

INTELIGÊNCIA ARTIFICIAL PARA ENRIQUECIMENTO SEMÂNTICO DE METADADOS

Enfrentando os Desafios dos Pântanos de Dados na Ciência de Dados

ARTIFICIAL INTELLIGENCE FOR SEMANTIC METADATA ENRICHMENT

Addressing Data Swamp Challenges in Data Science

Katllyn Carina Rezende

Faculdade de Tecnologia de Porto Ferreira, FATEC

E-mail: katllyn.rezende@aluno.cps.sp.gov.br

Thais Ferreira Lemos

Faculdade de Tecnologia de Porto Ferreira, FATEC

E-mail: thais.lemos@aluno.cps.sp.gov.br

Renata Firmino Lima Rocha

Faculdade de Tecnologia de Porto Ferreira, FATEC

E-mail: renata.rocha@cps.sp.gov.br

Felipe Ivo da Silva

Universidade Federal de São Carlos, UFSCar

Faculdade de Tecnologia, FATEC

E-mail: felipe.silva04@cps.sp.gov.br

RESUMO

O crescimento exponencial dos dados não estruturados transformou os repositórios de Big Data em ambientes complexos e desorganizados conhecidos como pântanos de dados. Nesse contexto, a catalogação manual de metadados torna-se impraticável devido ao volume e à diversidade das informações, comprometendo a governança de dados, a segurança e a conformidade regulatória. Este estudo propõe um modelo de arquitetura híbrida baseado em Inteligência Artificial para o enriquecimento automático de metadados, integrando Processamento de Linguagem Natural (PLN) e Visão Computacional para analisar documentos textuais e baseados em imagens. A solução atua de forma proativa durante a ingestão de dados e de forma reativa em repositórios legados, promovendo classificação, rastreabilidade e organização automatizadas. A metodologia baseia-se em uma revisão de literatura e análise conceitual da automação de metadados, do processamento distribuído e da orquestração de pipelines. Como resultado, o modelo proposto reduz as limitações associadas à curadoria manual, fortalece as práticas de

governança de dados e melhora o ciclo de vida analítico de ambientes orientados por dados. Além disso, o modelo contribui para a escalabilidade, a eficiência operacional e a gestão segura da informação, apoiando as organizações na extração de maior valor de grandes volumes de dados heterogêneos.

Palavras-chave: Inteligência Artificial, Metadados, Processamento de Linguagem Natural, Visão Computacional, Governança de Dados.

ABSTRACT

The exponential growth of unstructured data has transformed Big Data repositories into complex and disorganized environments known as data swamps. In this context, manual metadata cataloging becomes impractical due to the volume and diversity of information, compromising data governance, security, and regulatory compliance. This study proposes a hybrid architecture framework based on Artificial Intelligence for automatic metadata enrichment, integrating Natural Language Processing NLP and Computer Vision to analyze textual and image-based documents. The solution operates proactively during data ingestion and reactively on legacy repositories, promoting automated classification, traceability, and organization. The methodology is based on a literature review and conceptual analysis of metadata automation, distributed processing, and pipeline orchestration. As a result, the proposed model reduces the limitations associated with manual curation, strengthens data governance practices, and improves the analytical lifecycle of data-driven environments. Furthermore, the framework contributes to scalability, operational efficiency, and secure information management, supporting organizations in extracting greater value from large volumes of heterogeneous data.

Keywords: Artificial Intelligence, Metadata, Natural Language Processing, Computer Vision, Data Governance.

REFERÊNCIAS

- [1] HAI, R.; QUIX, C.; JARKE, M. Data Lake Management: Challenges and Opportunities. **Journal of Data Semantics**, Springer, vol. 12, no. 1, p. 23–38, 2023. Disponível em: <https://link.springer.com/article/10.1007/s44230-025-00106-5>.
- [2] KERN, C. J.; SCHÄFFER, T.; STELZER, D. Towards Augmenting Metadata Management by Machine Learning. **INFORMATIK 2021**, p. 455–468, 2021. Disponível em: <https://dl.gi.de/items/dc4718bc-c324-400a-bd3a-213ca21d2b2a>.

- [3] POLIMENO, A.; BRAGHIN, C.; ANISETTI, M.; ARDAGNA, C. A. Maximizing Data Quality While Ensuring Data Protection in Service-Based Data Pipelines. **Journal of Big Data**, vol. 13, no. 2, p. 88–105, 2025. Disponível em: <https://link.springer.com/article/10.1186/s40537-025-01118-5>.
- [4] SANTOS, J.; SILVA, R. Performance Evaluation of Distributed Task Queue Systems for Big Data Processing. **Journal of Systems and Software**, vol. 168, p. 110–122, 2020.
- [5] SMITH, R. An Overview of the Tesseract OCR Engine. In: **Proceedings** of the Ninth International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2007. p. 629–633. Disponível em: <https://ieeexplore.ieee.org/document/4376991>.
- [6] CELERY. Celery: Distributed Task Queue. Version 5.3 Documentation. 2023. Disponível em: <https://docs.celeryq.dev/>.
- [7] BRAZIL. **Law No.** 13,709, of August 14, 2018. General Personal Data Protection Law (LGPD). Brasília, DF: Office of the President, 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/113709.htm.
- [8] APACHE KAFKA. Apache Kafka Documentation. Version 3.6. 2023. Disponível em: <https://kafka.apache.org/documentation/>.
- [9] DEVLIN, J.; CHANG, M. W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: **Proceedings** of NAACL-HLT 2019. Minneapolis: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: <https://aclanthology.org/N19-1423/>.