From Acquisition to Interpretation: A Model for Creating Data Storytelling in Big Data

Victória T. Oliveira[1], **Rossana Maria de Castro Andrade**[1], **Miguel Franklin de Castro**[1], **Ismayle S. Santos**[1,2]

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

[2]State University of Ceará (UECE), Fortaleza, CE, Brazil

victoriat.oliveira@alu.ufc.br, rossana@ufc.br

miguel@ufc.br, ismayle.santos@uece.br

Abstract. *The processes involved in the planning, development, and optimization phases of a Big Data project pose challenges within the realm of Requirements Engineering (RE), emphasizing the complexities of requirements management in multidisciplinary teams. Streamlining activities related to requirements processes, as well as those involving other teams like the data science team, is crucial to the success of such projects. This article introduces a template designed to assist in creating data storytelling. The template encompasses everything from data collection sources to the types of visualizations to be used, and how to interpret these visualizations.*

Resumo. *Os procedimentos envolvidos nas etapas de planejamento, desenvolvimento e otimização de um projeto de Big Data apresentam desafios no escopo da Engenharia de Requisitos (RE), destacando as dificuldades inerentes ao gerenciamento de requisitos em equipes multidisciplinares. Otimizar as atividades relacionadas aos processos de requisitos e os processos que envolvem outras equipes, como a equipe de ciência de dados, contribui para o sucesso desses projetos. Este artigo apresenta um modelo para ajudar a criar data storytelling. O modelo abrange tudo, desde onde os dados são coletados, que tipo de visualização esses dados terão e como interpretar esse tipo de visualização.*

1. Introduction

The increase in the volume and velocity of data, coupled with the growing demand for automation, has made Big Data technology an efficient solution to several current challenges. This technological transformation has contributed to the popularization of Big Data systems. The development of software systems that incorporate Big Data components is on the rise and is being explored in several sectors. 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 [Davoudian and Liu 2020]. These systems have brought several challenges to software development, such as technical challenges related to the 5Vs (Volume, velocity, variety, veracity, and value).

Big Data applications, like traditional applications, meet the end user's needs, except that underlying the software system is the Big Data on which the system operates. Compared to traditional software development, where development processes are generally well established, the processes for developing applications involving Big Data are still unclear in the scientific literature – given the nature of the computation involved and data characteristics such as volume, variety, veracity and velocity. Recent studies have shown that traditional RE methodologies are commonly user-centric rather than data-centric. Traditional RE activities focus on requirements that are visible to users, rather than requirements obtained by data scientists after analyzing Big Data [Costa et al. 2024].

The process of creating software focused on Big Data differs from traditional development approaches, which makes the application of conventional techniques and tools more challenging. Requirements Engineering (RE) plays an essential role in the software engineering process, being considered one of the most critical phases of the software development life cycle. As we might expect, Requirements Engineering would play a similar role in the context of Big Data applications. Traditional requirements engineering involves requirements gathering, analysis, and detailed design for implementing an executable program, focusing primarily on writing code [Van Vliet et al. 2008]. On the other hand, software development with big data components requires consideration of additional configuration aspects that are not addressed by traditional requirements engineering. Examples include defining the architecture, collecting and processing data, choosing appropriate algorithms, training models, and presenting the results.

The software development process for Big Data needs to be detailed in the requirements phase, i.e., it is necessary to specify where the data will be collected from, how it will be processed, which metrics and indicators will be used, and how the data will be visualized. However, traditional requirements engineering does not provide specific guidelines for Big Data systems. As a result, these systems often lack well-defined requirements and appropriate requirements engineering (RE) techniques, which creates the need to adapt existing RE methods. Due to the difference in the development process, new challenges are emerging when managing requirements for Big Data systems.

In the case of the public sector, when a city manager decides to invest in the use of Big Data, there are a number of challenges to be faced. For example, there are several data sources with different formats, low data quality, and some studies that help with the requirements time to specify documents that contain data. These challenges include defining, eliciting, and specifying requirements involving data and data visualizations. This article presents a template to help create data storytelling. The template covers everything from where the data is collected, what type of visualization that data will have, and how to interpret that type of visualization.

2. Related Work

At the moment, researchers have studied the change in the methods of writing requirements engineering and data visualization using data storytelling. [Bosch et al. 2018] explained that companies are moving towards data-driven approaches as decisions become more dependent on data to determine system functionalities. This change has resulted in the demand to modify current RE practices to become more adaptable to data-driven approaches. In addition, data-driven RE is changing the way requirements are elicited and

obtained. The solution proposed by the authors can be summarized in some main approaches: Integration of Requirements with Data and Results, Adoption of an Iterative and Agile Process, Collaborative and Multidisciplinary Requirements, and Use of AI-Specific Requirements Models, among others. [Yasin et al. 2018] proposes an integrated approach to requirements analysis in Big Data, with the aim of ensuring that services meet user needs and are built efficiently and in compliance with regulatory requirements.

[Arruda and Laigner 2020] presents an analysis of Requirements Engineering practices and challenges in the context of software development for Big Data. The research is based on an initial case study, exploring how organizations deal with the demands of requirements in software projects involving large volumes of data. The authors discuss the difficulties development teams encounter when trying to map and meet requirements in Big Data projects, including complexity, rapid evolution of technologies and the need for greater collaboration between different company areas. The main challenges identified include the precise definition of requirements in a Big Data environment, the integration of multiple data sources, scalability and the flexibility needed to deal with the dynamic and unpredictable nature of these projects. They point out some strategies that can help solve or mitigate these challenges: Interdisciplinary Collaboration, Iteration and Flexibility, Specific Tools and Techniques, and Agile and Visual Documentation, among others.

[Nalchigar et al. 2021] presented a conceptual framework that offered three modeling views. The three views are the business view, analytical design view, and data preparation view. In the Business View, business-related objectives are elicited and modeled. The Analytical Design View focuses on the selection of technical resources, such as machine learning algorithms, and which qualities and tradeoffs should be considered when choosing an algorithm. Finally, the Data Preparation View helps in the selection and understanding of available data sets.

[Costa et al. 2024] explores the importance of requirements specification in the context of Big Data projects within the public sector. The authors discuss the particularities and challenges involved in collecting, organizing, and analyzing large volumes of data in government services. The main focus is on how an effective requirements specification can ensure the success of these projects, from defining user needs to issues related to data privacy and security. In addition, the article addresses the techniques, tools, and methodologies that can be used to understand demands and translate these needs into systems that are effective in using Big Data for decision-making in the public sector. All previous works proposed solutions related to requirements engineering but did not provide templates to assist in the creation of data storytelling.

In the field of data storytelling, data storytelling is often used for professional purposes, which makes its visual effectiveness a top priority. For this reason, data storytelling visualizations and visual styles in these fields are usually designed in a simplified way. Most research on data storytelling focuses on its application in the professional environment. While data visualization experts are starting to focus their attention on broader audiences, it is becoming increasingly crucial to explore the frameworks and techniques that enable data storytelling in informal contexts, creating engaging content from complex data sets [Lee et al. 2020]. The potential of visual data storytelling is still far from being fully explored. As a first step, our main goal is to provide a template for improving

data storytelling.

3. Requirements Elicitation and Big Data

In Big Data projects, traditional requirements engineering practices may not be sufficient due to the complexity and dynamic nature of these projects. Defining specific RE processes in the context of Big Data implies adjusting and customizing requirements engineering processes to deal with the peculiarities of these projects, such as the integration of large volumes of data and the need for real-time processing [Arruda and Madhavji 2017]. According to [Anderson 2015], requirements engineering must evolve to incorporate best practices in the development of systems that operate with large amounts of data, considering scalability, security, and efficiency.

Requirements Engineering is undoubtedly the most crucial phase in the software engineering life cycle and plays a significant role in each stage of software development [Hull et al. 2005]. In RE, understanding stakeholder requests is important, and requirements act as the communication channel between system developers and stakeholders [Wheatcraft and Ryan 2018]. RE acts as this conduit for gathering and documenting stakeholder needs [Nuseibeh and Easterbrook 2000]. Therefore, it is vital to establish requirements early on when building software systems to ensure that all stakeholder needs and specifications are captured and documented correctly. [Anderson 2015], discusses how software engineering is being transformed by the growing importance of large-scale data, known as Big Data. The author highlights that traditional approaches to software engineering need to be compensated to deal with the demands of Big Data, which include not only the volume of data, but also its diversity and velocity.

When designing big data software systems, new processes emerge, such as data management, model training, and visualization design; at Microsoft, the Requirements Engineering process is employed in a nine-step model. It includes requirements gathering, data collection, cleaning and labeling, feature engineering, model training and evaluation, and finally, model deployment and monitoring over time. Requirements are decided based on the feasibility of implementing features and finding appropriate models for specific problems [Amershi et al. 2019].

3.1. Data Storytelling

Data is a rich source of information, containing vast knowledge and insights in its “raw” and simplified form. However, for this data to be understood by humans, an analytical process is required that involves simplification and the selection of appropriate visualization methods. Raw data is often difficult for a lay audience to understand, but attractive images, graphs, gifs, and videos – common in everyday life and popular culture – are easy to understand, although they do not always carry much meaning, as they are often created for aesthetic or entertainment purposes. In this context, it seems possible to combine entertainment and data visualization, transforming “serious” information into entertaining content for the general public. This could broaden the reach of data visualization and make it more engaging for everyone [Zhang et al. 2022] [Ren et al. 2023].

As more and more data is collected and processed every day, the ways in which data-based information can be shared and communicated become increasingly meaningful. Exploring modern forms of expression provides new modalities for giving meaning

to data. Storytelling is a way of presenting data and, at the same time, organizing information in an accessible and easy-to-understand way. When moving from the data analysis phase to the presentation phase, storytelling becomes an effective method for transforming information extracted from data into a user-friendly format for non-specialist audiences [Matei and Hunter 2021].

Data storytelling in the public sector is a strategic approach to transforming complex data into understandable and impactful narratives, with the goal of clearly and effectively communicating information about public policies, program and service outcomes, or any government initiative. Public policies can be adjusted or improved based on insights derived from data. Data analysis and visualization help public managers make informed decisions about where to invest, how to prioritize projects, or how to optimize resource allocation. Creating clear visualizations, such as graphs or dashboards, makes information easier to understand. Instead of analyzing lengthy reports or spreadsheets, data can be presented in an interactive way, highlighting trends, patterns, and meaningful results.

We propose a three-part framework for organizing information derived from data into a model that supports data storytelling. The three parts of this framework are status, background information, and technical information. The following section presents the model we developed.

4. Context

At the City Hall of Fortaleza, one of the municipal agencies is the Fortaleza Planning Institute (IPLANFOR), which is dedicated to generating knowledge, monitoring and evaluating public policies, coordinating strategic planning, and encouraging innovative initiatives.

To guide the city's planning, the City Hall developed a tool with strategies for implementation in the short, medium, and long term, with a vision extending to 2040. As a result, Fortaleza 2040 was created: a participatory strategic plan designed to integrate physical and territorial development with social and economic growth. It fosters discussions about the city from multiple perspectives, sectors, territories, and levels of government. Fortaleza 2040 addresses pressing issues that need to be tackled, including child vaccination rates, infant mortality, and prenatal monitoring.

Among the challenges identified, early childhood development, focusing on children from zero to six years old, stands out. This area is cross-cutting and requires coordinated actions across health, education, social assistance, and other sectors. Given the challenges faced by the city of Fortaleza and the role of IPLANFOR, collecting and analyzing data is crucial for diagnosis and supporting decision-making by municipal authorities. Through intelligent data analysis, it becomes possible to enhance the quality of services offered to citizens, improving their quality of life. Therefore, having a robust infrastructure to collect, store, and process data from multiple sources is vital.

The development of the Big Data Fortaleza platform allows for analysis using data from various sectors related to Early Childhood policies. This is essential for ongoing monitoring and evaluation of key indicators, ideally following internationally recognized standards, and providing more accurate insights to guide data-driven and evidence-based decision-making.

The launch of the Big Data Fortaleza platform enabled comprehensive analysis using data from various early childhood-related sectors. It provided municipal authorities with over 27 dashboards reports and 3 notification alerts, covering areas such as the Early Childhood Secretariat, Education, Health, Social Assistance, and Drugs. As a result, municipal managers gained valuable insights that helped shape new public policies, such as the implementation of vaccination programs within daycare centers. Within a month of the platform's launch, more than 2,000 children had their vaccination schedules updated and brought up to date.

In the education sector, it was crucial for public authorities to ensure that the population had access to nurseries, daycare centers, and specialized early childhood education schools. In healthcare, providing comprehensive care—from prenatal to postpartum care for women, along with initial newborn monitoring and vaccinations—was essential. Regarding human rights and social development, it was necessary to identify and support families in situations of socioeconomic vulnerability or homelessness, offering social benefits to reduce disparities and ensuring access to services and resources that protect the rights of children and pregnant women.

5. Proposed Approach and Result

Requirements cannot be viewed in isolation, but must be tightly connected to the data and expected outcomes. When defining the requirements for a Big Data system, it is essential to consider how the data will be prepared, prepared, and used to train the models, as well as how this data will influence the outcomes that a Big Data system will deliver. Data requirements should be an integral part of the requirements definition process, to avoid the AI model being based on inadequate or envious data. In Big Data projects, specific diversity (with different interests and experiences) can hinder this common understanding. The proposed model helps to create a common vision, providing a starting point for alignment among all stakeholders, ensuring that everyone involved in the project has a clear and consistent understanding of the system requirements and their priorities.

A challenge for the requirements team was documenting data requirements. The system had features that processed City Hall data and presented it in a structured way using predefined visualizations in the form of a dashboard. The data and type of visualization are variable, but previously defined by stakeholders. These requests were treated as data requirements, as they required the participation of data scientists to analyze and process the data before incorporating the Platform.

However, this type of requirement was not easily met by the use cases, as they were generic scenarios that varied the use of the system's functionalities. The need to record these requests led to the creation of a new type of artifact. These described the data that would be displayed, the type of visualization, the source of the data, among other technical specifications.

The proposed model is divided into 3 stages. This model was created so that the requirements analyst can fill it out together with the data scientist after having collected all the necessary information from the project stakeholders. To develop this model, refinement meetings were held with Requirements and Data Science researchers. In these meetings, questions were asked about which attributes they defined as important for eliciting a dashboard.

In the first stage, the analysts answer questions related to the status of the request. The structure can be seen in the Table. In black and bold, this is the information that requires answers. In gray, we can see an explanation of what should be filled out.

Status	
Sprint	SCRUM methodology sprint to be implemented
Responsible	Responsible Data Scientist
Responsible	Requirements Analyst
Development Status	Not Started, In Progress, Under Review, Completed

Tabela 1. Request Status

In step 2, the required information is basic information related to the dashboard construction request, for example, which department and employee made the request, what the priority of this request is, among other things. This information is filled in by the requirements analyst.

Basic information	
Requesting Secretary	Secretary who requested the dashboard
Requesting Employee	Employee who requested the dashboard
Request Type	New Request, Removal or Edit
Priority Level	High, Medium, Low
Request Date	Date requested
Request Description	In detail, what is expected
Data storytelling	How to interpret this dashboard
Associated requirements	Functional, Non-functional requirements or associated Use Cases
Desired Access Restriction	Secretary/Agencies with viewing permission

Tabela 2. Basic information

In the third stage, we have technical information, this information is filled in by the data scientist. In this stage, the necessary information is more focused on the data and its architecture.

This template was used and helped to create 27 dashboards for the Big Data Platform Fortaleza. In Figure 1, one of these 27 dashboards can be seen, the HDI (Human Development Index) by neighborhood. The HDI by neighborhood was requested by IPLANFOR in order to monitor human development in the neighborhoods of Fortaleza. The HDI is calculated through the geometric mean of three indexes: health, education and income. Each index is calculated separately and then the geometric mean is calculated.

6. Final Considerations and Future Work

The development of Big Data systems, as demonstrated by the Big Data platform Fortaleza, is an essential tool for transforming raw data into valuable information that can guide more effective and evidence-driven public policies.

In addition, the process of adapting Requirements Engineering (RE) to Big Data systems is crucial to ensure that the solutions developed meet the specific needs of users

Technical information	
Metrics	A measure or set of measures used to generate the dashboard
Indicators	Measures calculated from metrics
Visualization type	Line chart, Bar chart, etc.
Required data sources	Which data sources are required to build the dashboard
ProcessingJob	Processing ETL
BusinessJob	Business-level ETL
Business-level data structure (schema)	Describe the logical structure of the database, i.e., the organization of tables, relationships, indexes, etc.
Pipilene	Responsible Automation
Observations (optional)	Observations on building the dashboard

Tabela 3. Technical information

and align with the strategic objectives of public administrations. The proposal of a data storytelling model, as presented in this article, can be an important aid in this process, as it allows stakeholders to visualize and interpret data in a clear, accessible and meaningful way. The implementation of robust Big Data systems, together with an adequate Requirements Engineering process, is essential for the creation of more effective public policies and for the development of more transparent and efficient management that prioritizes the well-being of the population.

For future work, the authors intend to use the template in two more government projects, namely in the State Treasury Department of Ceará and in the State Planning and Management Department of Ceará.

7. Acknowledgements

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).

8. References

Referências

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., and Zimmermann, T. (2019). Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300. IEEE.

Anderson, K. M. (2015). Embrace the challenges: Software engineering in a big data world. In *2015 ieee/acm 1st international workshop on big data software engineering*, pages 19–25. IEEE.

Arruda, D. and Laigner, R. (2020). Requirements engineering practices and challenges in the context of big data software development projects: Early insights from a case study. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 2012–2019. IEEE.

Arruda, D. and Madhavji, N. H. (2017). Towards a requirements engineering artefact model in the context of big data software development projects: Research in progress. In *2017 IEEE international conference on big data (big data)*, pages 2314–2319. IEEE.

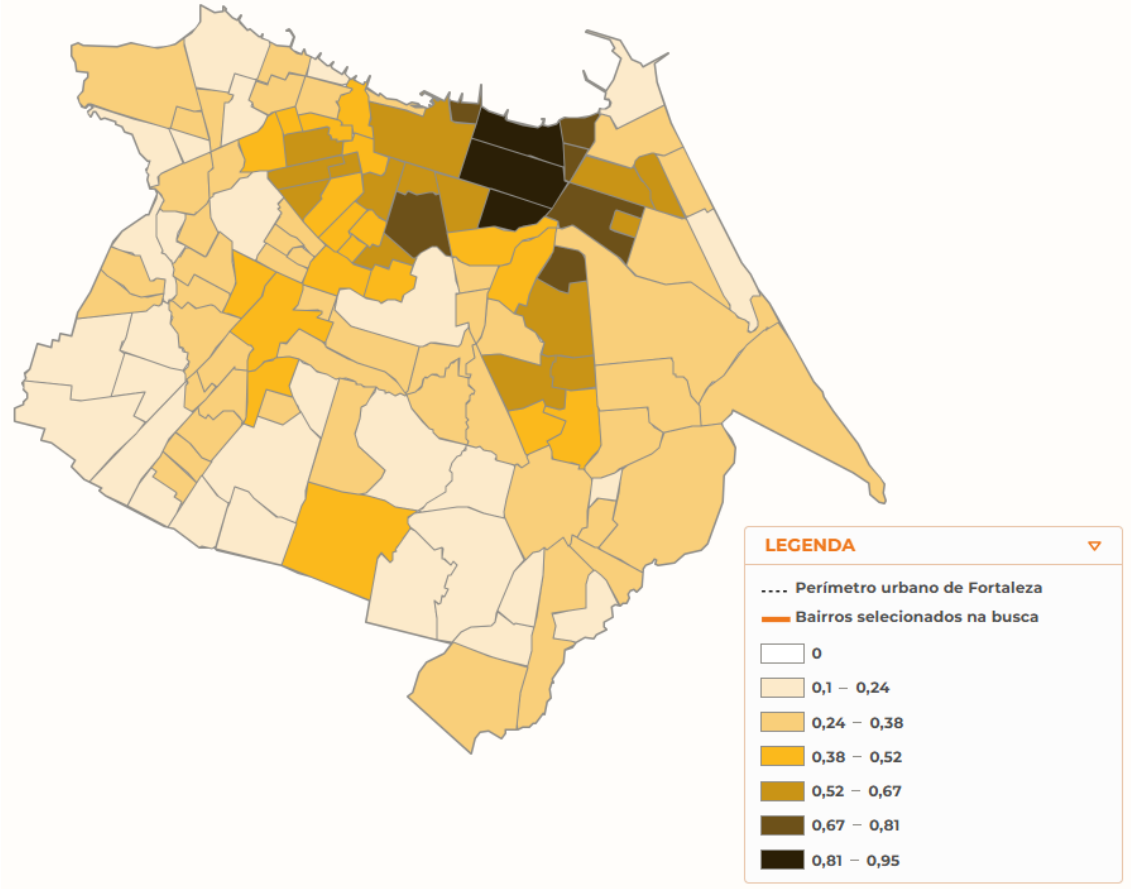

Figura 1. Dashboard - HDI (Human Development Index).

Bosch, J., Olsson, H. H., and Crnkovic, I. (2018). It takes three to tango: Requirement, outcome/data, and ai driven development. In *SiBW 2018, Software-intensive Business: Start-ups, Ecosystems and Platforms, Espoo, Finland, December 3, 2018*, pages 177–192. CEUR-WS. org.

Costa, A. F., Freitas, L. I., Cavalcante, D. A., Oliveira, V. T., Lelli, V., Santos, I. S., Oliveira, P. A., Nogueira, T. P., and Andrade, R. M. (2024). Especificaçao de requisitos em um projeto de big data no setor publico. In *Congresso Ibero-Americano em Engenharia de Software (CIbSE)*, pages 417–420. SBC.

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

Hull, E., Jackson, K., and Dick, J. (2005). *Requirements engineering in the solution domain*. Springer.

Lee, B., Choe, E. K., Isenberg, P., Marriott, K., and Stasko, J. (2020). Reaching broader audiences with data visualization. *IEEE computer graphics and applications*, 40(2):82–90.

Matei, S. A. and Hunter, L. (2021). Data storytelling is not storytelling with data: A framework for storytelling in science communication and data journalism. *The Information Society*, 37(5):312–322.

Nalchigar, S., Yu, E., and Keshavjee, K. (2021). Modeling machine learning requirements from three perspectives: a case report from the healthcare domain. *Requirements Engineering*, 26:237–254.

Nuseibeh, B. and Easterbrook, S. (2000). Requirements engineering: a roadmap. In *Proceedings of the Conference on the Future of Software Engineering*, pages 35–46.

Ren, P., Wang, Y., and Zhao, F. (2023). Re-understanding of data storytelling tools from a narrative perspective. *Visual Intelligence*, 1(1):11.

Van Vliet, H., Van Vliet, H., and Van Vliet, J. (2008). *Software engineering: principles and practice*, volume 13. John Wiley & Sons Hoboken, NJ.

Wheatcraft, L. S. and Ryan, M. J. (2018). Communicating requirements–effectively! In *INCOSE International Symposium*, volume 28, pages 716–732. Wiley Online Library.

Yasin, A., Liu, L., Cao, Z., Wang, J., Liu, Y., and Ling, T. S. (2018). Big data services requirements analysis. In *Requirements Engineering for Internet of Things: 4th Asia-Pacific Symposium, APRES 2017, Melaka, Malaysia, November 9–10, 2017, Proceedings 4*, pages 3–14. Springer.

Zhang, Y., Reynolds, M., Lugmayr, A., Damjanov, K., and Hassan, G. M. (2022). A visual data storytelling framework. *Informatics*, 9(4).