Towards Sustainable Fishing: Enhancing Data Management in the Araguaia-Tocantins Basin with AI Technologies

Felipe Castanheira[1], Saullo Andrade[1], Luna Loyolla[1], Marcela Souza[1], Hugo Kuribayashi[1], Keid Sousa[1], Cristiane Cunha[1]

[1]Universidade Federal do Sul e Sudeste do Pará (Unifesspa) - Marabá, PA - Brasil

{felipevcast,saullo.guilherme7,luna.loyolla}@unifesspa.edu.br

{marcela.alves,hugo,keid.sousa,crisvieira_cunha}@unifesspa.edu.br

Abstract. *The Araguaia-Tocantins basin, essential for riverine communities due to its importance in artisanal fishing, faces a substantial lack of information, preventing the formulation of effective policies. Although the Fishery Statistics System (SIEPE) exists as a traditional information system, the proposed integration with Artificial Intelligence (AI) platforms allows the formulation of natural language queries that are converted into Structured Query Language (SQL) commands, thereby increasing the accessibility of the available data. This innovation has the potential to improve the management of artisanal fishing and create new business opportunities, focusing on sustainability and the preservation of natural resources.*

1. Context

The Araguaia-Tocantins hydrographic basin is the second largest in Brazil and holds significant importance for the communities residing along its banks, as it provides employment, income, food security, and cultural identity [Prysthon et al. 2022]. The latest bulletin on extractive fishing data highlighted that the state of Pará is the second largest fish producer in Brazil, underscoring its relevance within the national context [MPA 2012].

Despite the related literature identifying research efforts aimed at monitoring fishing activities in the region, there is still a lack of initiatives that include dynamic solutions intended to enhance artisanal fishing in the Araguaia-Tocantins basin. Furthermore, Information Systems are the primary agents of economic growth and social transformation in Brazil and around the world, highlighting the essential need for a solution of this nature [Boscarioli et al. 2017].

In this context, SIEPE[1] has been promoting the storage, treatment, and transformation of fishing data in the Araguaia-Tocantins basin [da Silva et al. 2019]. Currently, the SIEPE stores data from 18 fishing communities in 7 municipalities in the state of Pará, managing information concerning 170 Production Units, which involve a total of 292 fishermen. The database encompasses various variables, including temporal aspects (such as the month of registration), productive factors (fishing costs), social factors (the identity and location of the fishermen), geographical features (characteristics of the fishing site) and instrumental elements (tools and vessels used). In addition, it contains detailed information on catch weight, profit and different types of fish caught, both for consumption and

[1] Available at: https://siepe-main-production.up.railway.app/view/inicio/

for sale. This data structure is essential for understanding and optimizing the management of artisanal fishing in the region [Cunha and Sousa 2021].

Although the SIEPE platform has the potential to integrate information into a software solution capable of representing the complexity of data, the data reading and analysis techniques are becoming increasingly complex and costly, leading to increased labor demands and a decrease in the number of people capable of understanding the information generated through statistical studies [Wong et al. 2021]. To mitigate this situation, this work proposes a solution that presents an automated data analysis approach, supported by Machine Learning (ML) techniques.

In particular, Large Language Modelss (LLMs) and Natual Language Processing (NLP) techniques are enabling technologies that render data observation and decision-making more accessible and dynamic. In this context, the updating of a traditional information system typically occurs more slowly and requires significant time to incorporate new information and analyzes. In contrast, data analysis through ML tends to facilitate data interpretation, allowing users without technical expertise to extract analyses quickly and efficiently, thus enhancing agility in decision-making.

2. Process

The SIEPE platform is a traditional PHP-developed information system, mainly consisting of nine (9) features intended to monitor fishing dynamics in the Araguaia-Tocantins basin. The application, executed within a Docker container that integrates the Web interface and the relational PostgreSQL database, establishes communication via JavaScript Objet Notation (JSON) with specialized services that operate in an integrated manner to coordinate the persistence, presentation, and analysis of the fishing data managed by the platform. The platform incorporates structured statistical functions that support fisheries research, including metrics such as total total fish catch per month, ranking of the top five fishing communities by production volume, and the Capture Per Unit Effort (CPUE) index, which enables efficiency evaluations of fishing activities.

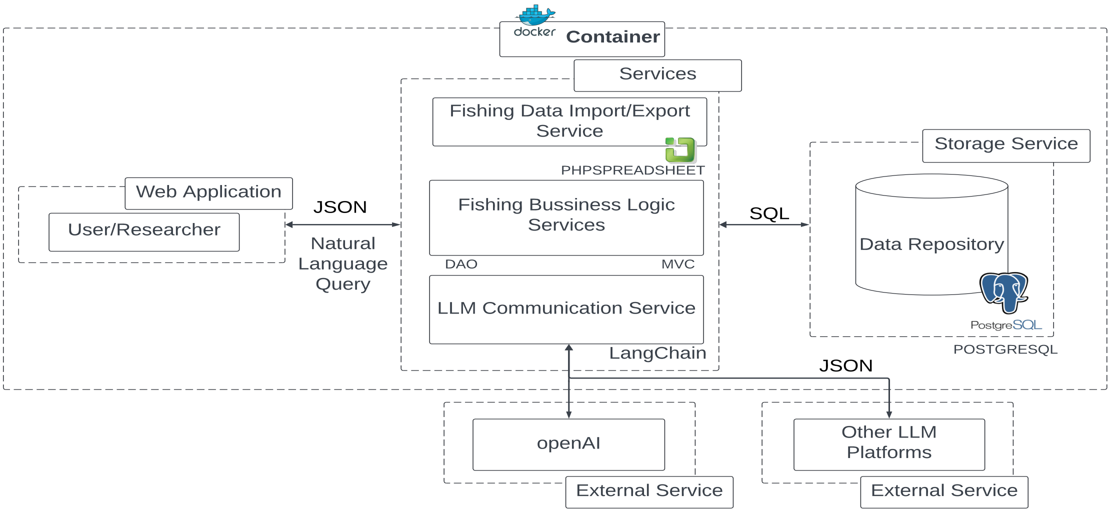


Figure 1. Representation of SIEPE's integration capabilities with LLM platforms.

Based on the perspectives of improving data analysis through integration with

Artificial Intelligence (AI) resources, Figure 1 presents a representation of the new implementation architecture of Sistema de Estatística Pesqueira (SIEPE), which includes integration with LLM platforms. The architecture highlights the integration of the interface (web application) with the services provided by its back-end, in order to demonstrate how these services aim to facilitate access to and manipulation of data by the end-users.

As a distinguishing feature of the new architecture, there is strong integration with the Langchain Framework [Langchain 2023], which provides a NLP infrastructure for data analysis on the SIEPE platform, such as the work by [Korat 2024]. The basic premise is the adoption of an interaction model that transforms queries in natural language into executable Structured Query Language (SQL) commands in the SIEPE database, taking into account the implementation schema details of the SIEPE data model. The main idea behind this approach is to provide a cost-effective way to analyze complex fish data, as it optimizes resource utilization while improving efficiency.

The process begins when a user inputs a query in natural language, which is subsequently processed by LangChain (as depicted in Figure 2). This processing involves executing a series of steps, including tokenization of the input, the identification of entities, and the interpretation of the user's intent (semantic parsing), culminating in the generation of an appropriate SQL query. The proposed schema utilizes a single SQL agent; however, employing multiple agents tends to facilitate the execution of various operations, allowing dynamic interactions with the data environment. Future improvements may include optimizations in the SQL generation process by incorporating domain-specific adaptations.

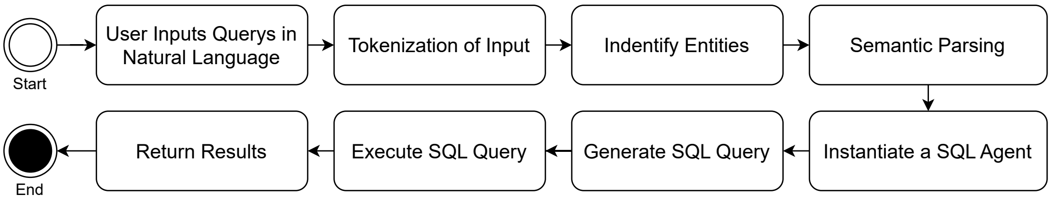
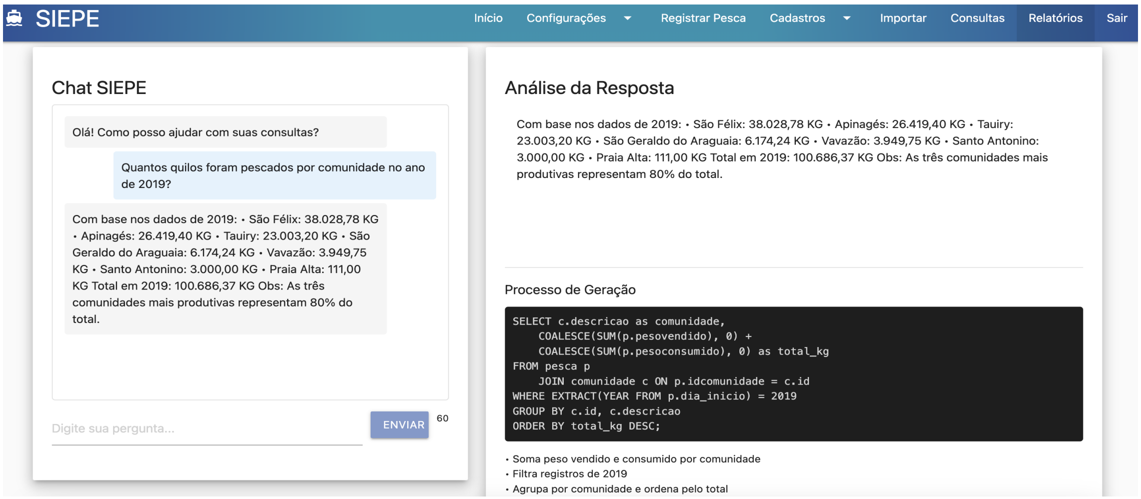


Figure 2. Operational flow representation.

Furthermore, this architecture incorporates a service that communicates with the LLMs platforms, which are combined with LangChain to ensure that the queries formulated in SQL can be executed in the SIEPE database. In the current reference implementation of the SIEPE, LangChain operates in conjunction with the OpenAI platform through the mini GPT-4 model to convert the formulations made by end users into SQL queries. In addition to the OpenAI platform, the proposed solution can be integrated with other LLM models and platforms, allowing a wide range of adaptable solutions, thus reducing the dependence on market players or specific solutions.

3. Solution

The proposed integration into the SIEPE architecture has the potential to enhance the collection and analysis of fishing data by allowing queries in natural language, eliminating the need for technical knowledge in SQL, as shown in Figure 3. This innovation makes obtaining information more accessible and efficient, facilitating data-driven decision-making. In small-scale fishing, where statistical analysis and monitoring are challenging, this solution improves data organization and optimizes its interpretation.



Chat SIEPE

Olá! Como posso ajudar com suas consultas?

Quantos quilos foram pescados por comunidade no ano de 2019?

Com base nos dados de 2019: • São Félix: 38.028,78 KG • Apinagés: 26.419,40 KG • Tauiry: 23.003,20 KG • São Geraldo do Araguaia: 6.174,24 KG • Vavazão: 3.949,75 KG • Santo Antonino: 3.000,00 KG • Praia Alta: 111,00 KG Total em 2019: 100.686,37 KG Obs: As três comunidades mais produtivas representam 80% do total.

Análise da Resposta

Com base nos dados de 2019: • São Félix: 38.028,78 KG • Apinagés: 26.419,40 KG • Tauiry: 23.003,20 KG • São Geraldo do Araguaia: 6.174,24 KG • Vavazão: 3.949,75 KG • Santo Antonino: 3.000,00 KG • Praia Alta: 111,00 KG Total em 2019: 100.686,37 KG Obs: As três comunidades mais produtivas representam 80% do total.

Processo de Geração

```sql
SELECT c.descricao as comunidade,
    COALESCE(SUM(p.pesovendido), 0) +
    COALESCE(SUM(p.pesoconsumido), 0) as total_kg
FROM pesca p
    JOIN comunidade c ON p.idcomunidade = c.id
WHERE EXTRACT(YEAR FROM p.dia_inicio) = 2019
GROUP BY c.id, c.descricao
ORDER BY total_kg DESC;
```

- Soma peso vendido e consumido por comunidade
- Filtra registros de 2019
- Agrupa por comunidade e ordena pelo total

Figure 3. Generation of an SQL query from natural language question.

The decision to integrate LLMs into SIEPE platform, rather than rely solely on traditional Business Intelligence (BI) solutions (for example), is grounded in the inherent limitations of conventional BI tools and the advanced capabilities that a natural language model can provide within the SIEPE context. While BI solutions are effective in generating dashboards and dynamic reports, offering descriptive analytics and interactive data visualization, they often require a rigidly structured data model and technical expertise for manipulation. Moreover, BI systems tend to function passively, requiring users to navigate dashboards or input structured queries. In contrast, a LLM-based approach facilitates dynamic interactions, allowing the system to interpret user intent and automatically refine responses, thus enhancing the querying and analytical experience.

The proposed approach demonstrates the impact of technology on environmental and economic management, promotes greater efficiency in the administration of natural resources, and directly benefits river communities. In the context of artisanal fishing, the proposed solution has the potential to improve data organization and optimize its interpretation, as well as act as a catalyst for the formulation of more effective public policies and the promotion of sustainable practices aimed at preserving natural resources and strengthening riverine communities.

This implementation represents an advance in the way complex data is manipulated and accessed within interconnected information ecosystems. In addition to facilitating research and analysis in the fishing context, the approach demonstrates potential for expansion into other areas that require intelligent processing and extraction of knowledge from large volumes of structured data. In this sense, there is an opportunity for the creation of a GreenTech focused on developing environmental monitoring solutions, in response to the lack of initiatives in the sector and the increasing relevance of these topics in contemporary discourse.

Finally, the proposed approach increases future expansion possibilities for the SIEPE platform, paving the way toward the inclusion of predictive analysis and recommendation systems based on machine learning mechanisms, thus improving the impact on the formulation of public policies and the sustainable management of fishery resources.

References

Boscarioli, C., Araujo, R., and Suzana, R. (2017). *I GranDSI-BR Grand Research Challenges in Information Systems in Brazil 2016-2026 Organized by*. Socieadade Brasileira de Computação.

Cunha, C. and Sousa, K. (2021). *Monitoramento Adaptativo da Pesca na Média Bacia Araguaia-Tocantins na Amazônia Brasileira, Pará, Brasil*, chapter 17. CRV.

da Silva, R. S., da Silva, R. R., Kuribayashi, H. P., da Cunha, C. V., Francês, C. R. L., and Sousa, K. N. S. (2019). Clusterização de dados mistos para análise da atividade pesqueira artesanal na bacia araguaia-tocantins. *Revista Brasileira de Computação Aplicada*, 11(3):155–164.

Korat, A. S. (2024). Ai-augmented langchain: Facilitating natural language sql queries for non-technical users. *Journal of Artificial Intelligence & Cloud Computing*, 3(3):1–5.

Langchain (2023). Langchain: A Framework for Building Applications with LLMs. Accessed: 2025-01-12.

MPA (2012). *Ministério da Pesca e Aquicultura: Boletim Estatístico da Pesca e Aquicultura - Brasil 2011*. ICMBio.

Prysthon, A., Ummus, M., Tardivo, T., Pedroza Filho, M., Chicrala, P., Kato, H., Dias, C., and Paz, L. (2022). *A Pesca Artesanal no Rio Araguaia, Tocantins, Brasil: Aspectos Tecnológicos e Socioeconômicos*. e-Publicar.

Wong, A., Joiner, D., Chiu, C., Elsayed, M., Pereira, K., Khmelevsky, Y., and Mahony, J. (2021). A Survey of Natural Language Processing Implementation for Data Query Systems. In *2021 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, pages 1–8.