TECHNICAL UNIVERSITY OF CLUJ-NAPOCA

ACTA TECHNICA NAPOCENSIS

Series: Applied Mathematics, Mechanics, and Engineering
Vol. 66, Issue Special I, September, 2023

THE DATA JOURNEY FRAMEWORK: PROPOSAL OF FORMAL REPRESENTATION OF DATA AGENTS, ASSETS, AND STATES FROM DATA GENERATION TO DATA SERVING IN A DATA-CENTRIC OPERATING ENVIRONMENT

Alexandre SURKUS-CASTRO, Fernando DESCHAMPS, Edson PINHEIRO DE LIMA, Déborah SARRIA, Anis ASSAD NETO, Sergio GOUVEA DA COSTA

Abstract: *In the ever-expanding landscape of data-intensive operations, such as Data Analytics, Machine Learning, and Deep Learning, where massive computational resources and extensive data are essential, the need for reliable and consistent attributes for data and operating agents becomes paramount. While frameworks like Archimate® and BPMN enable functional and semantic representation of organizational architecture and processes at a higher abstraction level, the data context lacks a comprehensive framework capable of capturing its components' semantic and functional aspects, from minor to big data infrastructures. This paper proposes a novel ontological framework that addresses this gap, offering a semantic and practical representation approach for data-centric business contexts. The proposed model encompasses the entire data journey, starting from data generation, moving through various stages of maturity within the value chain, and culminating in its utilization by consumer business platforms. The framework aims to be vendor-agnostic and intelligible across diverse organizations and technological ecosystems, fostering interoperability and collaboration. By leveraging this ontological framework, organizations can enhance data operations, ensuring reliability, consistency, and compliance while facilitating effective communication and decision-making within and between entities.*

Key words: *Big data, framework, business process, enterprise.*

1. INTRODUCTION

In recent decades, the advent of intensive data processing under the umbrella of Big Data has enabled significant advancements in data science and data engineering [1]. These disciplines, in turn, have facilitated the improvement and systematization of processes in the industry, both in terms of performance management and command and control of manufacturing processes. Several authors have highlighted aspects related to performance analysis in industries and services that operate with more sensitive data, such as healthcare. Considering the organizational architecture composed of ontological, functional, and semantic elements, frameworks like ArchiMate [2] and Business Process Management have provided sufficient elements to categorize and organize processes, business components, and involved actors. On the other hand, data architectures have followed a different direction, lacking structuring frameworks capable of expressing the semantic and functional representation of data components and data-operating components. In the parallel analysis with the frameworks mentioned earlier, we do not have frameworks that can establish a direct relationship with data elements for processes and organizational architecture. This gap, represented by the space between well-resolved representative processes and architecture and the functional infrastructure supporting operations, provides an opportunity to propose a new structuring framework that appropriately combines architectural-processual and data-oriented aspects. The proposal stems from extending the BPMN2.0 [3] representation with elements capable of representing aspects of the operating data structure. Some advantages of this approach include leveraging an already

established culture with existing organizational frameworks, expanding the existing representation to become more comprehensive, and primarily addressing the archetypes prevalent in several available data ecosystems. The proposition developed in this research, in its initial version, was used to structure and represent data, from acquisition (ingestion) to processing and *servitization*, in an automotive manufacturing plant. The research will be presented in the following sections: (i) presentation of functional, semantic, and structural elements in data-centric contexts that require the representation of processes and organizational architecture; (ii) elements of the BPMN 2.0 specification with a proposed extension of its representation; (iii) the adequacy of the representation; (iv) the description of the production context that served as a case for validation and regulation of the proposed model; (v) diagnosis and future proposals.

2. THEORETICAL BACKGROUND

BPMN 2.0, as discussed in, serves as a foundational framework for representing and modelling business processes. It provides a comprehensive set of graphical notations and symbols that aid in visualizing processes, facilitating process analysis, optimization, and decision-making. For Synthesis and Integration, the works [3–5], the authors present a comprehensive review of process design methodologies. The paper highlights the importance of synthesizing different process design approaches to integrate various process elements effectively. Process design involves systematically identifying and analyzing unit operations, equipment, and control systems to create efficient and reliable processes. For Manufacturing Measurement, the work [6] focuses on the critical aspect of manufacturing measurement and quality control. Measurement Systems Analysis is essential for ensuring accurate and reliable measurement data as the basis for process control and improvement. It comprehensively reviews the various measurement techniques, tools, and statistical methods used in manufacturing environments. By understanding the intricacies of measurement systems analysis, practitioners can identify and mitigate measurement errors and improve manufacturing processes' overall quality and reliabilities.

3. DATA COMPONENTS IN DATA-CENTRIC CONTEXTS

The data components commonly employed in a data-centric infrastructure can be categorized into two fundamental classes [7, 8]. The Agent Class refers to the active member of the infrastructure, capable of executing programmatic or on-demand actions, while the Data-State Class represents the passive entity within the infrastructure. Both classes exhibit decompositions that are dependent on functional specialization or maturity representation. Figures 1 and 2 and 2 depict instances of data component classification, encompassing both data-agents and data-state commonly employed within the industry.

3.1 Data-state

The Data-State class represents the passive entity of the infrastructure but with classifications dependent on the maturity level regarding business references or aspects. The ontological decomposition (subclasses) of the Agent class is presented in Figures 1 and 6, also representing the semantic and functional structure of each agent subclass.

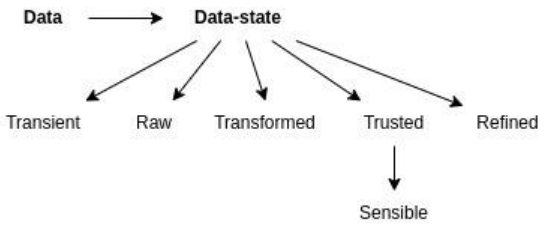

Figure 1: data-state components

The data-state components are passive entities, not capable of on-demand or programmatic actions. They represent the potential states of data in a data infrastructure. Transient: these are data in a transient state, either newly arrived or recently transferred or ingested from components external to the data solution domain. Their level of maturity is as rudimentary as possible, although it does not imply that they are not ready for consumption.

Data in the Raw state is considered ready and minimally organized, serving as consumable data for the initial data agents related to business rules. Transformed data results from minimal necessary transformations for the business process, typically involving size regularization, typology, and consistency. Trusted data represent more complex data structures composed of data originating from the preceding states. Aggregates may still contain open data. Depending on specific legislation or regulations (e.g., GDPR for the European Union), there is a subclassification of the Trusted state known as Sensible for these data states. The premise distinguishing data in the Trusted and Sensible states relates to anonymizing sensitive data, with the Sensible state referring to open and non-anonymized data.

3.2 Data-agent

The term "Data-Agent class" pertains to the active components of the data infrastructure, and they also possess an ontology dependent on their functional or structural relationship. Figures 2 and 3 illustrates the ontological structure of the data-agent class.

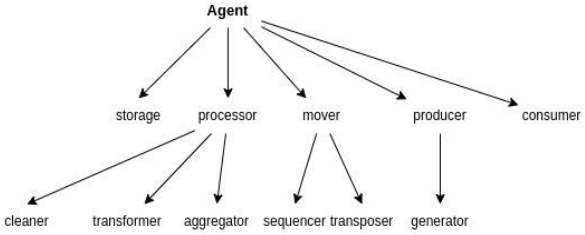

Figure 2: data-agent's components

Storage: This component retains data from persistence mechanisms, exhibiting an active nature by implementing elements of security, permission, quotas, and other restrictive aspects of the data infrastructure. Processor: This component is responsible for the processing of data within the infrastructure, with the ability to act as a cleaner (responsible for cleansing processes), data transformer, or aggregator (responsible for performing aggregations as defined by business rules). Mover: This component can perform data movements and can act as a sequencer (responsible for establishing the sequence of data across different stages) or as a transposer (a component that transposes data between the boundaries of data domains). Producer: This component can generate data constructs according to business rules. It can act as a generator, which is a producer that generates primary data such as random numbers and synthetic data.

4. ELEMENTS IN BPMN2.0 SPECIFICATION

BPMN 2.0, or Business Process Model and Notation 2.0 [3], offers a comprehensive set of standards and guidelines for designing and modelling various business processes, including manufacturing plant processes. It is versatility and graphical representation make it a valuable tool in the field of industrial engineering. By employing BPMN 2.0, engineers can visually depict the different stages, tasks, and interactions within a manufacturing plant process, enhancing the overall understanding and communication of the process flow. This standardized notation allows for the seamless integration of various elements such as activities, events, gateways, and decisions, enabling the representation of complex manufacturing processes in a clear and concise manner. Adapting BPMN components to illustrate the connection between Data Processing and Business Processes within inline manufacturing plants is crucial for comprehensive process modelling. Although BPMN needs more specific semantics for data, researchers have proposed approaches to address this gap [9]. In the proposed framework, the authors present extensions to BPMN that incorporates data objects and transformations, enhancing the representation of data-centric processes, and propose a methodology integrating BPMN with data modelling techniques to establish a coherent relationship between business processes and data processing activities. By leveraging these advancements, BPMN becomes a more powerful tool for capturing the intricacies of data-driven manufacturing processes.

5. FUNCTIONAL, SEMANTIC, AND ONTOLOGICAL EXTENSIONS PROPOSED

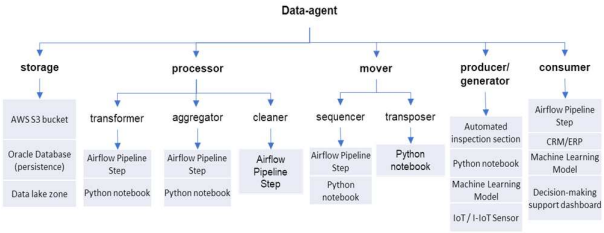

Figure 3: Examples of data agents

The BPMN's components related to Task, Start, End, and Sub-task lack explicit representation of data-centric aspects. The presented approach – henceforth called “Process Data Journey" or simply “Data Journey” – proposes an adaptation to BPMN 2 components, aiming to bridge the gap between business process representation and data-centric implementation. Adapting BPMN 2 components to include data-centric aspects brings several advantages to the representation and implementation of manufacturing processes. It allows for a more holistic view by explicitly capturing the interactions between process steps and data elements (figure 5). This facilitates a comprehensive understanding of how data flows through the manufacturing process and enables practitioners to identify bottlenecks, inefficiencies, and opportunities for process optimization.

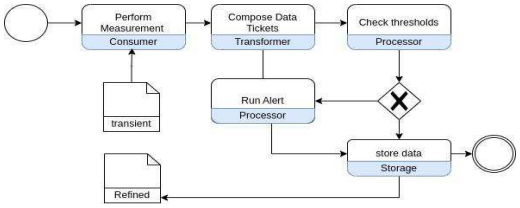

Figure 4: BPMN diagram with Data Journey Extensions

The Data Journey approach enables the integration of data modelling techniques with BPMN 2, fostering a closer alignment between business processes and data processing activities. This integration enhances data governance, data quality management, and overall data-driven decision-making within manufacturing plants.

By explicitly representing data transformations, data dependencies, and data flow paths, practitioners gain a deeper understanding of the impact of data on process execution, enabling them to make informed decisions regarding resource allocation, task assignment, and scheduling.

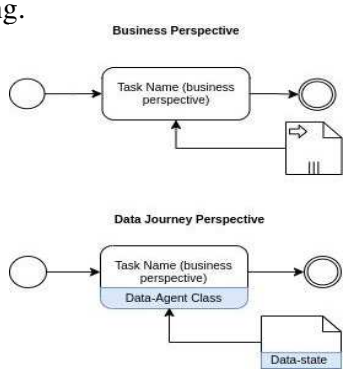

Figure 5: Business and Data Journey perspectives

This approach also may facilitate effective communication and collaboration between stakeholders by providing a shared visualization and interpretability of construct process-data, the interplay between business processes and data, empowering practitioners to identify opportunities for automation, intelligent data processing, and integration with emerging technologies like the Internet of Things (IoT) and Artificial Intelligence (AI). Figure 7 presents the representation of a segment of the process in which the approach was instantiated, starting from the measurement deck, passing through data transport, limit verification, alert conditional on failures, and ending with the storage of historical dataComponents now receive attributes that are concerned with governance items and component classes (Agent or State), added to their own graphical representations (figure 4 and 5). As for the agents, they are representative of the action type, not represented under the data domain in a vanilla BPMN. Data states can reference zoning in Data Lake or Data mesh structures.

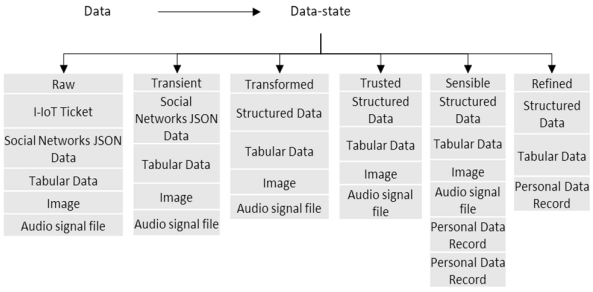

Figure 6. Examples of Data-states

6. STUDY CASE ON SHOP FLOOR

In Brazil's context of an inline automotive manufacturing plant, the Inline Automated

Inspection Section (AIS) serves as the origin of data. This section conducts automated inspections on car frames, capturing measurements and detecting any non-conformities or coordinate distance problems. The data from AIS is then transmitted to a data server, where it is organized and made available for further analysis (Figure 7). Machine learning models and applications utilize this data to validate the car frame's measurements against predefined limits, ensuring its conformity.

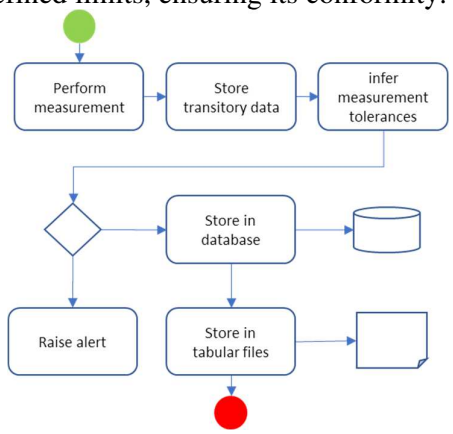

Figure 7. Measures Tolerance check

Any analysis results generated from this process are stored in a database within a data lake. Simultaneously, the data is also written to a Parquet file in the data-refined zone of the same data lake. To provide easy access and visualization of the historic car-frame measures, an automated dashboard is utilized by managers and operators.. The entire process, from data acquisition to storage and analysis, is designed and documented by technicians using BPM Notation, although it doesn't explicitly capture aspects such as data agents, their characteristics, or the data maturity stages. Parallel to the business process culture, various other approaches and techniques are employed to address data aspects and challenges. This data-centric approach, known as the Data Journey, can be applied in diverse contexts, ranging from small-scale solutions to large infrastructure setups. Within the manufacturing industry, it aids managers and operators in effectively identifying, classifying, and maintaining control of data components within the same business process framework. When dealing with the Industrial Internet of Things (I-IoT) and IoT sensors, as well as a wide range of control and automation assets, this framework assists in appropriately classifying them as suitable data components, considering the context, whether as agents or data states. Figure 8 provides a concise explanation of the data journey framework's application in the automotive manufacturing scenario. It outlines the entire process, which aligns with the factual inline operations, encompassing all the relevant assets and their classifications.

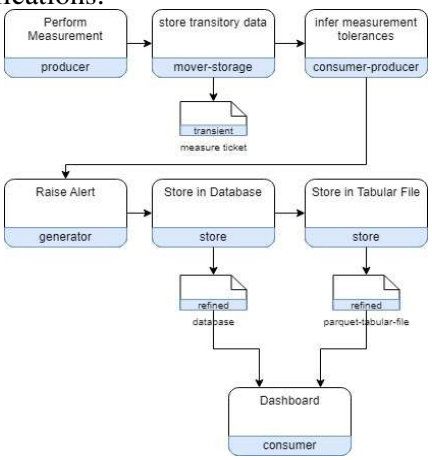

Figure 8. Data Journey for measure tolerance check

7. DISCUSSION

The presented context is an application case of the Data Journey framework. It demonstrates its applicability, potential, and the ability to adjust details in the proposal. The main advantage is the potential for asset documentation, enabling control and precise utilization. The framework optimizes data and application infrastructure to match expected data dimensions, from inactivity to high-demand operations. Experience highlights the need to segregate context domains, maintaining a distinct data-centric domain and plant asset domains. Future work aims to expand the framework to embrace multidomain contexts, avoiding double mapping. Customizing open-source BPM notation tools can enhance component options for new proposal diagram marks.

8. CONCLUSION

This paper introduces a novel approach to mapping and classifying assets within industrial contexts, specifically focusing on a case study conducted in an automotive inline plant. The

study successfully mapped and classified various components and data assets, providing significant benefits to both the research group and the practitioners involved. The proposed approach demonstrates the suitability of a unified tool capable of representing both data and process elements in a semantic and functional manner. This comprehensive view enables effective governance and data management practices within the context of asset mapping and classification. The tool's capabilities align with the envisioned governance vision, facilitating efficient decision-making processes and ensuring seamless integration between data and process management aspects.

9. REFERENCES

[1] Mehta S, Kothuri P, Garcia DL. A big data architecture for log data storage and analysis. In: *Stud. Comput. Intell.* Springer Verlag, pp. 201–209.

[2] Open Group TOG. The ArchiMate 3.2 Specification, http://www.opengroup.org (2022).

[3] Inc. OMG. Business Process Model and Notation (BPMN) 2.0, http://www.omg.org/spec/BPMN/2.0.

[4] Jacinto M, Rivera M, Viacava G. Lean Service and BPM to Increase the Efficiency of an Operational Process in the Insurance Sector. In: *ACM Int. Conf. Proc. Ser.* Association for Computing Machinery, pp. 218–222.

[5] Fernandes J, Reis J, Melão N, et al. The role of industry 4.0 and bpmn in the arise of condition-based and predictive maintenance: a case study in the automotive industry. *Appl Sci*; 11. Epub ahead of print 2021. DOI: 10.3390/app11083438.

[6] Esposito C, Cosenza C, Gerbino S, et al. Virtual shimming simulation for smart assembly of aircraft skin panels based on a physics-driven digital twin. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 2022; 16: 753–763.

[7] de Oliveira Cesar de Moraes HR, Sanchez O, Brown S, et al. Trust and distrust in big data recommendation agents. In: *Int. Conf. Inf. Syst., ICIS*. Association for Information Systems (2019).

[8] Guerrero-Prado JS, Alfonso-Morales W, Caicedo-Bravo EF. A data analytics/big data framework for advanced metering infrastructure data. *Sensors*; 21. Epub ahead of print 2021. DOI: 10.3390/s21165650.

[9] Zarour K, Benmerzoug D, Guermouche N, et al. A systematic literature review on BPMN extensions. *Business Process Management Journal* 2020; 26: 1473–1503.

CADRU PENTRU CĂLĂTORIA DATELOR: PROPUNERE DE REPREZENTARE FORMALĂ A AGENȚILOR DE DATE, A ACTIVELOR ȘI A STĂRILOR DE LA GENERAREA DATELOR PÂNĂ LA SERVIREA DATELOR ÎNTR-UN MEDIU DE OPERARE CENTRAT PE DATE

În peisajul în continuă expansiune al operațiunilor cu consum intens de date, cum ar fi Data Analytics, Machine Learning și Deep Learning, unde resursele de calcul masive și datele extinse sunt esențiale, nevoia de atribute fiabile și consecvente pentru date și agenții de operare devine primordială. În timp ce abordările precum Archimate® și BPMN permit reprezentarea funcțională și semantică a arhitecturii și proceselor organizaționale la un nivel de abstractizare mai înalt, contextului de date îi lipsește un cadru cuprinzător capabil să surprindă aspectele semantice și funcționale ale componentelor sale, de la infrastructuri minore la infrastructuri mari de date. Această lucrare propune un nou cadru ontologic care abordează acest decalaj, oferind o abordare de reprezentare semantică și practică pentru contextele de afaceri centrate pe date. Modelul propus cuprinde întreaga călătorie a datelor, pornind de la generarea datelor, trecând prin diferite stadii de maturitate în cadrul lanțului valoric și culminând cu utilizarea lor de către platformele de afaceri pentru consumatori. Cadrul își propune să fie independent de furnizor și inteligibil în diverse organizații și ecosisteme tehnologice, încurajând interoperabilitatea și colaborarea. Prin valorificarea acestui cadru ontologic, organizațiile pot îmbunătăți operațiunile cu date, asigurând fiabilitatea, consistența și conformitatea, facilitând în același timp comunicarea eficientă și luarea deciziilor în cadrul și între entități.

***Cuvinte cheie**: big data, cadru, proces de afaceri, întreprindere.*

[a] Alexandre SURKUS-CASTRO, M.Sc. Alexandre.Castro@pucpr.edu.br
[b] Fernando DESCHAMPS, Ph.D., fernando.deschamps@pucpr.br
[c] Edson PINHEIRO DE LIMA, Ph.D., pinheiro@utfpr.edu.br
[d] Déborah SARRIA, M.Sc., deborahsarria@gmail.com
[e] Anis ASSAD NET, M.Sc, anis.assad@gmail.com
[f] Sergio GOUVEA E. DA COSTA, Ph.D., gouvea@utfpr.edu.br

[a, b, d, e] Industrial and Systems Engineering, Polytechnical School - Pontifical Catholic University of Paraná, Brazil. Av. Imaculada Conceição, 1155 – Prado Velho - Bloco 2 Azul – 2º andar, CEP80.215 901, +55-41-3271-2579

[c, f] Industrial and Systems Engineering, Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brazil. Av. Sete de Setembro 3165, CEP80230-901, +55-41-3310-4626