

Um Estudo sobre Modelagem Poliglota de Dados

Hudson Afonso Batista da Silva, Luís Gustavo Bornia, Ronaldo dos Santos Mello

¹Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Departamento de Informática e Estatística (INE)
Universidade Federal de Santa Catarina (UFSC)
Florianópolis – SC – Brasil

hudson.silva@ifpa.edu.br, luis.gustavo.bornia@gmail.com, r.mello@ufsc.br

Abstract. *Polyglot persistence is seen as the future of database modeling, as it aims to adapt each part of a conceptual database modeling to logical and physical schemas with the best possible performance in terms of storage and access. However, polyglot data modeling brings new challenges to the designer, such as dealing with more than one database technology and choosing the best logical model or database technology to maintain and manage a certain part of a conceptual modeling. This article presents a systematic review of the literature on this research area, an overview of the found works and a comparative analysis of them. We did not find a study similar to this in the literature.*

Resumo. *Persistência poliglota é vista como o futuro das modelagens de bancos de dados, pois visa adequar cada parte de uma modelagem conceitual de BD para esquemas lógicos e físicos com o melhor desempenho possível em termos de armazenamento e acesso. Entretanto, uma modelagem poliglota de dados traz novos desafios ao projetista, como lidar com mais de uma tecnologia de banco de dados e escolher o melhor modelo lógico ou tecnologia de banco de dados para manter e gerenciar determinada parte de uma modelagem conceitual. Este artigo apresenta uma revisão sistemática da literatura sobre esta área de pesquisa, uma visão geral dos trabalhos encontrados e uma análise comparativa dos mesmos. Não encontramos na literatura um estudo similar a este.*

1. Introdução

Sistemas de Gerenciamento de Banco de Dados Relacionais (SGBDRs) têm sido uma tecnologia de longa data para uso em sistemas de gestão de *Big Data*, porém têm apresentado limitações quando se trata de certos tipos de modelos que utilizam dados heterogêneos, ou seja, a grande variedade e formato de dados impõe limites à abrangência de sua manipulação. Os Banco de Dados Relacionais (BDRs) exigem a garantia das propriedades ACID e algumas dessas propriedades podem ser proibitivas principalmente em termos de complexidade de tratamento desta heterogeneidade de representação dos dados e conseqüente escalabilidade.

Com o surgimento dos Bancos de Dados (BDs) NoSQL houve uma melhora substancial em termos de escalabilidade para gestão de dados complexos e heterogêneos. Neste cenário, algumas pesquisas buscam apresentar abordagens de migração de BDs SQL para NoSQL a fim de mitigar as deficiências encontradas nos SGBDRs [Schreiner et al. 2019]. As vantagens em relação a escalabilidade horizontal e a flexibilidade de estrutura que os sistemas de armazenamento NoSQL possuem os tornam opções

muito interessantes no momento de projetar a infraestrutura de dados de um sistema de informação. Isso não quer dizer que os SGBDRs ficaram ou ficarão obsoletos, visto que é importante considerar que se as propriedades ACID são requisitos necessários na gestão de dados complexos em alguns domínios de aplicação [Stonebraker 2012]. Por essa razão, a abordagem de persistência poliglota para grandes BDs é apropriada para que sejam atendidas diferentes objetivos de uma aplicação.

O conceito de *persistência poliglota* deriva do conceito de programação poliglota, onde uma mesma aplicação é desenvolvida em diferentes linguagens de programação [Fowler 2011]. Neste sentido, a persistência poliglota assume que uma determinada tecnologia de BD, por exemplo, BD relacional, mantenha e gerencie parte dos dados de uma aplicação, enquanto outra tecnologia, como um BD NoSQL, seja responsável pela gestão de outros dados que necessitem de mais flexibilidade.

Sadalage e Fowler [Sadalage and Fowler 2013] exemplificam um cenário de persistência poliglota no domínio de uma plataforma de *e-commerce* (Figura 1). Neste caso, partes dos dados do domínio são representados em modelos lógicos distintos de acordo com os seus requisitos. Dados de inventário e preços de itens são modelados no formato relacional, enquanto dados do carrinho de compras e compras finalizadas são armazenados em repositórios chave-valor e BDs de documentos, respectivamente. Ainda, serviços de recomendação utilizam BDs orientados a grafos.

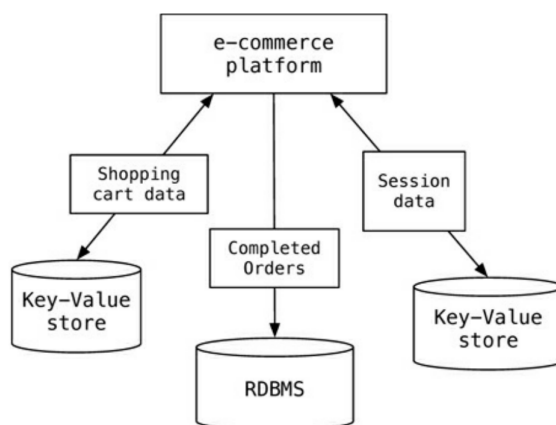


Figura 1. Exemplo de Domínio de Aplicação que utiliza Persistência Poliglota
Fonte: [Sadalage and Fowler 2013]

Kiehn et al. [Khine and Wang 2019] apresenta uma visão geral do estado da arte de soluções que manipulam dados poliglotas e destaca algumas oportunidades de pesquisa em aberto na área de persistência poliglota, como o suporte à manipulação de dados *ad-hoc* através de uma interface global, detecção de mudanças automáticas nos requerimentos dos sistemas ou carga de trabalho, sistema de cruzamento eficiente no planejamento e execução das consultas e o gerenciamento de esquemas multimodelo.

Embora existam soluções estabelecidas para o projeto de BDs Relacionais ou BDs NoSQL, percebe-se uma lacuna na literatura no que diz respeito à soluções consolidadas para a modelagem poliglota de dados ou o projeto de *BDs multimodelo*. Esses últimos se caracterizam pela utilização vários modelos lógicos derivados de acordo com possíveis requisitos, sejam eles provenientes de regras de negócios ou das características intrínsecas

dos dados e dos seus relacionamentos [Kolonko and Müllenbach 2020]. Dessa maneira, a derivação para um determinado esquema lógico deve levar em conta as vantagens e desvantagens da tecnologia ou modelo de dados alvo, visando eficiência no armazenamento e no acesso aos dados.

Entre as possíveis direções de pesquisa para a segmentação de dados em esquemas lógicos específicos, emergem questões-chaves, tais como: quais níveis de granularidade devem ser considerados para a segmentação? Como preservar os vínculos entre as partes segmentadas? Quais requisitos devem ser levados em consideração, abrangendo a estrutura do modelo conceitual, estimativa de carga de trabalho, restrições de integridade de domínio, restrições de integridade referencial e requisitos não-funcionais.

Assim sendo, este artigo apresenta e compara propostas de soluções disponíveis na literatura visando contribuir com a pesquisa relacionada à modelagem poliglota de esquemas de dados ou modelagem de dados considerando múltiplos modelos lógicos de dados. Não foi encontrada revisão sistemática similar na literatura.

Este artigo está organizado em mais 4 seções. Na Seção 2 descrevemos o método aplicado para a realização da revisão Sistemática de Literatura (RSL). Na Seção 3 apresentamos uma visão geral dos trabalhos selecionados pela RSL. Na Seção 4 comparamos os trabalhos selecionados e na Seção 5 concluímos o artigo.

2. Revisão Sistemática

Wazlawick [Wazlawick 2009] ressalta a importância dos procedimentos metodológicos para garantir a validade e a confiabilidade dos resultados obtidos em uma pesquisa na área da computação. Diante disso, uma RSL é uma metodologia que possibilita identificar, avaliar e interpretar a pesquisa disponível na literatura relacionada a uma pergunta de pesquisa, área temática ou fenômeno de interesse. Para realizar esta RSL, foi utilizado o guia proposto por Kitchenham [Kitchenham 2004], dada a sua grande aceitação na literatura.

O objetivo desta RSL é coletar, analisar, extrair e sintetizar a pesquisa realizada até agora sobre modelagem poliglota ou BDs multimodelo. A intenção é descrever as principais propostas de solução que as pesquisas na área de BDs sugerem para construir e obter uma modelagem poliglota.

O guia adotado para a RSL inicialmente sugere a definição de questões de pesquisa. Este trabalho propôs três questões de pesquisa (QP):

- **QP-1:** quais modelos lógicos de dados são considerados pela abordagem de modelagem poliglota de BDs?
- **QP-2:** quais requisitos (funcionais, não-funcionais, de domínio, etc) são considerados pela abordagem?
- **QP-3:** quais são os processos desenvolvidos pela abordagem com base nos modelos lógicos e requisitos considerados?

Na sequência, procedeu-se à busca de trabalhos relacionados em diversos repositórios de trabalhos acadêmicos. Cinco repositórios foram selecionados: *ACM*, *IEEE*, *Science Direct*, *SCOPUS* e *Springer Link*. Eles foram escolhidos como fontes primárias para a consulta de artigos visto que são amplamente utilizados em pesquisas semelhantes.

Para uma busca automatizada nesses repositórios, a segunda tarefa da RSL consistiu em definir *strings* de busca usando um conjunto de palavras-chave relacionadas ao assunto, ao corpo do texto, ao título e ao resumo dos trabalhos. Essas *strings* podem estar ligadas por conectores lógicos. As seguintes *strings* de busca utilizadas foram definidas:

- **Texto:** (“*polyglot persistence*“ AND *database* AND *modeling*) OR (*mult-model* AND *database* AND *modeling*) OR (“*polyglot persistence*“ AND *conceptual* AND *modeling*);
- **Título:** “*polyglot persistence*“ OR “*mult-model modeling*”;
- **Resumo:** *database* or “*polyglot persistence*”.

Essas *strings* de busca foram customizadas com base na interface de acesso do repositório consultado.

A partir do refinamento de busca nos títulos, resumo e texto dos artigos, a RSL requer a definição de critérios de *inclusão* e *exclusão* para reduzir o número de resultados e focar em estudos de interesse. Os seguintes critérios de inclusão foram propostos:

- Artigo de pesquisa publicado em uma veículo bem qualificado em língua inglesa;
- Estudos publicados entre 2010 e 2023;
- Assunto do estudo relacionado à modelagem poliglota de BDs.

Ainda, os seguintes critérios de exclusão foram propostos:

- Estudos secundários, ou seja, artigos que tratam do tema de persistência poliglota, porém sem foco na parte de modelagem de dados;
- Estudos duplicados;
- Artigos cujo conteúdo não seja acessível.

A Tabela 1 apresenta os repositórios pesquisados e os resultados obtidos. Inicialmente foram encontrados 596 trabalhos com a aplicação das cadeias de busca. A seguir, os critérios de inclusão e exclusão foram aplicados no título e resumo, reduzindo o número para 319. Ao aplicar o mesmo critério para o texto completo, o número final obtido foi 3.

Tabela 1. Repositórios Pesquisados e Resultados

Repositório	Seleção Inicial	Seleção Final
ACM-DL	51	0
IEEE Xplore	180	1
ScienceDirect	5	1
SpringerLink	79	1
SCOPUS	4	0
Total	319	3

Os 3 trabalhos selecionados são apresentados na Tabela 2. Esses trabalhos são detalhados na próxima seção.

Apesar de encontrarmos uma quantidade relevante de trabalhos relacionados ao tema persistência poliglota, apenas os 3 trabalhos apresentados na Tabela 2 possuem foco em modelagem poliglota de BDs. Os demais trabalhos tratam outras temáticas, como gerenciamento de dados políglotas [Lu and Holubová 2017, Khine and Wang 2019], gerenciamento de evolução de

Tabela 2. Trabalhos Selecionados através da RSL

Autores	Título	Ano	Veículo
Kolonko & Müllenbach	<i>Polyglot Persistence in Conceptual Modeling for Information Analysis</i>	2020	ACIT Conference
Roy-Hubara et.al.	<i>Selecting Databases for Polyglot Persistence Applications</i>	2022	Data and Knowledge Engineering Journal
Koupil & Holubová	<i>A Unified Representation and Transformation of Multi-model Data using Category Theory</i>	2022	Journal of Big Data

modelo de dados [Holubová et al. 2021, Vavrek et al. 2019], migração de dados políglotas [Niska 2024], inferência de esquemas [Koupil et al. 2022], desenvolvimento de linguagens e interfaces de consulta [Bakhtin 2023, Koupil et al. 2021] e *benchmarks* para avaliação de performance em BDs multimodelo [Ye et al. 2023].

3. Trabalhos Relacionados

Kolonko e Müllenbach descrevem uma proposta para visualização de um esquema conceitual particionado em múltiplas tecnologias de BDs (relacionais e não-relacionais) [Kolonko and Müllenbach 2020]. Essa divisão é feita de acordo com características estruturais dos dados em conjunto com necessidades de negócio e deve ser feita de acordo com algoritmos, existentes ou desenvolvidos para esta finalidade, pensados para lidar com sistemas políglotas. Os autores enfatizam não apenas a necessidade de adaptação de diversos algoritmos já conhecidos para a utilização em sistemas políglotas, como também mencionam que há a necessidade de desenvolvimento de algoritmos que sejam desenvolvidos com esta utilização em mente. Estes algoritmos irão, futuramente, gerar procedimentos semiautomáticos para a seleção da tecnologia recomendada ao modelador, inovação ainda não vista na proposta inicial dos autores.

A proposta dos autores é ampla, descrevendo que haverá um processo desde o esquema lógico até o esquema físico, cujo desenvolvimento será feito em outros artigos futuros, tanto pelos autores quanto por outros pesquisadores. Destaca-se, também, a importância de desenvolvimento de algoritmos agnósticos em relação ao paradigma, ou seja, que possam ser aplicados em qualquer combinação de tecnologias de BDs.

A Figura 2 mostra as etapas propostas para um projeto políglota de BD. Inicialmente um esquema conceitual é particionado utilizando-se algoritmos de transformação que consideram não apenas as características dos dados como também as necessidades do negócio. Os esquemas resultantes são esquemas lógicos específicos para cada tecnologia de BD utilizada no projeto que, por fim, são transformados em esquemas físicos.

O trabalho conclui ponderando a sua importância e dando indicações de quais áreas serão focadas em trabalhos subsequentes. Em especial, são mencionadas 2 áreas principais de contribuições futuras: (i) desenvolvimento de algoritmos para a transformação de esquemas conceituais em esquemas conceituais menores, focados em diferentes tecnologias de BDs utilizados em um suposto projeto políglota e; (ii) projeto e

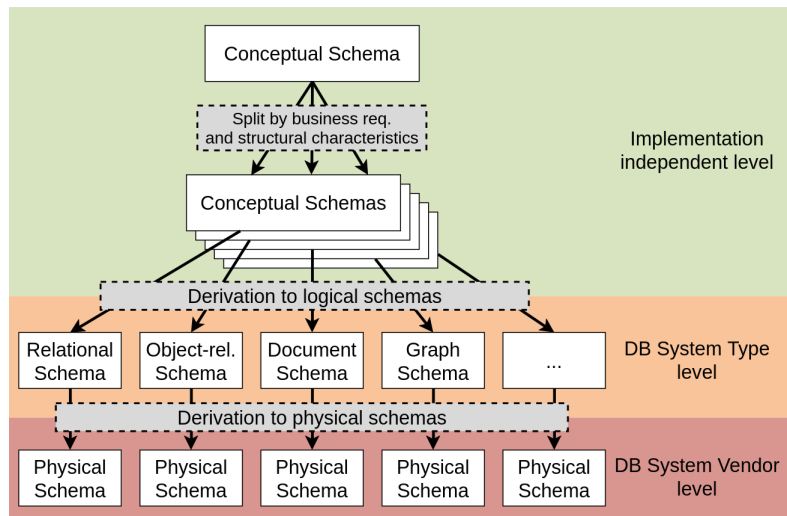


Figura 2. Etapas do processo proposto
Fonte: [Kolonko and Müllenbach 2020]

desenvolvimento de uma API centralizada para acesso a diferentes SGBDs considerados em um projeto poliglota de BDs.

O trabalho de Roy-Hubara et. al. propõe uma metodologia para a escolha de tecnologias de BD “ótimas” para um determinado projeto poliglota de BD [Roy-Hubara et al. 2022]. A metodologia apresenta 5 etapas: (i) determinação de requisitos do domínio da aplicação relacionados aos dados; (ii) determinação de requisitos funcionais do projeto; (iii) determinação dos requisitos não-funcionais do projeto; (iv) divisão da modelagem de dados em fragmentos diferentes que possuem características similares e; (v) escolha das melhores tecnologias de BDs para cada fragmento.

Em suma, os autores propõem a utilização de um método baseado em clusterização utilizando algoritmo de aprendizado de máquina (DBSCAN). O algoritmo tem como entradas as informações dadas pelo projetista do BD poliglota nas etapas (i) até (iii). Este *cluster* gerará, na etapa (iv) uma matriz de proximidade. A seguir, é calculada a *distância de Manhattan* para cada ponto desta matriz, sendo esta posteriormente normalizada entre 0 e 1, onde 0 é a maior medida de igualdade. Esta distância indica a tecnologia de BD que mais se aproxima dos requisitos (funcionais e não-funcionais) do projeto, cuja escolha é sugerida ao projetista. Uma exemplificação do método pode ser vista na Figura 3, onde, de acordo com a metodologia proposta, os BDs sugeridos seriam Relacional, para a primeira fração (RDBMS), e Documento, para a segunda fração.

Por fim, a proposta de Koupil & Holubová apresenta uma abordagem de transição de um modelo conceitual (ER ou UML) para um esquema lógico multimodelo de um SGBD através da unificação de terminologias dos modelos existentes e decomposição para os esquemas lógicos escolhidos [Koupil and Holubová 2022]. Para representação categórica estendida de dados multimodelo utilizou-se a *teoria de categorias*. A teoria de categorias é um formalismo que trata de maneira uniforme relações existentes entre diferentes estruturas matemáticas, como topologias, análise, álgebra e geometria, e ao longo do tempo vem sendo utilizada em outras áreas da Ciência [Mac Lane 2013].

O trabalho apresenta as seguintes contribuições no que diz respeito à modelagem

Distance for the fragments.

User								
	Consistency	Integrity	Flexibility	Volume	Velocity	Veracity	Query complexity	Distance
RDBMS	0	0	0.25	0.85	0.75	0	0.25	0.300
Key-value	1	1	0.75	0.1	0.25	1	0.25	0.621
Column based	1	1	0.75	0.4	0.25	1	0	0.629
Graph	1	1	0.75	0.7	0.25	1	0.75	0.779
Document	1	1	0.75	0.65	0.25	1	0.5	0.736
Rest								
	Consistency	Integrity	Flexibility	Volume	Velocity	Veracity	Query complexity	Distance
RDBMS	1	1	0.69	0.36	0.81	0.00	0.03	0.556
Key-value	0	0	0.31	0.59	0.19	1.00	0.47	0.366
Column based	0	0	0.31	0.09	0.19	1.00	0.22	0.259
Graph	0	0	0.31	0.21	0.19	1.00	0.53	0.320
Document	0	0	0.31	0.16	0.19	1.00	0.28	0.277

Figura 3. Tabela exemplificando o método proposto
Fonte: [Roy-Hubara et al. 2022]

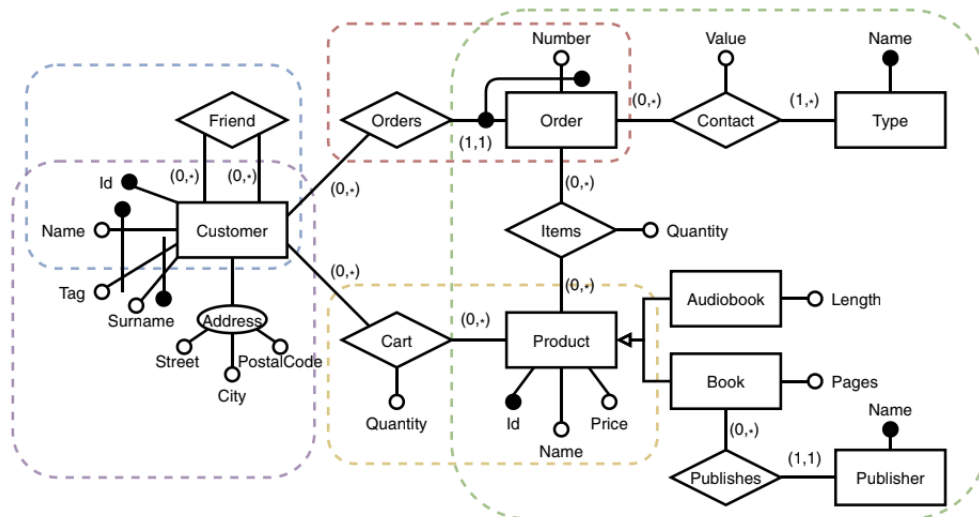


Figura 4. Marcação de partes do esquema ER para fins de mapeamento para múltiplos esquemas lógicos
Fonte: [Holubová et al. 2021]

poliglota de dados: (i) definição de uma terminologia padrão e unificada para os modelos lógicos de dados existentes; (ii) Um processo de transformação de partes de um esquema conceitual para diferentes modelos lógicos utilizando a teoria de categorias.

Destacamos aqui a contribuição (i), que sugere uma metodologia de migração de modelo de dados ER ou UML, conforme mostra a Figura 4. Para a definição dos esquemas lógicos de dados é feita uma seleção em tracejados de acordo com a cor associada a cada modelo de dados (modelo relacional - violeta, modelo de grafo - azul, modelo de documento - verde, modelo colunar - vermelho, modelo chave-valor - amarelo). O usuário é responsável pela definição manual das partes do esquema conceitual a serem mapeadas para um certo esquema lógico. Na sequência, é utilizada uma ferramenta desenvolvida pelos autores, denominada *MM-CAT*. A *MM-CAT* executa a migração do esquema conceitual para os esquemas lógicos, na qual o usuário define quais entidades e relacionamentos deseja migrar para um ou vários modelos de dados lógicos.

4. Comparativo

A Tabela 3 apresenta uma comparação entre os trabalhos relacionados, evidenciando: (i) as estratégias metodológicas adotadas para o projeto poliglota de BD; (ii) resultados gerados pela abordagem; (iii) forma de escolha das partições de entidades e relacionamentos conceituais a serem transformadas em um esquema lógico específico; bem como (iv) os requisitos considerados na transformação conceitual-lógico. Vale destacar aqui que havíamos definido um critério de comparação referente aos modelos lógicos de BD considerados pelos trabalhos em suas propostas. Entretanto, como todos os trabalhos consideram os mesmos modelos lógicos, retiramos esse critério da tabela para minimizar a redundância. Os modelos lógicos considerados pelos trabalhos são os *principais modelos NoSQL (documento, chave-valor, colunar e grafo)*, além do *modelo relacional*. Essa constatação responde a questão de pesquisa **QP-1** definida na Seção 2.

Tabela 3. Tabela de Comparação entre Trabalhos

Trabalho	Metodologia	Resultado	Forma de Escolha	Requisitos	Avaliação
Kolonko & Müllenbach	Proposta teórica	Marcações de particionamento	Manual	Domínio	-
Roy-Hubara et. al.	Aprendizado de máquina (<i>Clustering</i>)	Sugestão de modelos lógicos de BD	Semi-automática	Domínio, funcionais e não-funcionais	Qualidade da escolha dos modelos lógicos
Koupil & Holubová	Teoria de categorias	Definição de esquemas lógicos de BD	Semi-automática	Domínio	Facilidade de execução de consultas políglotas

Quanto aos requisitos, respondendo a questão de pesquisa **QP-2**, observa-se que os três estudos examinados empregam requisitos de domínio para uma definição estrutural mais adequada dos esquemas lógicos. Consideramos aqui como requisitos do domínio a estrutura e nomenclatura do esquema conceitual. Somente um trabalho considera adicionalmente requisitos funcionais e não-funcionais na definição de partições do esquema conceitual e seus enquadramentos em esquemas lógicos correspondentes.

Já com relação à questão de pesquisa **QP-3**, todos os trabalhos propõem metodologias mais ou menos detalhadas para particionar um esquema conceitual. O trabalho de Kolonko & Müllenbach apresenta uma estratégia manual onde o projetista define marcações de particionamento no esquema conceitual e particiona esse esquema em diversos sub-esquemas conceituais. Entretanto, ele não apresenta um processo completo de projeto poliglota de BD no sentido de gerar esquemas lógicos e físicos para as partições. Apenas uma sugestão de hierarquia de esquemas lógicos e físicos é mencionada. O trabalho de Roy-Hubara et. al., por sua vez, particiona (clusteriza) e sugere, de forma semiautomática, através de algoritmos de aprendizado de máquina (semiautomático), os modelos lógicos

de BD mais adequados para cada partição. Entretanto, este trabalho também não gera esquemas lógicos e físicos correspondentes. Por fim, o trabalho de Koupil & Holubová é o mais completo no sentido de gerar, também de forma semiautomática, os esquemas lógicos através de regras formais de transformação a partir do particionamento manual do esquema conceitual realizado pelo projetista.

Quanto à avaliação das metodologias propostas, Kolonko & Müllenbach não propõem métodos de avaliação. Koupil & Holubová apresentam uma validação do funcionamento da abordagem através de uma consulta formal expressa na teoria de categorias (consulta categórica), que demonstra a possibilidade dos usuários formularem consultas sem conhecer explicitamente a representação lógica dos dados nos diversos modelos lógicos. Roy-Hubara et. al. fazem avaliação da aplicabilidade dos *clusters* na determinação dos melhores modelos lógicos através da comparação destes *clusters* gerados com as escolhas realizadas por um usuário especialista.

Não foram encontrados na literatura trabalhos com processos totalmente automáticos de projeto poliglota de BD a partir de um esquema conceitual de entrada.

5. Conclusão

Este artigo apresenta um panorama das pesquisas relacionadas à modelagem poliglota de dados ou modelagem multimodelo, com o objetivo de identificar o que já foi feito, bem como lacunas e oportunidades de pesquisa. O cenário tecnológico, marcado pelo crescimento exponencial de dados heterogêneos, destaca a necessidade premente de estratégias eficientes de modelagem poliglota de BD, incluindo a melhor escolha de tecnologias.

Assim, ao apresentar e investigar trabalhos neste tópico de pesquisa, selecionados através de uma RSL, percebemos que a persistência poliglota não é apenas uma estratégia de armazenamento de dados, mas um paradigma dinâmico que oferece flexibilidade e adaptabilidade necessárias para enfrentar os desafios da crescente complexidade dos dados no cenário atual. Assim sendo, entendemos que este artigo contribui para uma fundamentação teórica neste tópico e proporciona um ponto de partida para futuras investigações e inovações.

Após análise dos trabalhos selecionados, constatamos que há uma grande lacuna no que diz respeito à proposta de metodologias de auxílio ao usuário de forma semiautomática, onde a escolha dos melhores modelos lógicos e físicos poderia ser feita de acordo com requisitos pré-estabelecidos e alinhados com esses modelos lógicos, como relacionamentos entre os dados, carga de trabalho, regras de negócio, custos e complexidade de manutenção, disponibilidade necessária dos dados, dentre outros. Entendemos que pesquisas futuras com este foco irão contribuir muito para a proposta de um projeto mais amplo de modelagem poliglota de BD, dada a dificuldade atual em escolher as melhores tecnologias de BD para determinado projeto de persistência de dados.

Referências

- Bakhtin, A. (2023). *A tool for querying multi-model data*. PhD thesis, Prague: Charles University, Faculty of Mathematics and Physics, Department of Software Engineering.
- Fowler, M. (2011). Polyglot persistence. <https://martinfowler.com/bliki/PolyglotPersistence.html>. Last access: 19 February 2024.

- Holubová, I., Vavrek, M., and Scherzinger, S. (2021). Evolution management in multi-model databases. *Data & Knowledge Engineering*, 136:101932.
- Khine, P. P. and Wang, Z. (2019). A Review of Polyglot Persistence in the Big Data World. *Information*, 10(4):141.
- Kitchenham, B. (2004). Procedures for Performing Systematic Reviews. *Keele University, UK*, 33(2004):1–26.
- Kolonko, M. and Müllenbach, S. (2020). Polyglot Persistence in Conceptual Modeling for Information Analysis. In *10th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 590–594. IEEE.
- Koupil, P. and Holubová, I. (2022). A Unified Representation and Transformation of Multi-model Data using Category Theory. *Journal of Big Data*, 9(1):61.
- Koupil, P., Hricko, S., and Holubová, I. (2022). A universal approach for multi-model schema inference. *Journal of Big Data*, 9(1):1–46.
- Koupil, P., Svoboda, M., and Holubová, I. (2021). Mm-cat: A tool for modeling and transformation of multi-model data using category theory. In *2021 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, pages 635–639.
- Lu, J. and Holubová, I. (2017). Multi-model data management: What’s new and what’s next? In *Proceeding of the 20th international conference on extended databases*.
- Mac Lane, S. (2013). *Categories for the working mathematician*, volume 5. Springer Science & Business Media.
- Niska, P. (2024). *Multi-model Database Migration*. PhD thesis, University of Helsinki, Faculty of Science.
- Roy-Hubara, N., Shoal, P., and Sturm, A. (2022). Selecting Databases for Polyglot Persistence Applications. *Data Knowl. Eng.*, 137(C).
- Sadalage, P. J. and Fowler, M. (2013). *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Pearson Education.
- Schreiner, G. A., Duarte, D., and Mello, R. d. S. (2019). When Relational-based Applications Go to NoSQL Databases: A Survey. *Information*, 10(7):241.
- Stonebraker, M. (2012). New Opportunities for NewSQL. *Communications of the ACM*, 55:10–11.
- Vavrek, M., Holubová, I., and Scherzinger, S. (2019). Mm-evolver: A multi-model evolution management tool. In *EDBT*, volume 19, pages 586–589.
- Wazlawick, R. S. (2009). *Metodologia de Pesquisa para Ciência da Computação*, volume 2. Elsevier.
- Ye, F., Sheng, X., Nedjah, N., Sun, J., and Zhang, P. (2023). A benchmark for performance evaluation of a multi-model database vs. polyglot persistence. *Journal of Database Management (JDM)*, 34(3):1–20.