A Survey on the State of Practice in the Adoption of Schema Matching in Brazilian Companies

Ricardo H. D. Borges
Universidade Federal de Goiás - Instituto de Informática
Goiânia, Goiás, Brasil
ricardoborges@ufg.br

Leonardo Andrade Ribeiro
Universidade Federal de Goiás - Instituto de Informática
Goiânia, Goiás, Brasil
laribeiro@inf.ufg.br

Valdemar V. Graciano Neto
Universidade Federal de Goiás - Instituto de Informática
Goiânia, Goiás, Brasil
valdemarneto@ufg.br

ABSTRACT

Context: Data integration is a fundamental activity for achieving full interoperability between Information Systems (IS). One of the most critical steps in this process is schema matching, which involves identifying correspondences between schema elements to ensure that equivalent information is correctly associated.
Problem: The schema matching activity can be extremely time-consuming, expensive, repetitive, labor-intensive, and error-prone, potentially leading to significant losses. In this context, companies can employ tools to support or partially/fully automate this activity. However, the current state of such practices in Brazil remains unclear.
Solution: Understanding the state of practice in Brazil can help guide research and development efforts by revealing how data integration is currently conducted in companies. This provides evidence of practical approaches and informs the structuring of training programs and the development of technologies to support such initiatives.
IS Theory: We rely on Socio-technical systems theory, since we consider social and technical aspects as interdependent parts of a complex system.
Method: A survey research was carried out.
Summarization of Results: Thirty-five data professionals from 12 states and the Federal District participated in the survey. The results show that, despite familiarity with the concept, practical adoption faces challenges such as schema complexity, a lack of tools, and the absence of specific training. Nevertheless, for those who apply the technique, the benefits are clear: (i) more efficient data integration, (ii) error reduction, and (iii) increased agility.
Contributions and Impact on the IS Area: The results contribute by providing evidence that can help guide research efforts and technological advancements to support schema matching in Brazil. This aims to enable full interoperability among information systems in the future, addressing Brazil's Grand Challenges in Information Systems, such as Systems-of-Systems, Smart Cities, and Full Interoperability, which serves as an enabler for the other two.

CCS CONCEPTS

• **Theory of computation** → **Data integration**; • **Information systems** → *Entity resolution.*

1 INTRODUÇÃO

A interoperabilidade entre Sistemas de Informação (SI) é uma necessidade premente, em particular com a iminente ascensão dos Sistemas-de-Sistemas e Cidades Inteligentes, que combinam múltiplos sistemas e demandam interoperabilidade plena e instantânea entre seus constituintes [4, 5, 8, 9, 17, 19]. *Schema matching*[1] é uma etapa inerente ao processo de estabelecimento de links de interoperabilidade, visto que a viabilização da comunicação entre sistemas exige que seus respectivos esquemas possuam compatibilidade semântica. Assim, é necessário, nesse processo, que sejam estabelecidas correspondências entre elementos de dois esquemas para garantir que informações equivalentes sejam corretamente associadas entre si [21]. Todavia, integrar dados pode ser um desafio, principalmente quando esses dados são provenientes de fontes que foram projetadas de maneira independente umas das outras. Nesse contexto, os esquemas, isto é, as descrições lógicas dos dados, frequentemente divergem, mesmo entre fontes associadas a um mesmo domínio de aplicação [12].

A literatura relata várias formas de integrar dados, geralmente envolvendo um processo complexo composto por várias etapas [3]. Como já indicado, a etapa de *schema matching* demanda a comparação de elementos de esquema como tabelas, colunas e tipos de dados em bancos de dados, para identificar correspondências semânticas entre eles. Uma vez identificadas as correspondências entre os esquemas, a próxima etapa é o *schema mapping* [11], que produz especificações descrevendo logicamente os relacionamentos entre elementos esquemáticos e seus valores correspondentes; tais especificações capturam a informação necessária para a geração de artefatos físicos que irão finalmente realizar a transformação de dados em um processo de integração.

O foco deste artigo é a etapa de *schema matching*. Essa etapa pode ser um gargalo em processos de integração de dados, mesmo para esquemas de tamanho e complexidade moderados. A resolução eficiente do *schema matching* tem motivado intensa pesquisa ao longo das décadas e uma variedade de técnicas foram propostas para identificação de correspondências [7, 21, 24]. Infelizmente, tais esforços científicos aparentemente ainda não produziram um impacto tecnológico escalado, tendo em vista a perceptível escassez de produtos maduros para *schema matching*. Por exemplo, enquanto existem ferramentas gratuitas e/ou de código aberto para especificação de mapeamentos entre esquemas, como Pentaho PDI[2] e Apache Hop[3], não se tem conhecimento de contrapartes com essas características para *schema matching*. Por esse motivo, é razoável supor que a realização completamente manual de *schema matching* ainda é uma prática comum, principalmente em projetos com orçamento limitado.

[1] O termo em inglês foi mantido por ser mais aderente ao jargão da área para denotar 'casamento dos esquemas de dados. O mesmo se aplica para *schema mapping*, visto à frente.
[2] https://www.hitachivantara.com
[3] https://hop.apache.org

Neste contexto, este artigo apresenta um estudo sobre o estado da prática em *schema matching* no cenário brasileiro. Para isso, foi conduzida uma pesquisa de opinião com a participação de 35 profissionais de 12 Estados e do Distrito Federal. As questões de pesquisa abordaram diversos aspectos sobre a prática de *schema matching*, incluindo contexto e complexidade do problema enfrentado, nível de esforço dedicado, abordagens predominantes, familiaridade e percepção dos participantes sobre o tópico, entre outros. Os resultados obtidos permitem delinear um panorama sobre o emprego de *schema matching* em processos de integração de dados nas instituições brasileiras e indicam oportunidades de pesquisa com elevado potencial de impacto prático.

O restante deste artigo está organizado como segue. A Seção 2 apresenta um breve referencial teórico e discute trabalhos relacionados. O método adotado na elaboração da pesquisa de opinião é descrito na Seção 3 e os resultados são reportados e discutidos na Seção 4. Possíveis ameaças à validade deste trabalho são apresentadas na Seção 5. Finalmente, a Seção 6 apresenta as conclusões.

2 FUNDAMENTOS E TRABALHOS RELACIONADOS

Existem três abordagens gerais para integração de dados: materialização, federação e busca [10]. Materialização realiza a consolidação e armazenamento das fontes de dados em um único banco de dados; esta é a abordagem usada em *data warehouses*. Federação constrói uma representação virtual dos dados integrados, que serão então acessados por alguma forma de mediação. Busca envolve a criação de um índice global sobre as fontes de dados. Estas abordagens podem ser combinadas em diferentes maneiras em arquiteturas como *data lakes* e *lakehouses* [13].

Schema matching é uma etapa necessária em soluções baseadas em materialização e federação. Ele consiste na identificação de correspondências semânticas entre elementos de dois esquemas. *Schema matching* é um problema difícil. Considerando-se apenas casamentos um-para-um, isto é, aqueles em que busca-se identificar correspondências envolvendo apenas um elemento de cada esquema, o espaço de comparação é quadrático no número de elementos. Se mais de um operador de comparação [4] for empregado, então a quantidade de comparações a serem realizadas aumentará ainda mais. Além disso, casamentos um-para-um são insuficientes quando as correspondências podem ocorrer entre composições de elementos esquemáticos. Por exemplo, o atributo `nome_completo` em um esquema pode corresponder à composição de `nome` e `sobrenome` em outra. Nesta situação, casamentos N:N (muitos-para-muitos) são necessários, o que expande substancialmente o espaço de comparações, podendo, inclusive, torná-lo ilimitado [7]. Com isso, devido à sua inerente complexidade, *schema matching*, assim como integração de dados de maneira geral, não pode ser plenamente resolvido por métodos completamente automatizados e a participação humana torna-se imprescindível [15]. Finalmente, *schema matching* é frequentemente uma tarefa repetitiva: novas identificações de correspondências devem ser feitas sempre que novas fontes de dados são adicionadas ou quando o esquema das fontes existentes é modificado.

[4]Tais operadores de comparação são referenciados como *Matchers* [21].

No que tange a **trabalhos correlatos**, os autores desconhecem outras iniciativas com a mesma proposta deste artigo. O trabalho mais próximo encontrado foi o de Atmatzides, Bedo e Oliveira (2022) [1], que apresenta um levantamento realizado com empresas brasileiras sobre a adoção de SGBDs NoSQL em suas atividades. No entanto, quanto ao estado da prática em Integração de Dados e, em particular, *schema matching*, não se sabe de outras iniciativas semelhantes publicadas na literatura científica.

3 MÉTODO DE PESQUISA

A pesquisa de opinião (do inglês *survey research*) é uma técnica de coleta de dados que visa obter informações de uma amostra representativa de uma população por meio de questionários, entrevistas ou outras formas de interação estruturada [23]. Realizar uma pesquisa de opinião é essencial para contextualizar a pesquisa no ambiente prático e obter *insights* [18]. No contexto deste trabalho, a diversidade de experiências na indústria é vasta, com diferentes setores apresentando nuances específicas nas suas necessidades de integração de dados, o que reflete na prática de *schema matching* adotada. Uma pesquisa de opinião permite capturar essa diversidade, garantindo uma análise que considera diferentes contextos e práticas. Essa característica é instrumental para a identificação de tendências e padrões emergentes no uso de *schema matching*.

Do ponto de vista de **caracterização da pesquisa**, este trabalho adotou uma abordagem quantitativa de análise dos dados sob a perspectiva epistemológica positivista, com método de pesquisa de pesquisa de opinião. O processo de condução da pesquisa de opinião adotado neste trabalho foi inspirado nas diretrizes propostas por Kasunic et al., Molléri et al. e Linaker et al. [14, 16, 18], com finalidade descritiva, usando as técnicas de coleta de dados de formulário para realizar a extração e síntese dos dados coletados e a análise desses por meio de estatística descritiva.

O método utilizado foi estruturado em oito etapas: (i) definir objetivo da pesquisa, (ii) definir a justificativa da pesquisa, (iii) definir questões de pesquisa, (iv) definir público-alvo da pesquisa, (v) elaborar questionário, (vi) teste piloto do questionário, (vii) distribuir questionário, e (viii) analisar resultados. Seguindo as etapas acima, cada um dos elementos do protocolo foram definidos, como segue.
Objetivo da Pesquisa. O objetivo da pesquisa é *investigar o estado da prática em integração de dados, e em particular, schema matching, na indústria brasileira.*

A pesquisa justifica-se pela necessidade de identificar as abordagens e técnicas usadas para realização de integração de dados em organizações, com foco principal na etapa de *schema matching*. Esta etapa é frequentemente um gargalo em projetos de integração e não dispõe de um repertório consolidado de ferramentas e metodologias para agilizar e sistematizar a sua execução.

A partir do objetivo, as **Questões de Pesquisa** derivadas foram:

(QP1) Quais são as abordagens predominantes que as empresas utilizam atualmente para *schema matching*?
(QP2) Quais são os benefícios que as empresas percebem na abordagem que elas utilizam para o *schema matching*?
(QP3) Quais são os principais desafios ou dificuldades enfrentados ao realizar *schema matching*?
(QP4) Qual é o grau de esforço desempenhado pelas empresas na integração de dados?

(QP5) Qual é o nível de complexidade das integrações de dados com que as empresas estão lidando atualmente?
(QP6) Qual é o processo de *schema matching* que as empresas estão empregando?
(QP7) Qual é o nível de satisfação das empresas ao utilizar o *schema matching*?

Justificativa. As questões de pesquisa buscaram investigar a experiência dos participantes em projetos de integração de dados, explorando o contexto, a complexidade das fontes, esforço investido e estratégias adotadas. Com foco em *schema matching*, essas questões buscaram identificar as ferramentas, desafios, satisfação e planos futuros para realização de atividades de integração de dados nas empresas brasileiras.

Público-alvo da pesquisa. A audiência alvo deste levantamento englobou profissionais especializados em diversas áreas, incluindo desenvolvedores de software, analistas de dados, engenheiros de dados, administradores de banco de dados, cientistas de dados, pesquisadores, professores e arquitetos de soluções.

Elaboração do questionário. Durante a fase inicial, empreendeu-se esforços para desenvolver perguntas que atendessem efetivamente aos objetivos delineados pelas questões de pesquisa. Este processo não se limitou apenas à formulação de questões, mas também incorporou a validação por especialistas em integração de dados, garantindo que as perguntas fossem tecnicamente precisas e alinhadas com as práticas da área.

Questões Éticas. Especial atenção foi dedicada à elaboração do Termo de Consentimento Livre e Esclarecido (TCLE). Este documento foi concebido com o propósito de assegurar a compreensão plena e informada por parte dos participantes, destacando seus direitos e delineando claramente os objetivos e procedimentos da pesquisa.

Table 1: Leiaute do questionário.

Partes	Conteúdo
Parte 1	Descrições dos termos utilizados.
	Termo de Consentimento Livre e Esclarecido.
	Opção de concordar com os termos.
Parte 2	Questionário demográfico fechado.
Parte 3	Questionário de estudo fechado.

A estrutura do questionário foi concebida conforme apresentado na Tabela 1. Foram elaboradas questões demográficas (QD) e questões de estudo (QE). O propósito das QDs é caracterizar o perfil dos participantes, enquanto as QEs solicitam aos(às) respondentes informações específicas para responder às questões de pesquisa elaboradas. O questionário foi elaborado com questões fechadas (QF). O intuito foi padronizar as respostas e facilitar a leitura e síntese das informações coletadas. Nas Tabelas 2 e 3, são apresentadas as questões de cada etapa do questionário, respectivamente das Partes 2 e 3, conforme apresentado na Tabela 1.

Table 2: Questões demográficas fechadas (QD).

ID	Questões de Caracterização
QD1	Qual sua faixa etária?
QD2	Qual é o seu nível de escolaridade?
QD3	Em qual unidade federativa você reside?
QD4	Qual é o seu sexo?
QD5	Quanto tempo de experiência profissional você tem no mercado de TI?
QD6	Qual é a sua principal ocupação no mercado de TI?
QD7	Quanto tempo de experiência profissional você tem na integração de dados?
QD8	Qual é o seu nível de familiaridade com o conceito de *schema matching*?

Na Tabela 4, é apresentada a relação entre as questões de pesquisa e as perguntas do questionário. Cada questão do questionário está associada a pelo menos uma questão de pesquisa, evidenciando a correspondência entre os objetivos da pesquisa e as indagações direcionadas aos participantes. Cada célula da tabela representa uma ponte entre a coleta de dados empíricos e os objetivos mais amplos da pesquisa. Ao examinar a tabela, torna-se evidente como as respostas obtidas a partir do questionário alimentam diretamente os elementos fundamentais que compõem as questões de pesquisa.
Teste piloto do questionário. Um teste piloto foi realizado para garantir que o questionário era claro e versava sobre os pontos mais importantes da prática de *schema matching*. Neste sentido, o questionário foi submetido à apreciação de profissionais da área provenientes de empresas. Adicionalmente, reconhecendo a sensibilidade de algumas questões pessoais contidas no questionário, foi realizada uma etapa específica de validação em colaboração com especialistas em anonimização de formulários. Esse processo teve como objetivo garantir que as informações de cunho pessoal dos respondentes fossem devidamente protegidas e que o questionário respeitasse as diretrizes éticas necessárias para a pesquisa.

Distribuição do questionário. A aplicação do questionário à amostra selecionada exigiu atenção cuidadosa para garantir que os participantes compreendam plenamente as perguntas e possam fornecer respostas de maneira adequada. Este estágio não apenas representou a interação direta com os respondentes, mas também influenciou diretamente a qualidade e a confiabilidade dos dados obtidos. Especial atenção foi dedicada à elaboração de instruções claras e simples, garantindo que a linguagem utilizada seja acessível e alinhada ao público-alvo.

Coleta de Dados. O questionário da pesquisa foi divulgado por meio de um link público no Google Forms, amplamente compartilhado em diferentes canais, como listas de e-mail, LinkedIn[5] e convites diretos. O formulário permaneceu ativo durante o período compreendido entre 21 de novembro de 2023 e 22 de março de 2024.
Análise dos resultados. A análise dos resultados foi realizada de modo quantitativo. Os resultados são discutidos na próxima seção.

[5]http://linkedin.com/

Table 3: Questões de estudo (QE).

ID	Questões de estudo
QE1	Quantos projetos de integração de dados você já participou?
QE2	Em qual contexto sua empresa realiza integração de dados?
QE3	Qual a complexidade (quantidade de atributos envolvidos) da fonte de dados com a qual você já trabalhou em uma atividade de integração de dados?
QE4	Qual a quantidade de pessoas envolvidas em projetos de integração na sua empresa?
QE5	Qual o esforço desempenhado em horas semanais dedicadas em projetos de integração na sua empresa?
QE6	Quais tipos de dados ou esquemas sua empresa precisa integrar com mais frequência?
QE7	Quais fatores você acredita que contribuem para o aumento do esforço em projetos de integração?
QE8	Qual estratégias sua empresa adota para gerenciar a complexidade na integração de dados?
QE9	Qual é o impacto da complexidade da etapa de integrações de dados no tempo para concluir um projeto?
QE10	Sua empresa atualmente está utilizando *schema matching* automático ou semi automático?
QE11	Qual o principal objetivo ou caso de uso para o *schema matching* na integração de sistemas em sua organização?
QE12	Como você avaliaria o desempenho do *schema matching* em sua organização em uma escala de 1 a 5, sendo 1 muito ineficaz e 5 muito eficaz?
QE13	Quais ferramentas ou tecnologias de *schema matching* sua organização utiliza atualmente?
QE14	Quais são os principais desafios que sua organização enfrentou ao implementar o uso do *schema matching*?
QE15	Em sua opinião, quais são os principais benefícios do uso de *schema matching* automático ou semi automático?
QE16	Quais são as principais razões para a não utilização do *schema matching*?
QE17	Quais ferramentas de *schema matching* você utilizou em projetos de integração de dados?
QE18	Como sua empresa lida com conflitos de correspondência ao usar *schema matching*?
QE19	Sua empresa tem estratégias ou planos para melhorar o uso de *schema matching* no futuro? Se sim, quais são esses planos?
QE20	Quais práticas sua empresa adota para documentar e monitorar as correspondências de esquemas ao longo do tempo?
QE21	Sua empresa realiza alguma validação ou verificação posterior ao processo de *schema matching* para garantir a precisão e a integridade dos dados?
QE22	Em uma escala de 1 a 5, onde 1 representa "totalmente insatisfeito" e 5 representa "totalmente satisfeito", quão satisfeito você está com o processo de *schema matching* em sua empresa?
QE23	Quais ferramentas de Schema Mapping (Transformação) você utiliza na sua empresa?

Table 4: Relação das questões de pesquisa com as questões do formulário.

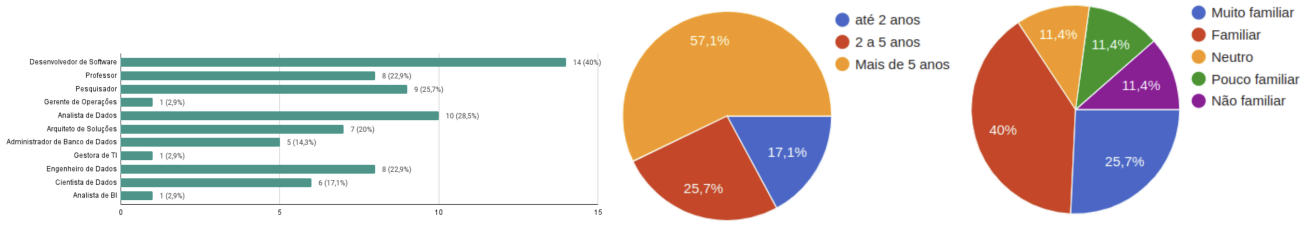
ID	Questões de pesquisa	Questão do formulário
QP1	Quais são as abordagens predominantes que as empresas utilizam atualmente para *schema matching*?	QE10, QE11, QE13, QE17
QP2	Quais são os benefícios que as empresas percebem na abordagem que elas utilizam para o *schema matching*?	QE12, QE15
QP3	Quais são os principais desafios ou dificuldades enfrentados ao realizar *schema matching*?	QE3, QE12, QE14, QE16
QP4	Qual é o grau de esforço desempenhado pelas empresas na integrações de dados?	QE1, QE6, QE18, QE19
QP5	Qual é o nível de complexidade das integrações de dados com que as empresas estão atualmente lidando?	QE2, QE3, QE4, QE5, QE7
QP6	Qual é o processo de *schema matching* que as empresas estão empregando?	QE8, QE9, QE23
QP7	Qual é o nível de satisfação das empresas ao utilizar *schema matching*?	QE20, QE21, QE22

4 RESULTADOS

Esta seção apresenta os resultados obtidos com a pesquisa de opinião realizada.

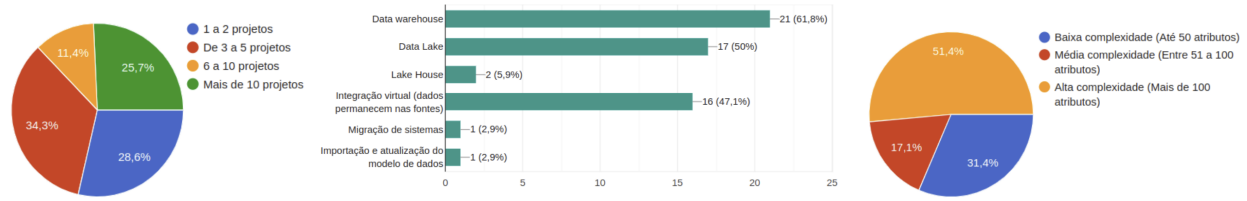
4.1 Resultados das Questões Demográficas

Trinta e cinco (35) profissionais de 12 Estados mais o Distrito Federal responderam a esta pesquisa. Não foi possível precisar a taxa exata de respondentes em relação aos convites enviados, mas a



(a) Dados sobre ocupação no mercado de TI (b) Tempo de experiência profissional. (c) Familiaridade com *schema matching*.

Figure 1: Resultados demográficos.



(a) Quantidade projetos de integração de dados. (b) Contexto da integração de dados. (c) Complexidade da fonte de dados.

Figure 2: Quantidade de projetos, contexto e complexidade das fontes de dados.

amostra parece razoável dentro do universo de profissionais de dados existentes no Brasil. Salienta-se que outros estudos publicados anteriormente no SBSI apresentam números semelhantes [2, 6, 20].

Doze unidades da federação além do Distrito Federal foram representados nesta pesquisa. Houve participantes de todos os Estados do Sul (Rio Grande do Sul, Paraná e Santa Catarina), de alguns Estados do Sudeste (São Paulo, Rio de Janeiro e Minas Gerais), Centro-Oeste (Goiás e Distrito Federal), e Nordeste (Bahia, Pernambuco, Paraíba, Maranhão e Rio Grande do Norte). Infelizmente, não houve respondentes da Região Norte.

A Figura 1 apresenta os principais resultados das questões demográficas; foi permitido aos participantes a escolha de mais de uma opção. Observa-se que mais da metade dos respondentes atuam diretamente com atividades fortemente relacionadas à análise de dados, como analistas e engenheiro de dados (Figura 1(a)). Uma aspecto interessante dos resultados é o fato de que 40% dos respondentes classificam-se como desenvolvedores de software, o que sugere que integração de dados insere-se no contexto de desenvolvimento de sistemas de maneira geral. Em relação à experiência profissional, a maior parte dos respondentes possui mais de 5 anos de experiência em integração de dados (Figura 1(b)). Finalmente, 65.7% do público da pesquisa possui familiaridade com *schema matching* (Figura 1(c)).

4.2 Resultados das Questões de Estudo

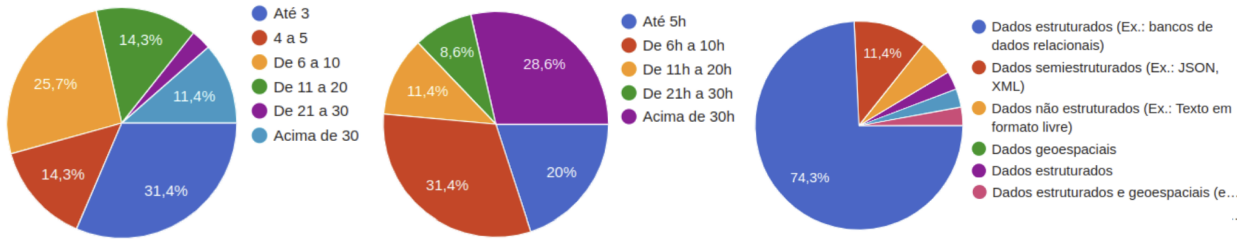
A seguir, são discutidos os resultados das questões apresentadas na Tabela 3; assim como nas questões demográficas, foi permitida a escolha de múltiplas opções em algumas questões. Devido a restrições de espaço, serão apresentadas figuras para ilustrar os resultados em somente parte das questões.

(QE1) De quantos projetos de integração de dados você já participou? A Figura 2(a) mostra que a maioria, cerca de 34,3% (12) dos respondentes participou de 3 a 5 projetos de integração de dados, indicando um nível de experiência moderado a alto, enquanto apenas 25,7% (9) estiveram envolvidos em mais de 10 projetos, sugerindo que profissionais com extensa experiência são menos comuns.

(QE2) Em qual contexto sua empresa realiza integração de dados? A maioria das empresas, cerca de 61,8% realiza integração de dados em contextos de *Data Warehouses* (Figura 2(b)). Além disso, metade dos respondentes também utilizam *Data Lakes*, destacando sua flexibilidade para grandes volumes de dados não estruturados, 50%. *Lakehouse*, uma arquitetura mais recente, é menos comum, enquanto a migração de sistemas é rara, indicando maior foco na integração entre sistemas existentes. Porém, muitas empresas combinam diferentes estratégias, refletindo maturidade ou necessidades específicas em suas operações de dados.

(QE3) Qual a complexidade (quantidade de atributos envolvidos) da fonte de dados com a qual você já trabalhou em uma atividade de integração de dados?

Figura 2(c) mostra que a maioria dos respondentes (51,4%) lidam com fontes de dados de alta complexidade em projetos de integração, indicando que as empresas frequentemente enfrentam grandes volumes de atributos. Apenas 17,1% trabalham com baixa complexidade, embora isso também exija gestão robusta e habilidades avançadas. Com 31,4% lidando com baixa complexidade, existe um mercado potencial para especialistas nesse nível. Esses dados apontam a necessidade de estratégias escaláveis para integrar dados com diferentes níveis de complexidade à medida que crescem em volume e variedade.



(a) Quantidade de pessoas. (b) Carga horária dedicada. (c) Tipos de dados.

Figure 3: Recursos humanos envolvidos e tipos de dados.

(QE4) Qual a quantidade de pessoas envolvidas em projetos de integração na sua empresa?

Figura 3(a) mostra que a maior parcela dos projetos de integração de dados envolve equipes de até 3 pessoas (31,4%), sugerindo que são geralmente projetos de menor escala ou que as empresas preferem equipes concisas e ágeis. Projetos com grandes equipes (11,4%) são menos comuns, indicando que integrações de maior escala são raras ou restritas a empresas maiores. O tamanho das equipes pode estar relacionado à complexidade do projeto e à cultura organizacional da empresa em relação à gestão de integração de dados.

(QE5) Qual o esforço desempenhado em horas semanais dedicadas em projetos de integração na sua empresa?

A maioria dos respondentes (31,4%) dedica de 6 a 10 horas semanais a projetos de integração de dados (Figura 3(b)), indicando que essa não é, para muitos, uma atividade de tempo integral. Há uma diversidade no tempo investido, refletindo diferenças no escopo dos projetos, no tamanho das empresas ou nos papéis dos profissionais. Cerca de 37,2% dedicam mais de 20 horas semanais, sugerindo a presença de projetos maiores ou funções específicas focadas exclusivamente na integração de dados.

(QE6) Quais tipos de dados ou esquemas sua empresa precisa integrar com mais frequência?

A integração de dados estruturados é a mais comum (74,3%), refletindo a centralidade dos bancos de dados relacionais nas operações empresariais (Figura 3(c)). Dados semiestruturados, como JSON e XML, representam 11,4%, indicando o uso de APIs e web services, enquanto dados não estruturados (5,7%) mostram a diversidade das fontes utilizadas. A integração de múltiplos tipos de dados destaca a necessidade de flexibilidade nas ferramentas e abordagens. Apesar do foco em dados estruturados, a variedade de esquemas e formatos mantém a complexidade dos desafios de integração.

(QE7) Quais fatores você acredita que contribuem para o aumento do esforço em projetos de integração?

Os principais fatores que aumentam o esforço em projetos de integração de dados incluem a falta de documentação adequada (77,1%) e a falta de padronização nos formatos de dados (68,6%), evidenciando a importância de informações claras e padrões consistentes. Mudanças frequentes nos requisitos (60%) e a complexidade dos sistemas ou plataformas (45,7%) refletem desafios em ambientes dinâmicos e sistemas complexos. O volume elevado de dados (34,3%) também é um obstáculo significativo. A falta de experiência ou treinamento é apontada por 20% dos respondentes, indicando a necessidade de desenvolvimento profissional contínuo, enquanto que a integração com dados de terceiros e ferramentas externas é mencionada por 34,3%, sugerindo a complexidade do processo como um fator relevante.

(QE8) Que estratégias sua empresa adota para gerenciar a complexidade na integração de dados?

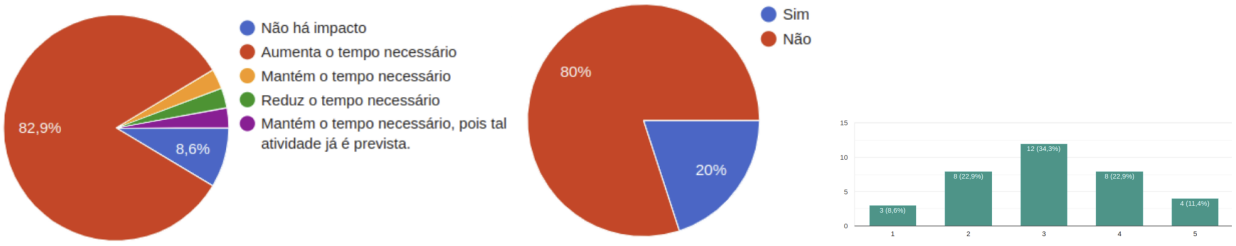
As estratégias mais adotadas para gerenciar a complexidade na integração de dados incluem a padronização de processos (65,7%) e o uso de ferramentas específicas (48,6%), indicando a importância de uniformidade e tecnologia. A contratação de especialistas (37,1%) reforça a relevância do conhecimento técnico especializado. Apenas 14,3% atualizam regularmente os requisitos de integração, apontando uma oportunidade de melhoria. Algumas empresas não possuem estratégias claras, sugerindo potencial para consultorias e soluções de integração. Além disso, há demanda por formação em padronização e ferramentas de integração.

(QE9) Qual é o impacto da complexidade da etapa de integrações de dados no tempo para concluir um projeto?

A maioria dos respondentes (82,9%) acredita que a complexidade na integração de dados aumenta o tempo de conclusão de projetos, destacando-a como um grande obstáculo à eficiência (Figura 4(a)). Isso reforça a importância de considerar a complexidade nas estimativas de prazo e planejamento de projetos. As empresas podem adotar melhores práticas, ferramentas especializadas e investir no treinamento de equipes para mitigar esse impacto. A tendência de subestimar a complexidade nas fases iniciais sugere a necessidade de avaliações mais precisas para melhorar as estimativas futuras. Reduzir o impacto no tempo de projetos pode se tornar um diferencial competitivo para empresas focadas em eficiência operacional.

(QE10) Sua empresa atualmente está utilizando *schema matching* automático ou semi automático?

Figura 4(b)) mostra que a maioria das empresas (28, ou seja, 80%) ainda não utiliza *Schema Matching* automático ou semi-automático, enquanto 7 empresas (20%) adotam essa tecnologia. Isso aponta para uma hesitação em adotar o *Schema Matching* automático, possivelmente devido a custos, complexidade ou falta de expertise. Essa situação cria uma oportunidade para o desenvolvimento e a adoção dessas ferramentas, com foco em educação e treinamento sobre seus benefícios e implementação. A falta de automação resulta em uma gestão manual com menor e eficiência e possível perda de efetividade em relação à qualidade dos dados e compatibilidade entre sistemas. A adoção do *Schema Matching* automático pode se tornar um diferencial competitivo, estimulando mais empresas a adotar essa prática.



(a) Complexidade da integrações de dados. (b) *Schema matching* automático? (c) Objetivos.

Figure 4: Complexidade, características e objetivos.

(QE11) Qual o principal objetivo ou caso de uso para o *schema matching* na integração de sistemas em sua organização?

Como ilustrado na Figura 4(c), os principais objetivos de uso do *schema matching* incluem a construção de Data Warehouses (48,6%) e a migração de bancos de dados (42,9%), evidenciando sua relevância em projetos de consolidação e transição de dados. A integração de bancos de dados (45,7%) e a melhoria da qualidade dos dados (31,4%) também se destacam como casos de uso importantes.

Menos frequentes, mas ainda significativos, estão a exploração e descoberta de dados (31,4%), a construção de bases de conhecimento (20%) e modificações de esquemas (11,4%), apontando para aplicações mais especializadas. Esses dados sugerem oportunidades para expandir o uso do *schema matching*, especialmente em áreas menos exploradas, conforme as organizações buscam aprimorar suas capacidades de gestão e análise de dados.

(QE12) Como você avaliaria o desempenho do *schema matching* em sua organização em uma escala de 1 a 5, sendo 1 muito ineficaz e 5 muito eficaz?

A avaliação do desempenho do *Schema Matching* nas organizações mostra uma distribuição variada, com 34,3% (11) das empresas classificando-o como abaixo da média (notas 1 e 2), indicando desafios na eficiência da ferramenta. Aproximadamente 34,3% (12) das empresas atribuíram notas 4 e 5, apontando que ainda há espaço para melhorias.

A avaliação do *Schema Matching* é polarizada, com 8,6% (3) considerando-o muito ineficaz e 11,4% (4) muito eficaz, sugerindo que as experiências variam entre as organizações. A complexidade dos processos de integração de dados e a qualidade dos dados podem estar influenciando essas percepções. Estratégias de integração mais maduras e dados mais estruturados podem melhorar a eficácia do *schema matching*.

(QE13) Quais ferramentas ou tecnologias de *schema matching* sua organização utiliza atualmente?

A maioria das organizações prefere desenvolver suas próprias ferramentas de *schema matching* (Figura 5(a)), indicando uma busca por soluções personalizadas ou a falta de ofertas completas no mercado. Há uma adoção limitada de ferramentas comerciais e de código aberto, o que representa uma oportunidade para fornecedores aprimorarem suas soluções. Cerca de 34,3% (12) das organizações não utilizam ferramentas de *schema matching*, sugerindo falta de conhecimento ou satisfação com métodos manuais. A alta utilização de ferramentas internas pode refletir a necessidade de customização devido à complexidade dos dados e processos, oferecendo uma oportunidade para inovação e parcerias com fornecedores.

(QE14) Quais são os principais desafios que sua organização enfrentou ao implementar o uso do *schema matching*?

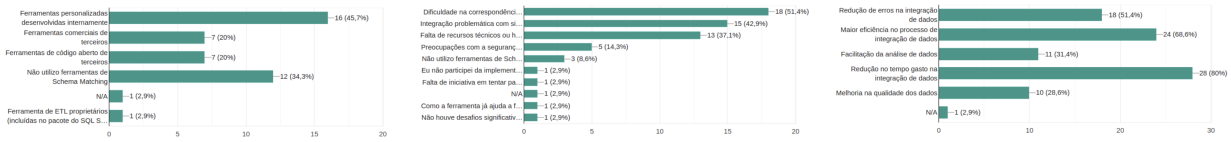
Figura 5(b) mostra que os principais desafios enfrentados pelas organizações ao implementar o uso do *schema matching* incluem a complexidade (51,4%), indicando a necessidade de abordagens mais eficazes e automação avançada. Cerca de 42,9% das empresas enfrentam dificuldades para integrar o *schema matching* com sistemas existentes, sugerindo a necessidade de soluções mais integradas. A falta de recursos técnicos ou humanos (37,1%) pode ser uma barreira para a adoção eficaz do *schema matching*. Além disso, a segurança (14,3%) é uma preocupação importante, refletindo a necessidade de soluções seguras por design. A baixa adesão à padronização de esquemas de dados (2,9%) pode ser um reflexo de problemas organizacionais ou culturais que dificultam a adoção. Isso representa uma oportunidade clara para melhorar as soluções de *schema matching* e oferecer suporte adicional às organizações.

(QE15) Em sua opinião, quais são os principais benefícios do uso de *schema matching* automático ou semi automático?

Os principais benefícios do uso de *schema matching* automático ou semi automático são a maior eficiência no processo de integração de dados (68,6%) e a redução no tempo gasto (80%), evidenciando a valorização dessa tecnologia por sua capacidade de acelerar e otimizar a integração (Figura 5(c)). Mais da metade dos respondentes (51,4%) reconhece que ela contribui para uma menor taxa de erro, resultando em integrações mais confiáveis. Além disso, a facilitação da análise de dados (31,4%) e a melhoria na qualidade dos dados (28,6%) também são vistas como vantagens substanciais. Apenas 2,9% dos respondentes se abstiveram de opinar, o que reflete um consenso positivo. Isso sugere que, apesar dos benefícios, o treinamento contínuo pode ser necessário para otimizar a utilização do *schema matching*.

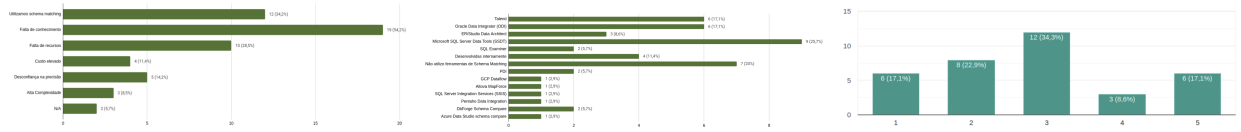
(QE16) Quais são as principais razões para a não utilização do *schema matching*?

Figura 6(a) mostra que as principais razões pelas quais as organizações não utilizam *schema matching* incluem a falta de conhecimento (51,4%), o que indica uma grande necessidade de educação e treinamento para capacitar as equipes. A desconfiança na precisão das ferramentas (14,3%) também é uma barreira importante, sugerindo que a confiabilidade é uma preocupação significativa.



(a) Tipos de ferramentas. (b) Principais desafios. (c) Principais benefícios.

Figure 5: Desempenho, ferramentas e desafios em *schema matching*.



(a) Motivação. (b) Ferramentas. (c) Satisfação.

Figure 6: Motivação, ferramentas e satisfação.

O custo e a falta de recursos (28,6%) também limitam a adoção, sendo percebidos como obstáculos financeiros. Por outro lado, a alta complexidade e o custo elevado são barreiras menos citadas (8,6%), indicando que, embora sejam desafios, não são as principais preocupações para a maioria das organizações. Isso revela uma oportunidade para fornecedores de soluções de *schema matching* promoverem suas ferramentas, enfatizando sua eficácia e abordando a falta de conhecimento e confiança. A diversidade das razões também destaca a necessidade de abordagens personalizadas para promover a adoção do *schema matching*, levando em consideração as necessidades específicas de cada organização.

(QE17) Quais ferramentas de *schema matching* você utilizou em projetos de integração de dados?

Como ilustrado na Figura 6(b), as ferramentas de *schema matching* mais utilizadas em projetos de integração de dados são as de grandes fornecedores, como Microsoft SQL Server Data Tools (SSDT) (25,7%), Oracle Data Integrator (ODI) (17,1%) e Talend (17,1%), o que reflete uma preferência por soluções consolidadas e confiáveis. No entanto, há também uma diversidade de outras ferramentas utilizadas, sem uma clara preferência entre as organizações. Algumas respostas indicam o uso de soluções internas, sugerindo uma necessidade de personalização ou integração específica com sistemas internos. Além disso, a presença de ferramentas de código aberto indica uma adesão crescente a soluções mais acessíveis, embora ainda não tão predominante quanto as ferramentas de grandes fornecedores. Esse cenário aponta para uma possível expansão no mercado de *schema matching*, tanto para soluções comerciais quanto de código aberto.

(QE18) Como sua empresa lida com conflitos de correspondência ao usar *schema matching*?

A maioria das empresas lida com conflitos de correspondência em *schema matching* de forma manual, indicando uma preferência ou necessidade de intervenção humana em conflitos complexos. Apenas uma fração adota ferramentas automáticas ou semi-automáticas, o que pode refletir falta de conhecimento, recursos limitados ou a preferência por controle manual. Uma pequena porcentagem (8,6%) não enfrenta conflitos, possivelmente devido a processos simplificados ou domínio de dados menos complexo. A predominância do método manual sugere uma oportunidade para promover o desenvolvimento de ferramentas. A dependência de processos manuais também evidencia preocupações com a qualidade e confiabilidade das soluções automatizadas disponíveis.

(QE19) Sua empresa tem estratégias ou planos para melhorar o uso de *schema matching* no futuro? Se sim, quais são esses planos?

O gráfico mostra que 28,6% das empresas planejam adotar novas ferramentas de *schema matching*, enquanto 20% focam no aprimoramento das habilidades da equipe. No entanto, 45,7% não têm planos de mudança, o que pode indicar satisfação com os sistemas atuais ou resistência a mudanças. A falta de respostas sugere incerteza sobre a evolução das práticas de *schema matching*. Isso representa uma oportunidade para fornecedores de soluções inovadoras e destaca a necessidade de estratégias claras para o futuro do *schema matching*.

(QE20) Que práticas sua empresa adota para documentar e monitorar as correspondências de esquemas ao longo do tempo?

As empresas adotam práticas variadas para documentar e monitorar o *schema matching*. A documentação de esquemas é a mais comum, com 48,6% (17) das empresas indicando sua importância para manter registros detalhados. Apenas 5,7% (2) realizam monitoramento contínuo, enquanto 11,4% (4) mantêm registros históricos, refletindo menor prioridade nessas áreas. A ausência de práticas em algumas empresas destaca riscos e oportunidades de melhoria, especialmente em setores regulamentados. Esse cenário evidencia espaço para ferramentas que automatizem a documentação e o monitoramento, além de diferentes níveis de maturidade na governança de dados entre as organizações.

(QE21) Sua empresa realiza alguma validação ou verificação posterior ao processo de *schema matching* para garantir a precisão e a integridade dos dados?

A validação ou verificação após o *schema matching* é realizada por 37,1% (13) das empresas, enquanto 57,1% (20) não adotam essa prática regularmente, o que destaca a ausência de uma etapa crítica

para muitas organizações. Além disso, há registros de desconhecimento sobre a realização de validação, o que pode refletir lacunas de comunicação interna ou compreensão limitada dos processos de gestão de dados.

Esses dados sugerem oportunidades para melhorias nas práticas de garantia de qualidade e para a adoção de ferramentas que automatizem esse processo. A validação regular pode ser integrada em estratégias de melhoria contínua, otimizando a precisão e a integridade dos dados ao longo do ciclo de integração.

(QE22) Em uma escala de 1 a 5, quão satisfeito você está com o processo de *schema matching* em sua empresa?

Figura 6(c) mostra que a insatisfação com os processos de *schema matching* é significativa, com 40% das respostas nas notas 1 e 2, enquanto a maioria avalia com nota 3, refletindo percepção mediana. Apenas uma pequena parcela está plenamente satisfeita (notas 4 e 5), indicando espaço para melhorias. Isso destaca a necessidade de melhores ferramentas, práticas mais eficazes e treinamento, além de oportunidades para inovação tecnológica e serviços especializados. Avaliações contínuas e *feedback* podem ajudar as empresas a identificar e abordar pontos críticos, aumentando a satisfação geral.

(QE23) Quais ferramentas de *schema mapping* (Transformação) você utiliza na sua empresa?

As ferramentas de *schema mapping* mostram diversidade na adoção pelas empresas. Pentaho PDI lidera como a mais utilizada, destacando-se por sua robustez e custo-benefício. Ferramentas de código aberto, como Apache NiFi e Apache Camel, estão ganhando espaço, enquanto o Microsoft SQL Server Integration Services (SSIS) mantém relevância em algumas organizações. A baixa adoção de muitas ferramentas indica preferências específicas e oportunidades para expansão ou desenvolvimento de novas soluções. Além disso, algumas empresas desenvolvem soluções internas ou não utilizam ferramentas específicas, sugerindo necessidades únicas. Essa diversidade ressalta a importância de capacitação e estratégias alinhadas aos recursos e sistemas existentes.

4.3 Respostas às Questões de Pesquisa do Survey

A seguir, será apresentada a análise dos resultados das questões de pesquisa.

(QP 1) Quais são as abordagens predominantes que as empresas utilizam atualmente para *schema matching*? Os gráficos revelam que a maioria dos projetos de integração de dados envolve pequenas equipes de até três pessoas (31,4%), sugerindo uma preferência por equipes menores e ágeis. Além disso, a maioria dos respondentes dedica de 6 a 10 horas semanais a esses projetos, indicando que a integração de dados não é uma atividade de tempo integral para muitos profissionais. Em termos de tipos de dados integrados, dados estruturados são os mais comuns, refletindo a importância contínua de bancos de dados relacionais, enquanto dados semiestruturados como JSON e XML também são frequentemente integrados, mostrando o uso de APIs e trocas de dados entre sistemas modernos. A necessidade de integrar vários tipos de dados exige ferramentas flexíveis e abordagens diversificadas, com uma atenção especial aos desafios específicos de integração que variam com a complexidade dos projetos e o tamanho das equipes.

(QP 2) Quais são os benefícios que as empresas percebem na abordagem que elas utilizam para o *schema matching*? A maioria dos respondentes (82,9%) acredita que a complexidade na integração de dados aumenta o tempo necessário para concluir um projeto, destacando a necessidade de planejamento cuidadoso e estratégias de mitigação. Apenas 20% das empresas utilizam *schema matching* automático ou semi-automático, indicando uma baixa adoção e sugerindo oportunidades de mercado e necessidade de conscientização sobre os benefícios dessa tecnologia. O principal uso de *schema matching* é na construção de Data Warehouses, seguido pela migração de bancos de dados e melhoria da qualidade dos dados, ressaltando sua importância na consolidação e transição de sistemas de dados.

(QP 3) Quais são os principais desafios ou dificuldades enfrentados ao realizar *schema matching*? A maioria (31,5%) considera a ferramenta abaixo da média (notas 1 e 2), com apenas 34,3% classificando-a como eficaz (notas 4 e 5), indicando necessidade de melhorias e treinamento. A maioria das organizações prefere desenvolver suas próprias ferramentas de *schema matching* personalizadas, com uma adoção limitada de soluções comerciais e de código aberto, apontando para uma oportunidade de mercado. Os principais desafios na implementação incluem complexidade, integração com sistemas existentes, falta de recursos técnicos e preocupações com segurança, sugerindo a necessidade de soluções mais eficazes e suporte aprimorado.

(QP 4) Qual é o grau de esforço desempenhado pelas empresas na integração de dados? (QP 5) Qual é o nível de complexidade das integrações de dados com que as empresas estão atualmente lidando? (QP 6) Qual é o processo de *schema matching* que as empresas estão empregando? (QP 7) Qual é o nível de satisfação das empresas ao utilizar o *schema matching*?

As principais razões para a não utilização do *schema matching* incluem a falta de conhecimento, a desconfiança na precisão das ferramentas, o custo e a falta de recursos. A alta complexidade é menos citada como uma barreira. Há uma oportunidade para os fornecedores de ferramentas promoverem o valor de suas soluções e a diversidade das razões aponta para a necessidade de abordagens personalizadas para incentivar a adoção do *schema matching*. Em relação às ferramentas utilizadas, as soluções de grandes fornecedores são as mais comuns, mas há uma variedade de ferramentas utilizadas, incluindo soluções internas e de código aberto. A satisfação com o processo de *schema matching* é geralmente baixa, com uma grande parte das empresas insatisfeitas ou apenas medianamente satisfeitas, indicando a necessidade de melhorias nos processos e ferramentas utilizadas.

5 AMEAÇAS À VALIDADE

Algumas ameaças à validade dos resultados obtidos neste trabalho foram identificadas, como discutido a seguir.

Tamanho da Amostra: Uma pequena amostra de 35 profissionais da indústria de software brasileira. Para maximizar o número de participantes, foi feito o convite ao maior número possível de profissionais dentro de um período fixo. Buscou-se obter uma representação diversificada das diferentes realidades regionais e nacionais. No entanto, o estudo foi limitado pelo número de respostas, uma dificuldade também observada em outros estudos de engenharia de software [22]. Além disso, o número e o tipo de questões podem ser considerados como possíveis limitações do estudo.

Análises e Inferências: As inferências obtidas também estão inerentemente sujeitas a interpretações e opiniões humanas. Para mitigar esta ameaça, três engenheiros de software e especialistas em integração de dados apoiaram a elaboração e avaliação das inferências. Outras limitações típicas de estudos de pesquisa de opinião também podem acontecer, como o baixo potencial de generalização dos resultados comunicados, visto que nem todos os Estados brasileiros foram cobertos, nem todas as regiões e poucas empresas de cada Estado foram consideradas. Investigações adicionais são necessárias para obter um panorama mais amplo que represente, de fato, o estado da prática no país.

6 CONCLUSÕES

A principal contribuição deste trabalho é fornecer um panorama inicial sobre estado da prática em integração de dados no Brasil, com foco na etapa de *schema matching*. Integração de dados é parte imprescindível do processo de interoperar Sistemas de Informação a fim de atingir a almejada interoperabilidade plena, que é reconhecida como um dos Grandes Desafios de Pesquisa em Sistemas de Informação [17]. Neste contexto, foi conduzida uma pesquisa de opinião envolvendo profissionais de diferentes setores da indústria. As respostas das questões foram analisadas quantitativamente. De um conjunto de 35 participantes, mais da metade (precisamente 23 participantes) tem familiaridade com o conceito de *schema matching*. Além disso, vários participantes (mesmo os menos experientes em *schema matching*) identificaram os benefícios na redução no tempo gasto com a integração de dados, redução de erros e uma maior eficiência no uso da técnica de forma automática. A análise realizada permitiu identificar que (i) os profissionais vislumbram diversas vantagens e oportunidades para aplicar o *schema matching* automático na integração de dados, (ii) o tamanho e a diversidade da base de dados deve fornecer um custo-benefício eficaz para a aplicação da técnica antes de sua adoção, e (iii) qualquer estratégia de adoção bem-sucedida deve ser avaliada de acordo com a complexidade da integração de dados.

Os resultados mostram que respondentes envolvidos em projetos de maior complexidade tendem a perceber de forma mais clara os benefícios do uso de técnicas de *schema matching*. Os benefícios identificados pelos participantes estão predominantemente associados à otimização de tempo, visto que a aplicação de *schema matching* automático reduz o tempo gasto na integração de dados, facilitando a conciliação de esquemas divergentes e acelerando a disponibilidade de informações para tomada de decisão. Quanto aos desafios e dificuldades, estes foram frequentemente relacionados à complexidade dos esquemas de dados envolvidos, à presença de dados inconsistentes que dificultam a integração automática, e à falta de treinamento específico para as equipes técnicas.

Trabalhos Futuros envolvem expandir a consulta para outros Estados Brasileiros, além de investigar como técnicas de Inteligência Artificial (IA) podem beneficiar este trabalho. Os resultados trazidos indicam a necessidade de (i) avanços científicos na área, (ii) desenvolvimento de ferramentas e (iii) treinamento especializado, servindo como uma agenda de trabalho para os especialistas da área no Brasil.

REFERENCES

[1] Nicolas Atmatzides, Marcos Bedo, and Daniel de Oliveira. 2022. Adoção de SGBDs NoSQL em Empresas Brasileiras: um Levantamento Preliminar. In *SBBD*. SBC, Búzios, 385–390. https://doi.org/10.5753/sbbd.2022.226015

[2] Matheus Batista, Andréa Magdaleno, and Marcos Kalinowski. 2017. A Survey on the use of Social BPM in Practice in Brazilian Organizations. In *Anais do XIII Simpósio Brasileiro de Sistemas de Informação* (Lavras). SBC, Porto Alegre, RS, Brasil, 436–443. https://doi.org/10.5753/sbsi.2017.6073

[3] Philip A. Bernstein and Laura M. Haas. 2008. Information Integration in the Enterprise. *Commun. ACM* 51, 9 (2008), 72–79.

[4] Flavia Cristina Berrnardini, José Viterbo, Dalessandro Vianna, Carlos Bazilio Martins, Adriana Pereira Medeiros, Edwin Meza, Patrick Moratori, and Carlos Alberto Malcher Bastos. 2017. *Grand Challenges for Information Systems in Brazil for the Decade 2016-2026.* SBC, Chapter General Features of Smart City Approaches from Information Systems Perspective and Its Challenges.

[5] Júlio Campos, Vitor Almeida, Elvismary Armas, Geiza Silva, Eduardo Corseuil, and Fernando Gonzalez. 2023. INSIDE: an Ontology-based Data Integration System Applied to the Oil and Gas Sector. In *Anais do SBSI 2023.*

[6] Roberto Dias, Rodrigo Zacarias, Jorge Luis Varella, and Rodrigo dos Santos. 2022. Investigating Information Security in Systems-of-Systems. In *Anais do XVIII Simpósio Brasileiro de Sistemas de Informação* (Curitiba). SBC, Porto Alegre, RS, Brasil. https://sol.sbc.org.br/index.php/sbsi/article/view/21353

[7] AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. 2012. *Principles of Data Integration.* Morgan Kaufmann.

[8] Ana Carolina Ferronato, Fernanda Pires, and Flavia Bernardini. 2016. Um Modelo para Integração e Disponibilização de Dados na Área de Saúde Governamental. In *Anais do SBSI 2016.* Florianópolis, 124–127.

[9] Valdemar Vicente Graciano Neto, Flavio Oquendo, and Elisa Yumi Nakagawa. 2017. *Grand Challenges for Information Systems in Brazil for the Decade 2016-2026.* SBC, Chapter Smart Systems-of-Information Systems: Foundations and an Assessment Model for Research Development.

[10] Laura M. Haas. 2007. Beauty and the Beast: The Theory and Practice of Information Integration. In *Proc. of the International Conference on Database Theory.* 28–43.

[11] Laura M. Haas, Mauricio A. Hernández, Howard Ho, Lucian Popa, and Mary Roth. 2005. Clio Grows Up: From Research Prototype to Industrial Tool. In *Proceedings of the SIGMOD Conference.* 805–810.

[12] Alon Y. Halevy. 2005. Why Your Data Won't Mix. *ACM Queue* 3, 8 (2005), 50–58.

[13] Ahmed A. Harby and Farhana H. Zulkernine. 2022. From Data Warehouse to Lakehouse: A Comparative Review. In *Proceedings of the IEEE BigData.* IEEE, 389–395.

[14] Mark Kasunic. 2005. Designing an effective survey. *Carnegie Mellon University, Software Engineering Institute Pittsburgh, PA* (01 2005), 142.

[15] Guoliang Li. 2017. Human-in-the-loop Data Integration. *Proceedings of the VLDB Endowment* 10, 12 (2017), 2006–2017.

[16] Johan Linaker, Sardar Muhammad Sulaman, Martin Höst, and Rafael Maiani de Mello. 2015. Guidelines for conducting surveys in software engineering v. 1.1. *Lund University* 50 (2015).

[17] Rita Suzana P. Maciel and Regina Braga José Maria N. David, Daniela Barreiro Claro. 2017. Full Interoperability: Challenges and Opportunities for Future Information Systems. In *I GranDSI-BR – Grand Research Challenges in Information Systems in Brazil 2016-2026,* Clodis Boscarioli, Renata M. Araujo, and Rita Suzana P. Maciel (Eds.). SBC, 107–118. https://doi.org/978-85-7669-384-0 Capítulo de eBook.

[18] Jefferson Seide Molléri, Kai Petersen, and Emilia Mendes. 2016. Survey guidelines in software engineering: An annotated review. In *Proc. of the 10th ACM/IEEE ESEM.* 1–6.

[19] Eduardo Soares Paiva, Kate Cerqueira Revoredo, and Fernanda Araujo Baião. 2016. DW-CGU: Integração dos Dados do Portal da Transparência do Governo Federal Brasileiro. *iSys* 9, 1 (2016), 6–32. https://doi.org/10.5753/isys.2016.298

[20] Ana Paula Perin, Deivid Silva, and Natasha Valentim. 2022. Investigating the Teaching of Block Programming in High School. In *Anais do XVIII Simpósio Brasileiro de Sistemas de Informação* (Curitiba). SBC, Porto Alegre, RS, Brasil. https://sol.sbc.org.br/index.php/sbsi/article/view/21373

[21] Erhard Rahm and Philip A. Bernstein. 2001. A Survey of Approaches to Automatic Schema Matching. *The VLDB Journal* 10, 4 (2001), 334–350.

[22] Edward Smith, Robert Loftin, Emerson Murphy-Hill, Christian Bird, and Thomas Zimmermann. 2013. Improving developer participation rates in surveys. In *6th CHASE.* IEEE, 89–92.

[23] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in software engineering.* Springer Science & Business Media.

[24] Yunjia Zhang, Avrilia Floratou, Joyce Cahoon, Subru Krishnan, Andreas C. Müller, Dalitso Banda, Fotis Psallidas, and Jignesh M. Patel. 2023. Schema Matching using Pre-Trained Language Models. In *Proceedings of the ICDE Conference.* IEEE, 1558–1571.