This Future Without SQL

Eduardo C. de Almeida, Eduardo H. M. Pena, Altigran S. da Silva

[1] Departamento de Informática – Universidade Federal do Paraná (UFPR)
Curitiba, PR, Brazil

[2]Universidade Tecnológica Federal do Paraná (UTFPR)
Brazil

[3]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Manaus, Amazonas, Brazil

{eduardo@inf.ufpr.br, eduardopena@utfpr.edu.br, alti@icomp.ufam.edu.br}

Resumo. *O futuro da gerência de dados está se voltando para sistemas que processam consultas diretamente em linguagem natural, eliminando a necessidade de SQL. Diferente dos sistemas NL-para-SQL, que traduzem a linguagem natural para SQL, essa nova abordagem permite a interação direta com bancos de dados, impulsionada pelos avanços em processamento de linguagem natural (NLP) e inteligência artificial (IA). Propomos que a gerência de dados do futuro aproveite o PLN para lidar com consultas complexas e sensíveis ao contexto, com foco em aprendizado adaptativo para refinar a interpretação das consultas, além de enfrentar os desafios de segurança e privacidade. Nossa visão de "Linguagem Natural para Bancos de Dados" (NL-para-DB) integra soluções de NLP com estruturas robustas, possibilitando uma adaptação intuitiva e personalizada das consultas. Essa mudança repensa a forma como acessamos informações, promovendo inclusão e novas possibilidades na gerência de dados.*

Abstract. *The future of data management is shifting towards systems that process queries directly in natural language, eliminating the need for SQL. Unlike NL-to-SQL systems, which translate natural language into SQL, this new approach allows direct interaction with databases, driven by advances in NLP and AI. We propose that future data management leverage NLP for complex, context-aware queries, focus on adaptive learning to refine query interpretation, and address security and privacy challenges. Our vision for a "Natural Language to Databases" (NL-to-DB) integrates NLP solutions with robust frameworks, enabling intuitive, user-specific query adaptation. This shift reimagines how we access information, promoting inclusivity and innovation in data management.*

1. Introduction

The evolution of data management has long relied on Structured Query Language (SQL) as the main interface for interacting with databases. SQL introduced a declarative approach to querying and manipulating relational data with intuitive table-based structures, establishing itself as the industry standard [Abadi et al. 2022]. While SQL provides powerful and precise query capabilities, it imposes a steep learning curve for non-technical users, limiting broader accessibility. As data becomes more complex and widespread, the need for more intuitive ways to interact with databases has grown. Natural Language

Processing (NLP) and Artificial Intelligence (AI) offer promising alternatives by enabling users to query databases in natural language, bypassing the SQL syntax constraints.

This paper discusses the potential of a SQL-free future, where natural language interfaces democratize access to data, allowing users across various industries to interact with databases more naturally and efficiently. It discusses the advantages of this data access shift, such as improved accessibility and query efficiency, while addressing the technical, security, and adoption challenges. We are not proposing yet another form of NoSQL database. Instead, this vision seeks to redefine declarative data interaction within the relational model, which remains the dominant data model [Abadi et al. 2022], making it more inclusive and intuitive for a wider audience.

Consider a lawyer or an accountant who wants to check the database of a financial system to verify whether a transaction complies with specific regulations. This professional may not understand the underpinnings of the relational model required to explore a large schema with several hundred relations correctly or be able to program a very long SQL code with multiple operators. For example, in the TPC-DS benchmark[1], many queries reflect such real-world, high-level questions. The textual specification for query #88 outlines a complex business rule in 3 lines of plain English. The implementation of the corresponding SQL code spans almost 100 lines with aggregations, inline views, and dozens of joins, among other operators. Programming such queries requires a database background, which limits non-experts' flexibility in exploring the data.

This paper overviews some of the limitations of SQL for database interactions and proposes natural language interfaces as a more accessible solution. It reviews the current state of SQL and Natural Language to SQL (NL-to-SQL) systems and then discusses the advantages of transitioning to a SQL-free future. Then, it explores the technical, security, and adoption challenges involved, offering strategies for gradual implementation. Finally, it concludes with key takeaways, emphasizing the potential for innovation and providing recommendations for stakeholders to support this transition.

2. Current State of Art and Practice

As data becomes increasingly democratic and complex, SQL's intricate syntax poses barriers for many non-technical users who now need to work with databases. Reflecting on SQL's original vision, Don Chamberlin, its co-inventor, recently emphasized, "Database queries should not look like programs that tell the computer what to do. We wanted to express queries in a high-level, non-procedural language," [Chamberlin 2024]. With recent advances in NLP, it might be the right time to revisit SQL and explore more intuitive ways for users to interact with data.

Text-to-SQL systems aim to bridge the gap between user proficiency and SQL-based data retrieval by translating natural language queries into SQL statements [Katsogiannis-Meimarakis and Koutrika 2021, Li et al. 2024]. This approach has gained traction recently, driven by advances in large language models like GPT and specialized benchmarks [Gkini et al. 2021]. These lines of research approach the text-to-SQL task as a form of language translation, training neural networks on large datasets of paired natural language questions and their cor-

[1] https://www.tpc.org/tpcds/

responding SQL queries. However, such systems must overcome several challenges [Katsogiannis-Meimarakis and Koutrika 2023], including ambiguity in natural language, difficulty with complex queries, and mismatches between user terms and database schema. They often struggle with advanced SQL functions like aggregates and domain-specific terminology. Additionally, they lack effective error handling and feedback mechanisms, making it hard to capture the user's intent in SQL form accurately.

Another alternative approach is to move beyond NL-to-SQL systems and focus directly on NL-to-DB systems, also known as natural language interfaces to databases (NLIDB) [Li and Jagadish 2014]. It resembles search engine interfaces and often overlaps with research in keyword search interfaces [Yu et al. 2010]. An orthogonal line of research is table question answering (TQA), which focuses on extracting information from structured data in response to NLP queries [Pasupat and Liang 2015]. TQA represents a multifaceted challenge that requires a blend of language comprehension, logical analysis, and data interpretation skills. The goal is to process a user's query, understand the underlying tabular structure, and provide precise responses through reasoning and data extraction [Zhang et al. 2023]. In industry, NLIDBs like Tableau's Ask Data, Power BI, and Cognos Assistant exemplify a relatively new wave of tools to allow non-technical users to ask questions and explore data sets through simple, conversational queries.

3. A Claim for This Future Without SQL

SQL as the primary database interface poses challenges for non-technical users, particularly as data complexity grows. While NL-to-SQL systems translate natural language into SQL, they still depend on SQL, limiting intuitive data interactions. We propose eliminating SQL as an intermediary, allowing users to query databases directly through natural language. Advances in NLP and AI enable this shift, making data access more seamless and intuitive for non-experts, reducing complexity, and enabling context-aware responses.

A SQL-free approach democratizes data access by allowing non-technical users to interact with databases naturally without needing to learn SQL. This broadens access to data-driven decision-making and reduces project delays in real-time environments. Thanks to advancements in NLP and AI, these systems provide more accurate, context-sensitive responses, capturing user intent more effectively. Additionally, they are adaptable, learning from user interactions and evolving to meet specific domain needs, improving accuracy in areas like healthcare and finance.

Adopting a SQL-free system requires retraining and infrastructure changes, which can be both costly and time-consuming. Users may view natural language interfaces as less reliable than SQL, necessitating a gradual transition to build trust in the new system. A hybrid approach is a practical solution, allowing organizations to retain SQL while gradually adopting natural language interfaces. This phased transition minimizes disruption and prepares the groundwork for a fully SQL-free future.

Implementing SQL-free systems transforms data access across healthcare, finance, and law industries. Non-technical users can interact with data directly, focusing on insights rather than query mechanics. This shift could create new services in customer support, personalized recommendations, and faster decision-making processes. As these systems mature, new applications for natural language-driven interactions will emerge. Indeed, a SQL-free future fosters interdisciplinary collaboration by making data systems

more accessible to professionals from diverse fields. This democratization of data access breaks traditional silos, encouraging more collaborative, data-driven projects. Designing these systems will require collaboration across AI, linguistics, and domain experts, enhancing the system's contextual accuracy.

The shift to a SQL-free paradigm drives innovation in databases, NLP, and AI. This challenges traditional database interaction, opening possibilities for more flexible and adaptive systems and pushing advances beyond databases into conversational agents and context-aware computing. SQL-free innovations will influence other areas like user interfaces and decision support systems. Natural language could be embedded into productivity tools, creating more intuitive environments where data insights are readily accessible, ultimately expanding the scope of AI-powered assistants and automation.

Key performance indicators are essential to successfully transitioning to an SQL-free system. These include improved accessibility for non-technical users, query precision in interpreting complex natural language, and operational efficiency regarding response time and resource usage. Measuring the success of SQL-free techniques can also be evaluated by query accuracy and response relevance in aligning generated results with user intent. Specific metrics include query success rate, user satisfaction, execution efficiency, and resource consumption. The system's adaptability and continuous learning from user interactions, measured by reduced error rates and improved handling of complex queries, are crucial for long-term effectiveness. Recent database management and information retrieval research has already started investigating and implementing some of these metrics [Xing et al. 2024, Afonso et al. 2024].

4. Challenges Ahead

Our vision aligns with the Seattle Report on Database Research on declarative language abstractions [Abadi et al. 2022]. However, we see numerous challenges ahead.

Technical Challenges. While advances in NLP have made it possible to write code in various programming languages, including SQL, writing complex queries that capture business rules and regulations remains an open challenge. Complex rules require new sophisticated compilers and substantial computing resources for execution. Even state-of-the-art DBMSs can take hours to process rules with just a few predicates due to inefficient intermediate memory representations [Pena et al. 2021]. A new NLP engine will face the same scalability problem. The intermediate representations to support NLP queries, such as Candidate Joining Networks(CJNs) and Query Matches(QMs) are still in initial development stages [Martins et al. 2023]. A key metric for evaluating progress in SQL-free systems is their increasing adoption.

Privacy and Security Concerns. Differential privacy and homomorphic encryption are the state-of-the-art approaches in SQL engines. The former ensures that statistical properties remain intact even without individual data [de Farias et al. 2020]. The latter supports unbounded database aggregation queries. However, recent research shows that the homomorphic encryption in relational databases is yet slower than plaintext processing in magnitudes [Ren et al. 2022], leaving room for new initiatives.

Databases are increasingly exposed to sophisticated cyberattacks. SQL injection is a well-known form of attack, but we are still in the early stages of understanding "NLP

injection" vulnerabilities like prompt injection or malicious queries. These challenges are new for relational databases if NLP becomes the query language. Assessing the progress of new solutions includes addressing new types of attacks that will certainly appear.

Adoption and Usability Challenges. Many data scientists rely on programming interfaces, such as Jupyter and Zeppelin, in their daily work despite their limitations in collaborative coding and operationalizing code. Non-technical users may not be familiar with these programming interfaces. Therefore, it is essential to design intuitive NLP interfaces to improve usability, similar to industry initiatives such as Cognos and Tableau.

We must also ensure usability with reliable performance to increase adoption, particularly for critical applications. NLP-to-DB can be initially implemented as extensions of SQL engines utilizing new compilation paradigms [Jungmair et al. 2022] to benefit from the internal representations of SQL engines, such as vectorization and data-centric code generation [Kersten et al. 2018], without requiring translation of NLP to SQL. For example, the DuckDB team introduced a database extension that allows querying CSV files using vectorization without uploading them to a database. Building pure NLP query engines is still being determined, but progress toward our vision can be initially measured by the development of NLP database extensions that benefit from the high performance and maturity of modern SQL engines.

Regulatory and Legal Challenges. Brazilians are engaged in a significant debate on artificial intelligence's legal and regulatory aspects, as outlined in Congress in Bill # 2338/2023. There are at least three key challenges: what should be regulated, who will be responsible for regulating, and how to implement the regulation. As for *what* to regulate, the following problems should be considered [Rodrigues 2020]: algorithmic transparency, cybersecurity vulnerabilities, unfairness, bias and discrimination, lack of contestability, legal personhood issues, intellectual property issues, adverse effects on workers, privacy and data protection issues, liability for damage and lack of accountability.

Concerning *who* and *how* to regulate, recent evidence shows that automated systems can produce unfair outcomes and exacerbate existing inequalities [Ghasemaghaei and Kordzadeh 2024, Sunyé 2020]. Additionally, the global rise of AI-generated fake news shows that much more work is needed. In Brazil, Bill # 2338/2023 formalized the National Data Protection Authority (ANPD) as the National System for Regulation and Governance of Artificial Intelligence (SIA) coordinating body. In our NLP vision, significant effort is required to create data collection and disposal frameworks that align with regulations and policy constraints.

5. Conclusions, Remarks, and Takeouts

A SQL-free paradigm offers a transformative shift in data management, making databases more accessible and intuitive. While SQL has been practical, it poses barriers for non-technical users. Natural language interfaces can democratize data access and improve efficiency. Still, significant challenges remain, as technical limitations in handling complex queries and security and privacy concerns still need to be appropriately addressed.

The SQL-free transition will require collaboration between researchers, industry leaders, policymakers, and technologists. A hybrid approach combining SQL and natural language can ease this shift. Though challenging, this future promises more inclusive, efficient data interactions, reshaping how we access and use information.

About the Proponents

Eduardo C. de Almeida is an Associate Professor at UFPR. He holds a Ph.D. in Computer Science from the University of Nantes, INRIA GDD Team, France. His research areas include data structures, query processing, and data management. He has experience in both industry and academia, having worked as a database engineer and held research positions in Luxembourg and Switzerland.

Eduardo H. F. Pena is a professor at UTFPR and a permanent faculty member at the UEM graduate program in Computer Science. His Ph.D. on data quality and metadata extraction earned the CAPES award for Best Thesis in Computing in 2021. He is currently a postdoctoral researcher at NYU, developing tools for biomedical data integration.

Altigran S. da Silva is a Full Professor at UFAM. He earned his Ph.D. from UFMG in 2002. His interests include Data Management, Information Retrieval, Data Mining, Machine Learning, and Language Models. He has coordinated and participated in dozens of research projects resulting in over 150 publications in journals and conference. He served as Dean for Research and Graduate Studies at UFAM (2007/2009), coordinator of CA-CC at CNPq (2023/2024), and adjunct coordinator of the computing area at CAPES (2011/2013). He was also a board member (2005/2015) and council member (2016/2019) of SBC. He co-founded companies such as Akwan (acquired by Google in 2005), Neemu (acquired by Linx Systems in 2015), and Teewa (acquired by JusBrasil in 2019).

References

Abadi, D., Ailamaki, A., Andersen, D. G., Bailis, P., Balazinska, M., Bernstein, P. A., Boncz, P. A., Chaudhuri, S., Cheung, A., Doan, A., Dong, L., Franklin, M. J., Freire, J., Halevy, A. Y., Hellerstein, J. M., Idreos, S., Kossmann, D., Kraska, T., Krishnamurthy, S., Markl, V., Melnik, S., Milo, T., Mohan, C., Neumann, T., Ooi, B. C., Ozcan, F., Patel, J. M., Pavlo, A., Popa, R. A., Ramakrishnan, R., Ré, C., Stonebraker, M., and Suciu, D. (2022). The seattle report on database research. *Commun. ACM*, 65(8):72–79.

Afonso, A., Martins, P., and da Silva, A. (2024). Sereia: document store exploration through keywords. *Knowledge and Information Systems*.

Chamberlin, D. (2024). 50 years of sql | don chamberlin computer scientist and co-inventor of sql. Accessed: 2024-08-31.

de Farias, V. A. E., Brito, F. T., Flynn, C. J., Machado, J. C., Majumdar, S., and Srivastava, D. (2020). Local dampening: Differential privacy for non-numeric queries via local sensitivity. *Proc. VLDB Endow.*, 14(4):521–533.

Ghasemaghaei, M. and Kordzadeh, N. (2024). Understanding how algorithmic injustice leads to making discriminatory decisions: An obedience to authority perspective. *Information and Management*, 61(2):103921.

Gkini, O., Belmpas, T., Koutrika, G., and Ioannidis, Y. (2021). An in-depth benchmarking of text-to-sql systems. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 632–644, New York, NY, USA. Association for Computing Machinery.

Jungmair, M., Kohn, A., and Giceva, J. (2022). Designing an open framework for query optimization and compilation. *Proc. VLDB Endow.*, 15(11):2389–2401.

Katsogiannis-Meimarakis, G. and Koutrika, G. (2021). A deep dive into deep learning approaches for text-to-sql systems. In Li, G., Li, Z., Idreos, S., and Srivastava, D., editors, *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, pages 2846–2851. ACM.

Katsogiannis-Meimarakis, G. and Koutrika, G. (2023). A survey on deep learning approaches for text-to-sql. *The VLDB Journal*, 32(4):905–936.

Kersten, T., Leis, V., Kemper, A., Neumann, T., Pavlo, A., and Boncz, P. A. (2018). Everything you always wanted to know about compiled and vectorized queries but were afraid to ask. *Proc. VLDB Endow.*, 11(13):2209–2222.

Li, F. and Jagadish, H. V. (2014). Constructing an interactive natural language interface for relational databases. *Proc. VLDB Endow.*, 8(1):73–84.

Li, H., Zhang, J., Liu, H., Fan, J., Zhang, X., Zhu, J., Wei, R., Pan, H., Li, C., and Chen, H. (2024). Codes: Towards building open-source language models for text-to-sql. *Proc. ACM Manag. Data*, 2(3):127.

Martins, P., Afonso, A., and da Silva, A. S. (2023). Pylathedb - A library for relational keyword search with support to schema references. In *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*, pages 3627–3630. IEEE.

Pasupat, P. and Liang, P. (2015). Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1470–1480. The Association for Computer Linguistics.

Pena, E. H. M., de Almeida, E. C., and Naumann, F. (2021). Fast detection of denial constraint violations. *Proc. VLDB Endow.*, 15(4):859–871.

Ren, X., Su, L., Gu, Z., Wang, S., Li, F., Xie, Y., Bian, S., Li, C., and Zhang, F. (2022). HEDA: multi-attribute unbounded aggregation over homomorphically encrypted database. *Proc. VLDB Endow.*, 16(4):601–614.

Rodrigues, R. (2020). Legal and human rights issues of ai: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology*, 4.

Sunyé, M. S. (2020). A quem servem os dados? *SBC Horizontes*, 4.

Xing, J., Wang, X., and Jagadish, H. V. (2024). Data-driven insight synthesis for multidimensional data. *Proc. VLDB Endow.*, 17(5):1007–1019.

Yu, J. X., Qin, L., and Chang, L. (2010). Keyword search in relational databases: A survey. *IEEE Data Eng. Bull.*, 33(1):67–78.

Zhang, L., Zhang, J., Ke, X., Li, H., Huang, X., Shao, Z., Cao, S., and Lv, X. (2023). A survey on complex factual question answering. *AI Open*, 4:1–12.