Métricas para Análise de Esquemas em Banco de Dados NoSQL Orientado a Documentos

Harley Vera-Olivera[1,2], **Maristela Holanda**[1]

[1]Departamento de Ciências da Computação – Universidade de Brasília
Brasília – DF, Brazil

[2]Departamento de Ingeniería Informática – Universidad Nacional de San Antonio Abad del Cusco – Cusco, Perú.

harley.vera@unsaac.edu.pe, mholanda@unb.br

Abstract. *The data schema is crucial in software development, impacting the final product. Finding the ideal schema is challenging due to the numerous alternatives. Metrics have been proposed to determine the ideal schema but focus on specific aspects, such as query or schema evaluation, without addressing both simultaneously. This article proposes a metric that considers queries and schemas, including subschemas. Relationships and attributes are also considered, with weighting coefficients for each. The results show that the metric can identify complex schemas, assigning them higher scores, while simpler schemas receive lower scores.*

Resumo. *O esquema de dados é crucial no desenvolvimento de software, impactando o produto final. Encontrar o esquema ideal é desafiador devido às inúmeras alternativas. Métricas têm sido propostas para determinar o esquema ideal, mas focam em aspectos específicos, como a avaliação de consultas ou esquemas, sem abordar ambos simultaneamente. Este artigo propõe uma métrica que considera tanto consultas quanto esquemas. Relacionamentos e atributos também são considerados, com coeficientes de ponderação para cada um. Os resultados mostram que a métrica pode identificar esquemas complexos, atribuindo-lhes pontuações mais altas, enquanto esquemas mais simples recebem pontuações mais baixas.*

1. Introdução

O esquema de dados é uma das etapas mais críticas no desenvolvimento de software. Embora definir o esquema seja apenas uma etapa, seu impacto no produto final é mais significativo do que qualquer outra etapa [Moody 2005]. Portanto, deve-se encontrar o esquema mais apropriado para um determinado problema. No entanto, essa não é uma tarefa fácil devido às múltiplas alternativas de esquemas que podem ser produzidas para um problema dado. Cada modelo pode representar uma solução válida, mas pode ter implicações drásticas para o banco de dados. Na prática, a escolha de um esquema de dados adequado geralmente é baseada em bom senso, opinião e experiência dos projetistas.

A análise de esquemas de dados em bancos de dados NoSQL orientados a documentos enfrenta um desafio significativo devido à falta de métricas que considerem

as características de um esquema, como tipos de relacionamentos, tipos de atributos e mais de um esquema como solução a um problema. Os estudos de [Gómez et al. 2021] e [Kuszera et al. 2020] exemplificam essa lacuna, pois, embora abordem a análise de esquemas no formato JSON (*JavaScript Object Notation*), cada um deles se concentra em aspectos específicos, como as consultas ou esquemas.

Este artigo propõe quatro métricas de avaliação: *completude*, *padrão de acesso*, *custo de recuperação* e *redundância*. Diferentemente dos trabalhos relacionados, essas métricas consideram não apenas consultas e esquemas, mas também relacionamentos referenciados e aninhados, além de atributos simples e complexos. Adicionalmente, considera-se que mais de um esquema pode ser usados como uma solução para um problema, proporcionando uma visão mais completa sobre sua complexidade. Por fim, coeficientes de ponderação são aplicados aos relacionamentos e atributos, levando em consideração a real influência de cada uma dessas características na complexidade de um esquema.

O artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados ao tema de pesquisa. A Seção 3 apresenta nossa proposta da métrica de avaliação de esquemas. Em seguida, a Seção 4 mostra dois cenários de validação da métrica. Finalmente, na Seção 5, são apresentadas as conclusões do estudo.

2. Trabalhos relacionados

Diversos estudos têm sido propostos para a análise de esquemas, incluindo [Gómez et al. 2016], [Reis et al. 2018], [Mior et al. 2017], [Imam et al. 2020], [Chen et al. 2022] e [Reniers et al. 2020]. Contudo, apenas dois trabalhos destacam-se por suas propostas de métricas para a avaliação de esquemas: [Kuszera et al. 2020] e [Gómez et al. 2021]. O primeiro concentra-se no suporte do esquema para consultas previamente definidas, enquanto o segundo se dedica à análise da estrutura do esquema sem considerar consultas.

No estudo de [Kuszera et al. 2020], os esquemas são avaliados para determinar se respondem adequadamente às consultas; no entanto, os fatores de qualidade considerados não são explicitados. Para realizar essa avaliação, consultas em formato SQL(*Structured Query Language*) e estruturas de documentos em formato JSON foram representadas por grafos acíclicos direcionados. Dessa forma, são calculados caminhos, sub caminhos e caminhos indiretos nas consultas para analisar a cobertura de arestas em relação às arestas de uma coleção de esquemas fornecida. A métrica proposta atribui uma pontuação a cada esquema com base nesses cálculos. Embora o estudo se concentre na análise das consultas, o esquema não é avaliado através de fatores de qualidade e não considera esquemas com relacionamentos referenciados.

Por outro lado, o estudo de [Gómez et al. 2021] avalia o esquema por meio de fatores de qualidade como consumo de memória, redundância, e custo de navegação. Essa abordagem parte de um diagrama UML (*Unified Modeling Language*) e gera possíveis esquemas no formato JSON, cada um com uma estrutura de árvore equivalente. A métrica proposta avalia a complexidade do esquema, analisando métricas estruturais, como a presença de coleções, profundidade de aninhamento de documentos, largura de documentos, taxa de referência e redundância. No entanto, essa proposta não leva em consideração as consultas e o uso de mais de um esquema.

Enquanto [Kuszera et al. 2020] examina os esquemas com base em consultas, [Gómez et al. 2021] considera os relacionamentos referenciados ou aninhados. No entanto, nenhum desses estudos aborda ambos os aspectos simultaneamente, deixando uma lacuna na compreensão dos esquemas em questão. Além disso, o uso mais de um esquema é considerada apenas por [Kuszera et al. 2020], acrescentando uma complexidade adicional à análise que não é contemplada por [Gómez et al. 2021]. A análise de atributos simples e complexos também não é abordada por nenhum dos estudos, o que ressalta ainda mais a incompletude das métricas anteriores. Diferentemente dessas duas abordagens, consideramos o uso de mais de um esquema, analisando o conjunto de esquema como um todo. Além disso, empregamos coeficientes de ponderação para relacionamentos e atributos.

3. Métricas Proposta

O cenário em que se aplica nossa proposta são os bancos de dados NoSQL orientado a documentos. Esse cenário é composto por três conjuntos: coleções, consultas e esquemas. O conjunto de coleções inclui todas as coleções que participam do caso analisado, tanto nas consultas quanto nos esquemas. O conjunto de consultas contém as consultas previamente definidas. Finalmente, o conjunto de esquemas está composto pelos esquemas que dão suporte às consultas. Um esquema é composto, no mínimo, por uma coleção e uma coleção é composta por documentos. Os documentos podem conter atributos simples(valores atômicos) ou complexos (*arrays* ou objetos).

A métrica proposta avalia os esquemas levando em conta consultas por meio de quatro métricas: completude, padrão de acesso, custo de navegação e redundância. Para calcular essas métricas, são analisados relacionamentos referenciados e aninhados, bem como atributos simples e complexos, com a aplicação de coeficientes de ponderação para ambos. Além disso, a análise considera o conjunto completo de esquemas.

A proposta recebe como entrada um conjunto de consultas e um conjunto de esquemas. Como saída obtém-se uma avaliação para as consultas e outra para os esquemas. A métrica *completude* está relacionada à avaliação das consultas, enquanto o *padrão de acesso*, *custo de navegação* e *redundância* estão relacionados à avaliação dos esquemas. Para avaliar a *completude* é necessário avaliar as coleções existentes nas consultas e sua correspondência com as coleções e relacionamentos dos esquemas. Para avaliar o *padrão de acesso* são avaliados os relacionamentos dos esquemas. Para o *custo de navegação* são avaliados os relacionamentos e atributos e para a *redundância* as coleções dos esquemas.

As métricas devem ser usadas considerando que o *padrão de acesso*, *custo de navegação* e *redundância* identificam com pontuações mais altas os esquemas mais complexos na sua estrutura. Por outro lado, menores valores indicam esquemas mais simples. No entanto, uma pontuação menor não garante necessariamente que o esquema seja o melhor, sendo crucial também considerar a métrica de *completude*.

3.1. Definição Formal

Definem-se três conjuntos fundamentais para a análise: o conjunto de coleções $\boldsymbol{C}$, o conjunto de esquemas $\boldsymbol{E}$ e o conjunto de consultas $\boldsymbol{Q}$. O conjunto de coleções $\boldsymbol{C}$ é denotado como $\boldsymbol{C} = \{c_1, c_2, ..., c_k\}$, onde $k \geq 1$. O conjunto de esquemas $\boldsymbol{E}$, definido como $\boldsymbol{E} = \{e_1, e_2, ..., e_m\}$, onde cada esquema e_j é um subconjunto não vazio de $\boldsymbol{C}$, ou seja,

$e_j \subseteq \boldsymbol{C}$ e $|e_j| \geq 1$ para $j = 1, 2, ..., m$ com $m \geq 1$. Finalmente, o conjunto das consultas $\boldsymbol{Q} = \{q_1, q_2, ..., q_n\}$, onde cada consulta q_i é um subconjunto não vazio de $\boldsymbol{C}$, ou seja, $q_i \subseteq \boldsymbol{C}$ e $|q_i| \geq 1$ para $i = 1, 2, ..., n$ com $n \geq 1$.

3.2. Métrica Completude

Esta métrica verifica se os esquemas definidos em $\boldsymbol{E}$ atendem todas as consultas definidas em $\boldsymbol{Q}$. Assim, para cada consulta q_i é verificado se coleções que participam em q_i existem em $\boldsymbol{E}$. Também é verificado se existe uma relação entre as coleções que compõem q_i em algum esquema de $\boldsymbol{E}$. Se todas as consultas de $\boldsymbol{Q}$ são atendidas retorna 1, caso contrário retorna um valor entre 0 e 1. A *completude* é definida da seguinte forma:

$$\textbf{completude}(\boldsymbol{E}) = \frac{Q_a}{m} \tag{1}$$

Onde: Q_a, quantidade de consultas atendidas
m, quantidade de consultas definidas

$Q_a = \sum atende(\mathbf{E}, q_i), \quad \forall q_i \in \mathbf{Q}$ é usado para determinar a quantidade de consultas atendidas pelos esquemas de $\mathbf{E}$, onde.

$atende(\mathbf{E}, q_i) = existeColecao(\mathbf{E}, q_i) \times existeRelacionamento(\mathbf{E}, q_i)$

- $existeColecao(\mathbf{E}, q_i)$ verifica a existência de todas as coleções que compõem q_i em algum esquema e_i de $\mathbf{E}$. Se as coleções existem, retorna 1; caso contrario, retorna 0.
- $existeRelacionamento(\mathbf{E}, q_i)$ verifica a existência de um relacionamento (referenciado ou aninhado) entre as coleções que compõem q_i. Se existe relacionamento, retorna 1; caso contrario, retorna 0.

3.3. Métrica Padrão de Acesso

O padrão de acesso analisa como estão estruturadas as coleções (referenciadas ou aninhadas) em $\boldsymbol{E}$. Então para todos os esquemas e_i de $\boldsymbol{E}$ são contabilizados os relacionamentos devidamente ponderados. O uso de ponderações nos relacionamentos é devido a diferença de tempo de recuperação de dados entre documentos referenciados e aninhados. A ponderação é aplicada através de coeficientes de ponderação para relacionamentos referenciados ($coef_r$) e aninhados ($coef_a$). Assim, tem-se o padrão de acesso da seguinte forma:

$$\textbf{padraoAcesso}(\boldsymbol{E}) = contarRel(\boldsymbol{E}) \tag{2}$$

Onde:

$$contarRel(\mathbf{E}) = \sum contar(e_j) \quad \forall e_j \in \boldsymbol{E} \tag{3}$$

Assim $contar(e_j)$ contabiliza os relacionamentos referenciados e aninhados com as ponderações para cada e_j de $\boldsymbol{E}$, da seguinte forma:

$$contar(e_j) = \sum rel\_ref \times coef_r + \sum rel\_ani \times coef_a$$

3.4. Métrica Custo de Recuperação

Esta métrica analisa a estrutura e o tamanho das coleções, então para todos os esquemas e_j de $\boldsymbol{E}$ são contabilizados os relacionamentos considerando os coeficientes de ponderação usando a Equação 3. Adicionalmente, são contabilizadas os atributos simples e complexos considerando também coeficientes de ponderação para os atributos simples ($coef_s$) e complexos ($coef_c$). Com tudo, a equação para a obtenção do custo de recuperação é:

$$\textbf{custoRecuperacao(}\boldsymbol{E}\textbf{)} = contarRel(\boldsymbol{E}) + contarAtr(\boldsymbol{E}) \tag{4}$$

Onde: $contarRel(\boldsymbol{E})$, conta os relacionamentos do esquema $\boldsymbol{E}$ (Equação 3). $contarAtr(\boldsymbol{E}) = \sum contarAtributos(e_j) \quad \forall e_j \in \boldsymbol{E}$.

Assim, $contarAtributos(e_j)$ conta os atributos simples e complexos considerando suas ponderações para cada esquema e_i de $\boldsymbol{E}$, da seguinte forma:

$$contarAtr(e_j) = \sum attr_{simple} \times coef_s + \sum attr_{complexo} \times coef_c$$

3.5. Métrica Redundância

A redundância de dados pode ocorrer devido à duplicação de coleções em diferentes esquemas, cada coleção contendo cópias idênticas ou quase idênticas dos mesmos dados. Então, esta métrica analisa possíveis dados redundantes no conjunto de esquemas $\boldsymbol{E}$. Para cada esquema e_j de $\boldsymbol{E}$, são contabilizados as coleções repetidas. A equação para a obtenção da métrica é:

$$\textbf{redundancia(}\boldsymbol{E}\textbf{)} = \sum repetidas(\boldsymbol{E} - \{e_j\}, e_j) \quad \forall e_j \in \boldsymbol{E} \tag{5}$$

Onde: $repetidas(\boldsymbol{E} - \{e_j\}, e_j)$ conta as coleções repetidas do esquema e_j em $\boldsymbol{E}$.

4. Estudos de caso

Nesta seção, dois cenários são utilizados para aplicação e validação das métricas. O primeiro cenário é extraído de [Gómez et al. 2018] e inclui três coleções: *companies*, *departments* e *employees* assim como sete consultas definidas na Tabela 1. Com base nas três coleções, nove esquemas são gerados para avaliação, conforme mostrado na Figura 1. O segundo cenário, descrito em [Kuszera et al. 2020], é composto por sete coleções denominadas *customers*, *orders*, *reorder*, *orderlines*, *products*, *categories*, e *inventory* e sete consultas apresentadas na Tabela 2. Da mesma forma, com base nas sete coleções quatro esquemas são gerados para avaliação Figura 2. Os esquemas e consultas do primeiro e segundo cenário foram gerados nos trabalhos originais, sem alterações realizadas neste estudo.

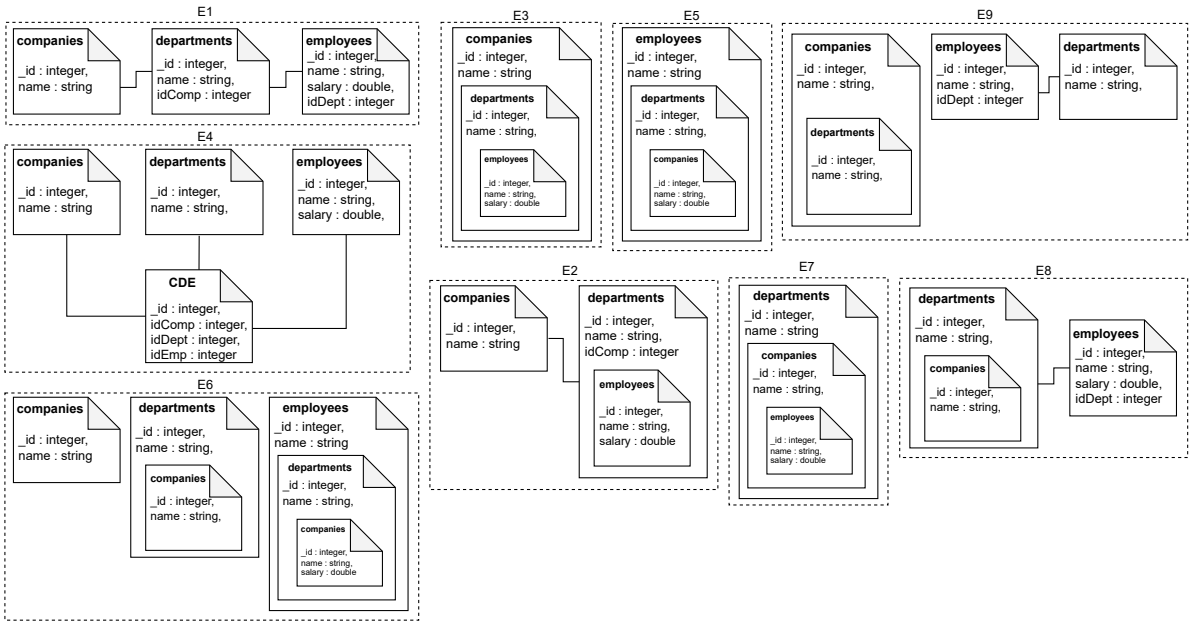

Figura 1. Esquemas para o primeiro cenário (link para os esquemas).

Tabela 1. Consultas definidas para o primeiro cenário.

No	Consulta	Coleções
q_1	*Funcionários com um salário igual a $1000*	funcionário, salário
q_2	*Funcionários com um salário superior a $1000*	funcionário, salário
q_3	funcionários com o maior salário	funcionário
q_4	*Funcionários com o maior salário por empresa e o ID da empresa*	funcionário, salário, empresa
q_5	*Funcionários com o maior salário por empresa e o nome da empresa*	funcionário, salário, empresa
q_6	*O salário mais alto*	salário
q_7	*Informações das empresas, incluindo o nome de seus departamentos*	empresa, departamentos

Tabela 2. Consultas definidas para o segundo cenário.

No	Consulta	Coleções
q_1	*Selecionar todos os dados dos clientes onde o id_cliente é igual a 1.*	cliente
q_2	*Selecionar todos os dados dos produtos juntamente com o inventário, e onde o identificador do produto é igual a 1*	produto, inventário
q_3	*Selecionar todos os dados dos pedidos juntamente com as linhas de pedido, onde o id do pedido é igual a 1*	pedido, linha_pedido
q_4	*Selecionar todos os dados dos clientes juntamente com os pedidos, linhas de pedidos e produtos, e a data do pedido está entre '2009-01-01' e '2009-01-02'*	cliente, pedido, linha_pedido, produto
q_5	*Selecionar todos os dados dos produtos juntamente com as linhas de pedidos, os pedidos e os clientes, e onde o preço do produto está entre 29 e 30*	produto, linha_pedido, pedido, cliente
q_6	*Selecionar todos os dados dos pedidos juntamente com os clientes, e as linhas de pedidos, onde a data do pedido está entre '2009-01-01' e '2009-01-02'*	pedido, cliente, linha_pedido
q_7	*Selecionar todos os dados do inventário juntamente com as linhas de pedido, onde o identificador do pedido é igual a 1*	linha_pedido inventário

Para analisar cada cenário, segue-se o seguinte processo: primeiro, são analisadas as consultas para calcular a métrica da *completude*, utilizando a Equação 1. Em seguida, é analisado os *esquemas* para calcular as métricas de *padrão de acesso*, *custo de recuperação* e *redundância*. Uma vez que as métricas que avaliam os esquemas estão baseadas

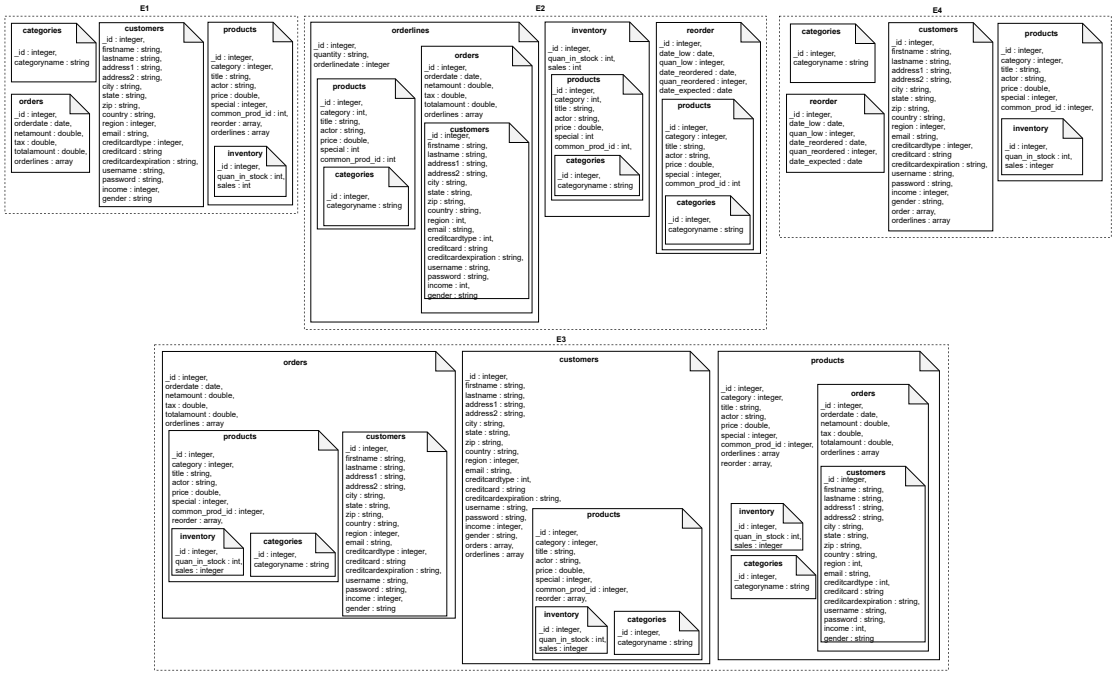

Figura 2. Esquemas para o segundo cenário (link para os esquemas).

na análise dos relacionamentos, atributos e coleções uma análise prévia é realizado para o cálculo de $contarRel(E)$, $contarAtr(E)$ e $repetidas(E - \{e_j\}, e_j)$. Nesta proposta consideramos $coef_r$ com o valor de 0.6 e o valor de 0.4 para o $coef_a$. Enquanto, os valores para os coeficientes de ponderação de atributos foram 0.51 para $coef_s$ e 0.49 para $coef_c$. Os valores dos coeficientes foram obtidos através de uma análise de tempos de resposta com regressão múltipla, aplicando o método proposto em [Vera-Olivera et al. 2023].

4.1. Análise primeiro cenário

A Tabela 3 apresenta a análise da *completude* para cada consulta nos nove esquemas do primeiro cenário. A coluna "eC" avalia a existência das coleções que participam de uma consulta q_i em pelo menos um esquema e_j de E ($existeColecao(E, q_i)$), enquanto a coluna "eR" avalia a existência de um relacionamento entre as coleções de uma consulta q_i em pelo menos um esquema e_j de E. Na terceira coluna, é avaliado $atende(\mathbf{E}, q_i) = existeColecao(\mathbf{E}, q_i) \times existeRelacionamento(\mathbf{E}, q_i)$.

Tabela 3. Análise de $existeColecao(E, q_i)$ e $existeRelacionamento(E, q_i)$ para cada consulta nos nove esquemas.

	E1			E2			E3			E4			E5			E6			E7			E8			E9		
	eC	eR		eC	eR		eC	eR		eC	eR		eC	eR		eC	eR		eC	eR		eC	eR		eC	eR	
q_1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
q_2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
q_3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
q_4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
q_5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
q_6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
q_7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

A Tabela 4 apresenta os resultados da métrica $completude$ para cada esquema.

Dos resultados pode-se observar que os esquemas E1 ao E8 atendem a todas as consultas. No entanto, o esquema E9 não atende as consultas q_4 e q_5 devido à falta de identificação de um relacionamento que suporte as coleções participantes dessas consultas em nenhum esquema e_j de E9.

Tabela 4. Resultados da métrica completude para os nove esquemas.

E1	E2	E3	E4	E5	E6	E7	E8	E9
1	1	1	1	1	1	1	1	0.7

Para o cálculo das métricas associadas a um esquema (*padrão de acesso*, *custo recuperação* e *redundância*), inicialmente, analisamos os relacionamentos, atributos e coleções de cada esquema e_j de E. A Tabela 5 apresenta a análise dos relacionamentos referenciados e aninhados em todos os esquemas. Por exemplo, E1 possui apenas um esquema, no qual há dois relacionamentos referenciados e nenhum aninhado. Usando os coeficientes de ponderação para cada relacionamento o resultado final é de 1.2. Por outro lado, E6 apresenta três esquemas, o esquema e_1 não possui nenhum relacionamento, enquanto os esquemas e_2 e e_3 apresentam um e dois relacionamentos aninhados, respectivamente. O resultado final para o E6 é a soma de todos os cálculos dos esquemas e_1, e_2, e e_3. Dos resultados, pode-se perceber que o esquema e_1 de E4 possui o maior valor devido ao uso de três relacionamentos referenciados. Por outro lado, E3, E5 e E7 apresentam os menores valores devido ao uso exclusivo de relacionamentos aninhados.

Tabela 5. Análise de relacionamentos primeiro cenário.

Soluções	Esquemas	Relacionamentos referenciados	Relacionamentos Aninhados	Cálculo com coeficientes	Total
E1	e_1	2	0	$2 \times 0.6 + 0 \times 0.4 = 1.2$	1.2
E2	e_1	1	1	$1 \times 0.6 + 1 \times 0.4 = 1$	1
E3	e_1	0	2	$0 \times 0.6 + 2 \times 0.4 = 0.8$	0.8
E4	e_1	3	0	$3 \times 0.6 + 0 \times 0.4 = 1.8$	1.8
E5	e_1	0	2	$0 \times 0.6 + 2 \times 0.4 = 0.8$	0.8
E6	e_1	0	0	$0 \times 0.6 + 0 \times 0.4 = 0$	1.2
	e_2	0	1	$0 \times 0.6 + 1 \times 0.4 = 0.4$	
	e_3	0	2	$0 \times 0.6 + 2 \times 0.4 = 0.8$	
E7	e_1	0	2	$0 \times 0.6 + 2 \times 0.4 = 0.8$	0.8
E8	e_1	1	1	$1 \times 0.6 + 1 \times 0.4 = 1$	1
E9	e_1	0	1	$0 \times 0.6 + 1 \times 0.4 = 0.4$	1
	e_2	1	0	$1 \times 0.6 + 0 \times 0.4 = 0.6$	

De forma similar, é realizada a análise dos atributos simples e complexos de todos os esquemas. A Tabela 6 apresenta os resultados, onde pode-se perceber que o conjunto de esquemas E6 tem o maior valor devido ao uso de três esquemas, gerando repetições de esquemas e, consequentemente, de atributos.

Tabela 6. Análise de atributos primeiro cenário.

Soluções	Esquemas	Atributos simples	Atributos complexos	Cálculo com coeficientes	Total
E1	e_1	9	0	$9 \times 0.51 + 0 \times 0.49 = 4.59$	4.59
E2	e_1	8	1	$8 \times 0.51 + 1 \times 0.49 = 4.57$	4.57
E3	e_1	7	2	$7 \times 0.51 + 2 \times 0.49 = 4.55$	4.55
E4	e_1	11	0	$11 \times 0.51 + 0 \times 0.49 = 5.61$	5.61
E5	e_1	7	2	$7 \times 0.51 + 2 \times 0.49 = 4.55$	4.55
E6	e_1	2	0	$2 \times 0.51 + 0 \times 0.49 = 1.02$	8.1
	e_2	4	1	$4 \times 0.51 + 1 \times 0.49 = 2.53$	
	e_3	7	2	$7 \times 0.51 + 2 \times 0.49 = 4.55$	
E7	e_1	7	2	$7 \times 0.51 + 2 \times 0.49 = 4.55$	4.55
E8	e_1	8	1	$8 \times 0.51 + 1 \times 0.49 = 4.57$	4.57
E9	e_1	4	1	$4 \times 0.51 + 1 \times 0.49 = 2.53$	5.08
	e_2	5	0	$5 \times 0.51 + 0 \times 0.49 = 2.55$	

Por último, a Tabela 7 apresenta o número de coleções repetidas em cada esquema usando a Equação 5. Os resultados mostram que E6 apresenta 3 coleções repetidas entre os três esquemas que possui.

Tabela 7. Análise de coleções repetidas em cada esquema.

E1	E2	E3	E4	E5	E6	E7	E8	E9
0	0	0	0	0	3	0	0	1

A Tabela 8 apresenta os resultados por métrica. O *padrão de acesso* é obtido da análise dos relacionamentos na Tabela 5, enquanto o *custo de recuperação* é obtido da soma dos resultados da análise dos relacionamentos e atributos (Tabelas 5 e 6). A métrica *redundância* é obtido da análise da contagem de coleções repetidas apresentada na Tabela 7. Finalmente, a coluna "Total" apresenta a soma de todos os valores das métricas. Adicionalmente, a coluna *Completude* mostra os resultados da análise do suporte às consultas em cada esquema apresentados na Tabela 4.

Dos resultados obtidos, observa-se que E6 possui o maior valor entre todos. Isso pode ser atribuído à repetição das coleções nos três esquemas, resultando em repetição de atributos. Das três métricas no E6, o *custo de recuperação* apresenta o maior valor, possivelmente devido à profundidade dos relacionamentos aninhados e ao tamanho dos documentos especialmente nos esquemas e_2 e e_3. O segundo com maior valor é o E4, destacando-se principalmente na métrica *padrão de acesso*, o que pode ser atribuído ao uso de três relacionamentos referenciados nos quatro esquemas que o compõem, além da criação de atributos adicionais num quarto esquema nomeado "CDE". Por outro lado, os menores valores são o E3, E5, e E7; essas propostas são semelhantes, apresentando apenas um esquema e relacionamento aninhados. Em relação às consultas, E9 é o único que não suporta todas as consultas, oferecendo suporte apenas a 70% delas.

Tabela 8. Resultados final das métricas no primeiro cenário.

Soluções	Padrão acesso	Custo recuperação	Redundância	Total	Completude
E1	1.2	5.79	0	6.99	1
E2	1	5.57	0	6.57	1
E3	0.8	5.35	0	6.15	1
E4	1.8	7.41	0	9.21	1
E5	0.8	5.35	0	6.15	1
E6	1.2	9.3	3	13.5	1
E7	0.8	5.35	0	6.15	1
E8	1	5.57	0	6.57	1
E9	1	6.08	1	8.08	0.7

4.2. Análise segundo cenário

De forma semelhante à análise do primeiro cenário, para o segundo cenário são analisados a *completude*, assim como os relacionamentos, atributos e coleções para as métricas *padrão de acesso*, *custo de recuperação* e *redundância*. Nesse sentido, as Tabelas 9 e 10 apresentam as análises das consultas e os resultados para a métrica *completude*. Dos resultados (Tabela 10), observa-se que nenhuma das propostas oferece suporte completo às consultas. Assim, E2 é o único a atingir um 90% de suporte às consultas, enquanto os demais alcançam apenas 30%.

Tabela 9. Análise de $existeColecao(E, q_i)$ e $existeRelacionamento(E, q_i)$ para cada consulta nos quatro esquemas.

	E1			E2			E3			E4		
	eC	eR		eC	eR		eC	eR		eC	eR	
q_1	1	1	1	1	1	1	1	1	1	1	1	1
q_2	1	1	1	1	1	1	1	1	1	1	1	1
q_3	0	0	0	1	1	1	0	0	0	0	0	0
q_4	0	0	0	1	1	1	0	0	0	0	0	0
q_5	0	0	0	1	1	1	0	0	0	0	0	0
q_6	0	0	0	1	1	1	0	0	0	0	0	0
q_7	0	0	0	1	0	0	0	0	0	0	0	0

Tabela 10. Resultados da métrica completude para os quatro esquemas.

E1	E2	E3	E4
0.3	0.9	0.3	0.3

Na análise dos relacionamentos, os resultados estão apresentados na Tabela 11. Observa-se que nenhuma das propostas apresenta relacionamentos referenciados. O maior valor enquanto a relacionamentos é da proposta E3. Em relação aos atributos, a análise é apresentada na Tabela 12 para as quatro propostas. Observa-se que o segundo

cenário apresenta uma maior quantidade de atributos, sendo os atributos simples predominantes em relação aos complexos. Notavelmente, E3 se destaca ao apresentar a maior quantidade de atributos simples e complexos entre todas as propostas analisadas. Em relação às coleções repetidas, a Tabela 13 apresenta a contagem em todas as propostas. Destaca-se que E2 possui quatro coleções repetidas, enquanto E3 apresenta nove.

Tabela 11. Análise de relacionamentos segundo cenário.

Soluções	Esquemas	Relacionamentos referenciados	Relacionamentos aninhados	Cálculo com coeficientes	Total
E1	e_1	0	0	0 x 0.6 + 0 x 0.4 = 0	0.4
	e_2	0	0	0 x 0.6 + 0 x 0.4 = 0	
	e_3	0	1	0 x 0.6 + 1 x 0.4 = 0.4	
	e_4	0	0	0 x 0.6 + 0 x 0.4 = 0	
E2	e_1	0	4	0 x 0.6 + 4 x 0.4 = 1.6	3.2
	e_2	0	2	0 x 0.6 + 2 x 0.4 = 0.8	
	e_3	0	2	0 x 0.6 + 2 x 0.4 = 0.8	
E3	e_1	0	4	0 x 0.6 + 4 x 0.4 = 1.6	4.4
	e_2	0	3	0 x 0.6 + 3 x 0.4 = 1.2	
	e_3	0	4	0 x 0.6 + 4 x 0.4 = 1.6	
E4	e_1	0	0	0 x 0.6 + 0 x 0.4 = 0	0.4
	e_2	0	0	0 x 0.6 + 0 x 0.4 = 0	
	e_3	0	1	0 x 0.6 + 1 x 0.4 = 0.4	
	e_4	0	0	0 x 0.6 + 0 x 0.4 = 0	

Tabela 12. Análise de atributos segundo cenário.

Soluções	Esquemas	Atributos simples	Atributos complexos	Cálculo com coeficientes	Total
E1	e_1	6	1	$6 \times 0.51 + 1 \times 0.49 = 3.55$	20.32
	e_2	2	0	$2 \times 0.51 + 0 \times 0.49 = 1.02$	
	e_3	10	3	$10 \times 0.51 + 3 \times 0.49 = 6.57$	
	e_4	18	0	$18 \times 0.51 + 0 \times 0.49 = 9.18$	
E2	e_1	38	0	$38 \times 0.51 + 0 \times 0.49 = 19.38$	33.15
	e_2	12	0	$12 \times 0.51 + 0 \times 0.49 = 6.12$	
	e_3	15	0	$15 \times 0.51 + 0 \times 0.49 = 7.65$	
E3	e_1	36	5	$36 \times 0.51 + 5 \times 0.49 = 20.81$	59.69
	e_2	30	5	$30 \times 0.51 + 5 \times 0.49 = 17.75$	
	e_3	36	6	$36 \times 0.51 + 6 \times 0.49 = 21.13$	
E4	e_1	2	0	$2 \times 0.51 + 0 \times 0.49 = 1.02$	19.38
	e_2	18	2	$18 \times 0.51 + 2 \times 0.49 = 10.16$	
	e_3	10	0	$10 \times 0.51 + 0 \times 0.49 = 5.1$	
	e_4	6	0	$6 \times 0.51 + 0 \times 0.49 = 3.06$	

Tabela 13. Análise de coleções repetidas em cada esquema.

E1	E2	E3	E4
0	4	9	0

O resultado final por métricas é apresentado na Tabela 14. Observa-se que a proposta E3 apresenta os maiores valores em todas as métricas, enquanto apenas 30% das consultas são atendidas. Por outro lado, a proposta com o menor valor é o E4; entretanto, esta proposta também atende apenas 30% das consultas. Entre as quatro propostas, uma possibilidade seria selecionar E2 como o mais adequado devido à sua segunda menor pontuação no esquema e ao atendimento de 90% das consultas. No entanto, uma análise adicional seria necessária para reduzir a redundância e garantir que todas as consultas sejam atendidas. Em geral, para selecionar a proposta mais adequada utilizando a métrica proposta, deve-se considerar pontuações baixas, verificando também o atendimento às consultas no fator de *completude*.

Tabela 14. Resultados final das métricas no segundo cenário.

Soluções	Padrão acesso	Custo recuperação	Redundância	Total	Completude
E1	0.4	20.72	0	21.12	0.3
E2	3.2	36.35	4	43.55	0.9
E3	4.4	64.09	9	77.49	0.3
E4	0.4	19.78	0	20.18	0.3

Em relação aos trabalhos relacionados, o estudo de [Gómez et al. 2021] identifica a E9 como o melhor, nossa proposta identifica a E3, E5 e E7 como as três melhores, com igual pontuação. Em nossa análise, E9 é classificado como o segundo melhor. Essa diferença pode ser explicada pelo fato de que, no estudo de [Gómez et al. 2021], não são considerados o fator da *completude* e *redundância*, aspectos em que E9 apresenta diferenças. Por outro lado, no segundo cenário o estudo de [Kuszera et al. 2020] identifica como a melhor proposta a E3, enquanto nossa proposta identifica ao E2 (embora tenha a segunda menor pontuação porém 90% de *completude*). O E3, avaliado com nossa métrica, apresenta o maior pontuação e apenas 30% de *completude.* Deve-se considerar que o foco do estudo de [Kuszera et al. 2020] foram as consultas, não os esquemas.

5. Conclusões

Neste trabalho, foi proposto uma métrica para avaliar consultas e esquemas de dados na etapa do desenho lógico. Enquanto a métrica *completude* avalia se todas as consultas definidas são atendidas ou não pelos esquemas, as métricas *padrão de acesso*, *custo recuperação* e *redundância* avaliam os esquemas de dados. As métricas que avaliam os esquemas estão baseadas na análise dos relacionamentos, atributos e coleções. Considerando que escolher entre um relacionamento referenciado ou aninhado tem impacto nos esquemas, coeficientes de ponderação são usados para refletir a complexidade real dos esquemas. Da mesma forma, coeficientes de ponderação são aplicados aos atributos simples e complexos.

Em nossa métrica, os maiores valores estão associados aos esquemas menos adequados, enquanto os menores valores estão associados a esquemas mais simples. No entanto, uma pontuação menor pode não ser necessariamente indicativa do melhor esquema, pois é essencial verificar a métrica de *completude*. Assim, nossa métrica pode fornecer uma indicação mais completa, em comparação com a literatura, sobre qual proposta de esquemas poderia ser o mais adequado para um determinado problema.

Como trabalhos futuros, pretendemos aplicar as métricas em mais cenários de validação e incluir coeficientes de ponderação para as cardinalidades. Além disso, a métrica proposta pode viabilizar uma avaliação automatizada dos esquemas por meio de algoritmos heurísticos para determinar os esquemas mais adequado para um determinado problema.

Referências

Chen, L., Davoudian, A., and Liu, M. (2022). A workload-driven method for designing aggregate-oriented nosql databases. *Data & Knowledge Engineering*, 142:102089.

Gómez, P., Casallas, R., and Roncancio, C. (2016). Data schema does matter, even in nosql systems! In *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*, pages 1–6. IEEE.

Gómez, P., Roncancio, C., and Casallas, R. (2018). Towards quality analysis for document oriented bases. In *International Conference on Conceptual Modeling*, pages 200–216. Springer.

Gómez, P., Roncancio, C., and Casallas, R. (2021). Analysis and evaluation of document-oriented structures. *Data & Knowledge Engineering*, 134:101893.

Imam, A. A., Basri, S., Ahmad, R., Wahab, A. A., González-Aparicio, M. T., Capretz, L. F., Alazzawi, A. K., and Balogun, A. O. (2020). Dsp: Schema design for non-relational applications. *Symmetry*, 12(11):1799.

Kuszera, E. M., Peres, L. M., and Didonet Del Fabro, M. (2020). Query-based metrics for evaluating and comparing document schemas. In *International Conference on Advanced Information Systems Engineering*, pages 530–545. Springer.

Mior, M. J., Salem, K., Aboulnaga, A., and Liu, R. (2017). Nose: Schema design for nosql applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2275–2289.

Moody, D. L. (2005). Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. *Data & Knowledge Engineering*, 55(3):243–276.

Reis, D. G., Gasparoni, F. S., Holanda, M., Victorino, M., Ladeira, M., and Ribeiro, E. O. (2018). An evaluation of data model for nosql document-based databases. In *World Conference on Information Systems and Technologies*, pages 616–625. Springer.

Reniers, V., Van Landuyt, D., Rafique, A., and Joosen, W. (2020). A workload-driven document database schema recommender (dbsr). In *International Conference on Conceptual Modeling*, pages 471–484. Springer.

Vera-Olivera, H., Alvarez-Mamani, E., and Holanda, M. (2023). Análise de desempenho em banco de dados nosql orientado a documentos: Um Índice para comparação de modelos de dados. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 26–38, Porto Alegre, RS, Brasil. SBC.