Series: Earth and Environmental Science*, Vol. 717. IOP Publishing, 012009.

[15] Syed Iftikhar Hussain Shah, Vassilios Peristeras, and Ioannis Magnisalis. 2021. Government big data ecosystem: definitions, types of data, actors, and roles and the impact in public administrations. *ACM Journal of Data and Information Quality* 13, 2 (2021), 1–25.

[16] Bhagya Nathali Silva, Murad Khan, Changsu Jung, Jihun Seo, Diyan Muhammad, Jihun Han, Yongtak Yoon, and Kijun Han. 2018. Urban planning and smart city decision management empowered by real-time data processing using big data analytics. *Sensors* 18, 9 (2018), 2994.

[17] Tom Van Eijk, Indika Kumara, Dario Di Nucci, Damian Andrew Tamburri, and Willem-Jan Van den Heuvel. 2024. Architectural design decisions for self-serve data platforms in data meshes. In *2024 IEEE 21st International Conference on Software Architecture (ICSA)*. IEEE, 135–145.

[18] RD Wahyunengseh and S Hastjarjo. 2021. Big Data Analysis of Policies on Disaster Communication: Mapping the issues of communication and public responses in the government social media. In *IOP conference series: Earth and environmental science*, Vol. 717. IOP Publishing, 012004.