Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise de Métodos, Técnicas e Ferramentas para a Engenharia de Requisitos em Big Data

Victor Carneiro Seidel

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientadora
Prof.a Dr.a Edna Dias Canedo

Brasília
2023

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise de Métodos, Técnicas e Ferramentas para a Engenharia de Requisitos em Big Data

Victor Carneiro Seidel

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof.a Dr.a Edna Dias Canedo (Orientadora)
CIC/UnB

Prof.a Dr.a Aletéia Patricia Favacho de Araujo
CIC/UnB

Dr. Daniel de Paula Porto
CIC/UnB

Prof. Dr. Marcelo Grandi Mandelli
Coordenador do Bacharelado em Ciência da Computação

Brasília, 18 de Julho de 2023

Dedicatória

Dedico este trabalho ao meu avô Aloysio Victor Seidel (in memorian), que em seus 89 (oitenta e nove) anos de vida sempre foi um exemplo de perseverança, força de vontade e solidariedade para com todos os seus familiares e desconhecidos. Com muito bom humor, energia e entusiasmo, transmitia a importância de um trabalho bem feito.

Estendo também homenagem ao meu finado tio Alberto de Miranda Carneiro, o qual sempre se manteve otimista mesmo diante de muitas situações difíceis.

Enfim, a todos os familiares, amigos e colegas de curso, pela paciência e auxílio ao longo desta jornada.

Agradecimentos

Aos meus pais Alba e Aloysio, pelos ensinamentos e motivação ao longo da vida.
Aos familiares e amigos pelo apoio e companheirismo.
Aos professores do Departamento de Ciência da Computação que atuaram e possibilitaram minha formação no ensino superior.

Resumo

A Engenharia de Requisitos é considerada pela academia e pela indústria como a fase mais importante do processo de desenvolvimento de software, por possibilitar logo no início entender o problema, as necessidades dos envolvidos e o objetivo a ser alcançado, de forma coesa, mantendo a integridade do projeto. Na construção de um sistema que envolve *Big Data*, a importância da engenharia de requisitos se torna ainda maior, frente aos desafios de armazenar, processar, analisar diversos e volumosos dados válidos, para que seja possível extrair valor confiável para o negócio. Este documento realiza uma investigação do uso total ou parcial de métodos, técnicas e ferramentas da engenharia de requisitos para *Big Data* em instituições financeiras nacionais de grande porte. Para isto, foram identificados 314 estudos por meio de uma revisão sistemática da literatura, dos quais 11 foram selecionados como estudos primários para realizar a análise de dados. Assim foi realizado um *survey* contendo 22 questões, com o objetivo de obter a percepção dos profissionais de Tecnologia da Informação das instituições financeiras brasileiras quanto à usabilidade e apoio dos estudos identificados e das técnicas existentes na literatura e na indústria. A aplicação do *survey* ocorreu durante 35 dias e obteve a resposta de 52 participantes. Os resultados da pesquisa demonstraram uma boa aceitação das ferramentas identificadas durante a revisão de literatura, e os profissionais mostraram se dispostos em usá-las na indústria. Os profissionais destacaram a necessidade de uma ferramenta web colaborativa que abrangesse todo o ciclo de desenvolvimento de software no contexto de *Big Data*, assim como um *framework* para apoiar a elicitação automática de requisitos orientado a dados de fontes externas à organização. Contudo, nenhuma das propostas dos estudos é aplicada em sua totalidade nas instituições dos participantes que responderam o *survey*.

Palavras-chave: *Big Data*, Engenharia de Requisitos, Métodos, Técnicas, Ferramentas, Revisão Sistemática da Literatura, *Survey*

Abstract

Requirements Engineering is considered by academia and industry as the most important phase of the software development process, as it enables the early understanding of the problem, the stakeholders' needs, and the intended objective in a cohesive manner, while maintaining the project's integrity. In the construction of a system involving Big Data, the importance of requirements engineering becomes even greater, given the challenges of storing, processing, and analyzing diverse and voluminous valid data, in order to extract reliable business value. This document investigates the total or partial use of methods, techniques, and tools of requirements engineering for Big Data in large national financial institutions. For this purpose, 314 studies were identified through a systematic literature review, of which 11 were selected as primary studies for data analysis. A survey was conducted consisting of 22 questions, with the aim of obtaining the perception of Information Technology professionals from Brazilian financial institutions regarding the usability and support of the identified studies and the existing techniques in literature and industry. The survey was conducted over a period of 35 days and received responses from 52 participants. The research results demonstrated a good acceptance of the tools identified during the literature review, and the professionals showed willingness to use them in the industry. The professionals emphasized the need for a collaborative web tool that encompasses the entire software development lifecycle in the context of Big Data, as well as a framework to support the automatic elicitation of requirements from external data sources to the organization. However, none of the proposals from the studies are fully implemented in the institutions of the participants who responded to the survey.

Keywords: Big Data, Requirement Engineering , Methods, Techniques, Tools, Systematic Literature Review, Survey

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A importância da Engenharia de Requisitos (RE) é amplamente reconhecida na academia e no mercado [7]. É a chave de sucesso para que o entendimento do escopo do projeto seja alcançado e aprovado entre todos os envolvidos, e se derive custo e prazo para a sua concepção, conforme as metas e objetivos estabelecidos [8], [9].

Nos projetos de desenvolvimento de sistemas *Big Data* (BD) não é diferente, a presença da engenharia de requisitos agrega suma importância para encontrar os valores de negócios [8] e manter a qualidade e a confiabilidade dos dados [9]. Esses valores ajudam as partes interessadas a entender a importância do projeto e seu valor no mercado [10], com o desafio de garantir a conversão efetiva de dados em conhecimento aplicável [11], evitar falhas de concepção [12] e permitir o desenvolvimento de um sistema resiliente [13].

O termo RE se refere a um dos processos cruciais na montagem de qualquer sistema em razão de possibilitar logo no início do desenvolvimento do projeto, o levantamento dos requisitos, obtidos das necessidades dos usuários, do entendimento do problema ou visando alcançar um objetivo. Abrange também descrever a compreensão destes requisitos de forma coesa, gerenciando suas mudanças, mantendo a integridade do projeto [14]. Pressman e Castro apresentaram estudos sobre quanto mais tarde se descobre os erros cometidos nesta fase inicial, mais alto é o custo de sua correção. Neste contexto, a Engenharia de Requisitos, uma disciplina da Engenharia de Software, foi criada com o objetivo de definir normas na utilização e na gestão destes requisitos [14].

Os requisitos são classificados como funcionais (o que o sistema deve fazer) e os não funcionais, que descrevem como os requisitos funcionais são implementados, suas restrições, aspectos de desempenho, segurança, confiabilidade, padrões, etc. As metas, políticas e estratégias da empresa são consideradas como requisitos organizacionais, com o intuito de abranger também os aspectos gerenciais, organizacionais, econômicos, sociais e ambientais no contexto da organização [14]. Cabe ao Engenheiro de Requisitos, ao se comunicar de forma eficaz e frequente com as partes interessadas, elicitar, analisar, modelar e gerenciar

os requisitos definidos durante o projeto [8].

A expressão *Big Data*, por sua vez, é um conceito inicialmente aplicado para descrever uma grande quantidade de dados (estruturados, não estruturados e semi-estruturados), internos ou externos de uma instituição, que possuem alta velocidade de uso, alto volume e uma alta variedade (diversidade) [15]. Posteriormente, foram adicionadas mais duas características relevantes na descrição de um sistema de BD: a veracidade das informações e o valor a elas atribuído [16]. O volume se refere à enorme capacidade de armazenamento dos dados, com crescimento exponencial diário. A velocidade se refere à rapidez que os dados gerados precisam ser processados e analisados. A variedade se refere aos tipos de formatos, obtidos de diversas fontes de dados. A veracidade se refere à confiabilidade dos dados, a duração em que os dados são válidos, com a filtragem dos irrelevantes. O valor é o resultado esperado da combinação das demais características, obtendo benefícios com as informações relevantes aplicadas ao negócio [8], [17].

O conceito de BD se confunde com o de *Big Data Analytics*. Este último se refere ao processo de coleta, organização e análise dos grandes volumes de dados. Armazenar e analisar BD é um procedimento complexo, pois o armazenamento e processamento de dados de BD requer tecnologia específica (hardware, rede e sistemas). A análise, efetuada por cientistas de dados [8], também precisa de ferramentas analíticas de BD, que se caracteriza pelos seguintes tipos de análise: descritiva (o que aconteceu no passado); diagnóstica (procura entender por que algo aconteceu); preditiva (prevê o que pode acontecer no futuro) e prescritiva (analisa o que aconteceu, por que aconteceu e o que pode acontecer para determinar o que deve ser feito) [15].

Vale ressaltar que, a solução BD permite analisar os dados em tempo real e podem ser provenientes de diferentes fontes e formas, tais como, *cloud computing* (computação em nuvem), *data mining* (mineração de dados) e o grande volume de dados online, como e-mails, gravações de log, mensagens instantâneas, imagens, redes sociais, dentre outros aspectos [18].

No início da "era *Big Data*", os objetivos mais comuns que faziam as organizações apreciarem projetos de BD estava na obtenção de *insights* de estatísticas e otimização da tomada de decisão. Posteriormente, constatou-se que o uso de BD na indústria, por exemplo, não garantia vantagem competitiva, caso não fossem realizadas melhorias nos processos, serviços e na organização do negócio [17]. Atualmente, estamos vivenciando o compartilhamento de dados e informações entre as organizações parceiras e concorrentes.

No Brasil, a Associação Brasileira das Empresas de Software (ABES) divulgou, em abril de 2022, a atualização do estudo "Mercado Brasileiro de Software 2022 - Panorama e Tendências", que o uso de dados no Brasil vai gerar R$ 14,9 bilhões de investimentos em 2022, com a maior parte dos recursos destinada à aquisição de soluções de *Big Data*

Analytics [19].

Com a implantação, no Brasil, da tecnologia de Rede Móveis 5G, impulsionando o mercado de soluções integradas entre *Big Data*, Internet das Coisas (IoT), e Inteligência Artificial (IA), torna-se oportuno investigar a aplicação no mercado brasileiro de abordagens, métodos, técnicas e ferramentas de RE para BD propostas pela literatura.

1.1 Problema de Pesquisa

Nos projetos de BD, os desafios de RE são potencializados, em razão das características e da qualidade dos dados de BD, da complexidade da arquitetura tecnológica envolvida [8], [9], da dificuldade de identificar os requisitos das diversas fontes de dados não intencionais, limitadas em termo de completude, podendo ser ambíguas, não estruturadas, fornecendo apenas um retorno implícito sobre os requisitos, exigindo alto esforço ou até impossibilidade de ser efetuada manualmente [20],[1].

Estudos apontam também que as técnicas de RE tradicionais não são suficientes para obtenção e gerenciamento de todos os requisitos de BD [8], [21], [1] e que poucas propostas de solução constantes na literatura foram aplicadas no mercado [22], [17]. Inclusive, se constatou num estudo de caso exploratório sobre projeto de BD na área de petróleo e gás, que 35% (trinta e cinco por cento) dos requisitos de BD foram identificados nas fases de design, arquitetura, codificação, gerando retrabalho e prejudicando a escalabilidade da arquitetura [21].

Portanto, este trabalho visa identificar na literatura, estudos contendo os métodos, as técnicas e ferramentas de RE para o desenvolvimento de projetos de BD, analisando suas vantagens e desvantagens, assim como obter a percepção de equipes da TI, que trabalham com BD em instituições financeiras nacionais, quanto à aplicação desses métodos, técnicas e ferramentas durante o processo de desenvolvimento de software de BD.

1.2 Justificativa

A pesquisa se propõe a demonstrar que são necessários mais estudos para abordar adequadamente RE em BD, por serem mais complexas as atividades nesse contexto [17].

Obter na literatura soluções de RE que auxiliam a construção de aplicações no contexto de BD, inclusive os trabalhos que não apresentam aplicação em um contexto real, podem auxiliar novos estudos e pesquisas e proporcionar um diferencial para os problemas identificados. Para o mercado, esta contribuição se reflete em economia e vantagem com o sucesso do projeto.

Portanto, diante do volume de investimento em BD e de sua importância para os negócios da organização, identificar as vantagens e as desvantagens, possibilidades de melhorias na utilização de abordagens, métodos, técnicas e ferramentas de RE em BD, pode agregar valor e subsidiar o desenvolvimento destes projetos no mercado nacional.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo geral desse trabalho é identificar a aplicabilidade, total ou parcial, de métodos, técnicas e ferramentas propostos pela literatura referentes a RE na área de projetos de BD, por meio da utilização de questionários direcionados às equipes de Tecnologia da Informação (TI), que atuam em soluções de BD nas instituições financeiras nacionais de grande porte.

1.3.2 Objetivo Específico

Para atingir o objetivo geral deste trabalho, os seguintes objetivos específicos foram definidos:

- Identificar na literatura, estudos que tratam da RE no contexto de BD, seus desafios e as respectivas propostas de solução;
- Identificar os métodos, as técnicas e as ferramentas de RE em projetos de BD propostos pelos estudos identificados;
- Realizar uma análise comparativa dos métodos, técnicas e ferramentas de RE em BD identificadas pelos estudos selecionados, apontando seus aspectos comuns, e se existem pontos positivos ou negativos a destacar em cada uma delas;
- Investigar se os métodos, as técnicas e as ferramentas de RE em BD selecionados são utilizados, total ou parcialmente, nas instituições financeiras.

1.4 Resultados Esperados

Espera-se com este trabalho obter um maior conhecimento sobre RE relacionada ao contexto de BD. Outra expectativa relevante está na identificação dos métodos, técnicas e ferramentas de RE propostos na literatura para o desenvolvimento de projetos de BD, bem como verificar se estão sendo utilizados nas instituições financeiras, considerando suas vantagens e desvantagens.

1.5 Metodologia de Pesquisa

Esta seção apresenta o método utilizado para a condução do estudo e as respectivas questões de pesquisa (RQ), quais sejam:

- **RQ.1**: Quais são as abordagens de RE no contexto de *Big Data* existentes na literatura?
- **RQ.2**: Quais são os métodos, as técnicas e as ferramentas de RE no contexto de *Big Data* existentes na literatura?
- **RQ.3**: Quais destes métodos, técnicas e ferramentas de RE em BD selecionados são utilizados, total ou parcialmente, nas instituições financeiras?

Para responder a primeira questão de pesquisa (RQ.1) e a segunda questão (RQ.2) realizou-se uma revisão sistemática da literatura (SLR), com o objetivo de identificar as abordagens e os métodos, técnicas e ferramentas de RE para BD, realizar uma análise comparativa, assim como suas vantagens, desvantagens e desafios. Os resultados desta etapa são apresentados no Capítulo 3.

Para responder a terceira questão de pesquisa (RQ.3), utilizou-se o método de pesquisa *survey* [23], por meio da utilização de questionários direcionados às equipes de TI, que atuam em soluções de BD nas instituições financeiras nacionais de grande porte. Os resultados desta etapa são apresentados no capítulo 5.

1.6 Estrutura do Trabalho

Este trabalho está organizado da seguinte maneira: O Capítulo 2, contém a revisão sistemática da literatura realizada para a seleção de estudos relevantes. O Capítulo 3 apresenta os resultados das questões de pesquisa de acordo com os resultados da revisão sistemática da literatura. O Capítulo 4 apresenta os detalhes do *survey* estruturado a partir dos documentos selecionados no capítulo anterior. O Capítulo 5 irá descrever a parte prática dos resultados encontrados e discorrerá sobre a análise dos resultados de maneira total. O Capítulo 6 refere-se as conclusões finais acerca do documento.

Capítulo 2

Revisão Sistemática da Literatura

Este capítulo apresenta o detalhamento da revisão de literatura para identificar estudos que têm como base aspectos de RE em BD, com o objetivo de selecionar aqueles considerados relevantes para ser investigada a sua aplicabilidade no mercado.

2.1 Definição da Revisão Sistemática da Literatura

O tipo de revisão escolhida para realizar a seleção dos estudos objeto desta pesquisa foi a Revisão Sistemática da Literatura (SLR) [24] que visa responder uma questão de pesquisa utilizando métodos sistemáticos para consolidar as evidências relevantes que atendam aos critérios pré-definidos. A presente SLR contou com o apoio da ferramenta *Parsifal*[1] que segue todos os passos das diretrizes propostas por Kitchenham e Charters [24], incluindo planejamento, condução e relatório, conforme detalhamento a seguir:

2.1.1 Planejamento

No planejamento da SLR, executamos as etapas de delineação do objetivo, nomeação dos termos PICOC [25], estabelecimento das questões de pesquisa, seleção das palavras-chave e seus sinônimos, formação da *string* de busca, destaque das fontes de pesquisa, definição dos critérios de inclusão e exclusão dos estudos pesquisados e triagem das perguntas para avaliação da qualidade de cada estudo encontrado nas bases digitais.

- **Objetivo geral da revisão**

  O primeiro passo para a conceitualização da SLR é a descrição de um objetivo principal da revisão a ser arquitetada, por isso ele deve ser descrito com clareza e objetividade. O objetivo geral da presente SLR é encontrar na literatura possíveis

[1] `https://parsif.al/`

métodos, técnicas ou ferramentas de RE, que podem ser utilizados em um ambiente real e auxiliar no processo de desenvolvimento de projetos de software de BD.

- **Delineação dos termos PICOC**

  A categoria a seguir delimita o escopo do trabalho e seus objetivos principais. O enquadramento dos parâmetros é fundamental para a construção da *string* de busca, da demarcação das questões de pesquisa e dos objetivos principais da revisão.

  - **Population** (população) : estudos na área de Big Data;
  - **Intervention** (intervenção): estudos os quais contém engenharia de requisitos;
  - **Comparasion** (comparação): revisão sistemática da literatura ou mapeamento sistemático;
  - **Outcomes** (resultados): métodos, técnicas ou ferramentas;
  - **Context** (contexto): pesquisa acadêmica.

  Embora utilizar os termos para a comparação na definição do PICOC não seja comum em pesquisas na área da engenharia de software [26], o seu uso foi aplicado com o objetivo de identificar estudos comparativos similares, reconhecidos e aderentes ao tema. Também possibilitou limitar a quantidade de estudos encontrados nas bases após a formação da *string* de busca, identificar palavras e sinônimos relacionados. Ao observar diferentes revisões (SLR) e mapeamentos (SMS) efetuados nos estudos apresentados na população indicada, foi possível estabelecer estratégias mais concisas, facilitar o processo de pesquisa e diminuir o tempo necessário para a completude da tarefa. Além disso, ao adicionar os termos de comparação na *string*, resultados mais significantes foram encontrados nas bases definidas, diminuindo a quantidade total de milhares para centenas de estudos a serem avaliados por apenas um pesquisador.

- **Questões de pesquisa**

  As seguintes definições das questões de pesquisa possibilitaram a constatação se serão respondidas após o término da SLR.

  - **RQ.1**: Quais são as abordagens de RE no contexto de Big Data existente na literatura?
  - **RQ.2**: Quais são os métodos, técnicas e ferramentas de RE no contexto de Big Data existente na literatura?
  - **RQ.3**: Quais técnicas e ferramentas da RE no contexto de BD são utilizadas nas instituições financeiras?

- **Palavras-chave e formação da *string* de busca**

  Após a delimitação dos termos PICOC e a numeração das principais questões de pesquisa é efetuado o detalhamento das palavras ou termos necessários para compor a *string* de busca, que é gerada automaticamente pela ferramenta *Parsifal*. A Tabela 2.1 apresenta as palavras-chave e sinônimos definidos para compor a *string* de busca.

  Tabela 2.1: Tabela para palavras-chave e sinônimos para formação da *string*.

Palavras-chave	**Sinônimos**
Big Data	Big Data Analytics, Big Data Applications, Big Data Projects, Big Data Software, Big Data System
Requirement	Requirement Engineering, Requirements Elicitation, Requirements Gathering, Requirements Software
Tools	Methods, Techniques
systematic literature review	systematic mapping

  String de busca gerada :

  (“Big Data” OR “Big Data Analytics” OR “Big Data Applications” OR “Big Data Projects” OR “Big Data Software” OR “Big Data System”) AND (“Requirements” OR “Requirements Elicitation” OR “Requirements Engineering” OR “Requirements Gathering” OR “Requirements Software”) AND (“systematic literature review” OR “systematic mapping”) AND (“Methods” OR “Techniques” OR “Tools”)

- **Fontes principais de pesquisa**

  Com a *string* de busca definida, foi possível efetuar consulta automatizada nas quatro bases das seguintes fontes de pesquisa escolhidas: 1)ACM Digital Library[2], 2)DBLP computer science bibliography[3], 3)IEEE Xplore[4], 4) Scopus[5]. No Google Scholar[6], considerada a quinta fonte de pesquisa, foi efetuada busca manual, utilizando a lista de referências ou citações dos estudos para identificar trabalhos

[2] `https://dl.acm.org/`
[3] `https://dblp.org/`
[4] `https://ieeexplore.ieee.org/Xplore/home.jsp`
[5] `https://www.elsevier.com/pt-br/solutions/scopus/`
[6] `https://scholar.google.com/`

a considerar. Estas fontes foram selecionadas devido ao considerável volume de publicações em congressos e periódicos dentro do contexto da área do objeto da pesquisa.

- **Critérios para seleção do estudo**

  Os critérios de inclusão definem as características que um estudo deve conter para ser considerado relevante à pesquisa, enquanto que os critérios de exclusão estabelecem características para se excluir aqueles irrelevantes ao contexto definido. Assim, foram definidos os seguintes critérios de inclusão e exclusão nesta pesquisa:

  - Categoria de inclusão:
    - Estudo que trata de pelo menos uma das questões de pesquisa (RQ.1 ou RQ.2).
  - Categoria de exclusão:
    - Estudo fora do escopo.
    - Estudo publicado antes de 2012.

- **Perguntas para a avaliação da qualidade dos estudos**

  Como processo complementar, seguindo as recomendações propostas por Kitchenham e Charters [24], foram definidas novas questões para avaliar a qualidade dos estudos selecionados após aplicação dos critérios de inclusão e exclusão na etapa anterior, visando obter mais um filtro. Definiu-se também uma pontuação para cada resposta às questões para avaliar a qualidade a serem aplicadas por estudos. Estas questões para avaliar a qualidade e as respostas com a respectiva pontuação são:

  - Questões de qualidade (QQ):
    - QQ.1: Os autores descrevem a limitação do estudo?
    - QQ.2: Os objetivos estão claramente descritos?
    - QQ.3: Os métodos, técnicas ou ferramentas de RE propostos para projetos de software de BD estão claramente definidos?
  - Respostas:
    - Sim (1.0 ponto)
    - Parcialmente (0.5 ponto)
    - Não (0.0 ponto)

2.1.2 Condução da Revisão Literária e seus Resultados

A fase de condução da revisão foi realizada utilizando a *string* de busca gerada originalmente, conforme descrito no Subitem 2.1.1: **("Big Data" OR "Big Data Analytics" OR "Big Data Applications" OR "Big Data Projects" OR "Big Data Software" OR "Big Data System") AND ("Requirements" OR "Requirements Elicitation" OR "Requirements Engineering" OR "Requirements Gathering" OR "Requirements Software") AND ("systematic literature review" OR "systematic mapping") AND ("Methods" OR "Techniques" OR "Tools")** para consulta nas bases: 1)ACM Digital Library, 2)DBLP computer science bibliography, 3)IEEE Xplore, 4) Scopus. Porém, como não apresentou resultado na base DBLP foi aplicada a seguinte *string* reduzida, **("requirement big data")**. A Tabela 2.2 apresenta a quantidade de estudos encontrados por meio da utilização das *strings* de busca em cada base literária.

Tabela 2.2: Tabela das fontes de pesquisa utilizadas.

Base Literária	**Número de estudos**
ACM Digital Library	202
DBLP computer science bibliography	76
Scopus	33
IEEE Xplore	3
Manual	7

Conforme apresentado na Figura 2.1, que descreve as etapas da condução desta pesquisa, 314 (trezentos e quatorze) estudos foi o resultado inicial obtido com os termos definidos pela *string* de pesquisa, sendo que 6 (seis) estudos duplicados foram excluídos automaticamente pela ferramenta *Parsifal*. Os referidos estudos foram salvos em um arquivo tipo "bibtex" e importados para a referida ferramenta. As buscas foram executadas em dezembro de 2022.

O título, abstract e palavras-chave dos estudos desta seleção inicial foram lidos com o objetivo de aplicar os critérios de inclusão e exclusão. A ferramenta registra o resultado desta seleção por estudo. Assim, foram excluídos 292 (duzentos e noventa e dois) estudos, restando 22 (vinte e dois), com a adição de mais 7 (sete) estudos incluídos por via manual para análise integral do texto e aplicação dos critérios de avaliação da qualidade totalizando 29 (vinte e nove) estudos.

Muitos estudos da etapa inicial de seleção foram excluídos pelo critério de exclusão "fora do escopo" por estarem direcionados às tecnologias capazes de lidar com grandes volumes de dados, tais como, implantação do ecossistema Big Data com suas ferramentas de análise, visualização, integração com sistemas, hardware, rede, infraestrutura, conexões de aplicativos, arquitetura, padrões de segurança da informação e privacidade, uso de "Blockchain", Pipeline de BD, Nuvem, inclusive *frameworks* de automação da cadeia de produção industrial, agricultura e pecuária, medicina e outros domínios de negócios, visando à comunicação entre máquinas e sensores.

A ferramenta também registra o resultado dos critérios de avaliação da qualidade aplicado nos 29 (vinte e nove) estudos, classificados anteriormente pelo critério de inclusão e exclusão. Nesta etapa, foi efetuada a leitura integral dos estudos. O critério de seleção final considerou o valor superior a 2.0, obtido do somatório da pontuação de cada resposta às três questões formuladas por estudo. Dezoito (18) estudos não obtiveram o valor total da pontuação definida. A Tabela 2.3 apresenta os restantes 11 (onze) estudos selecionados nesta etapa final, título e ano de publicação.

Outrossim a Tabela 2.4 apresenta o detalhamento com o somatório da pontuação total, de acordo com o atributo qualidade de cada estudo referenciado, contendo o identificador de cada estudo, a pontuação obtida em cada uma das 3 (três) questões definidas, bem como o total acumulado. Observa-se que todos os estudos totalizaram a pontuação máxima, 3.0 (três) pontos, com exceção do E10[6] que ficou com 2.5 (dois e meio), devido às limitações estarem restritas as escolhas das bases de dados dos estudos relacionados. O referido estudo não apresenta seção específica sobre limitações ou ameaças à validade, por isso, a sua pontuação final reflete tal fator.

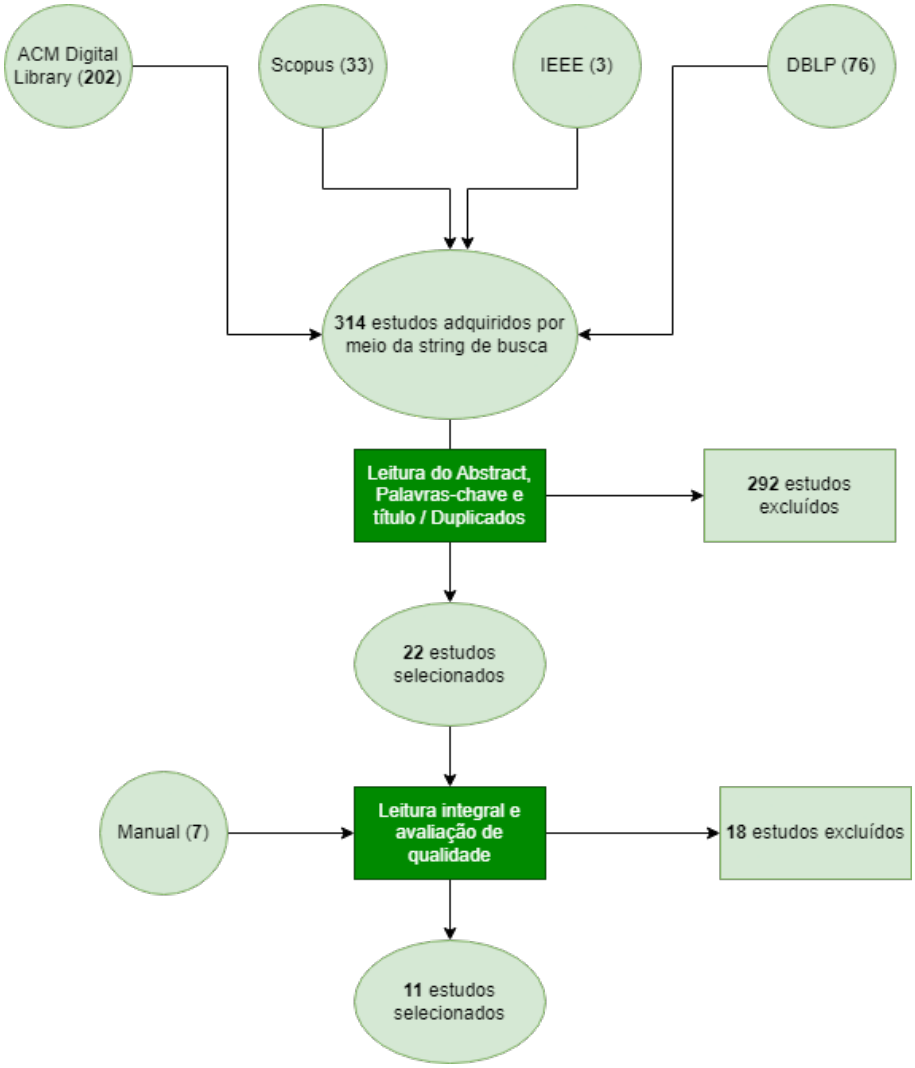


Figura 2.1: Fluxo de condução da SLR do autor deste estudo.

Tabela 2.3: Tabela para destaque dos estudos selecionados.

ID	**Estudo**	**Ano**
E1	Holistic data-driven requirements elicitation in the big data era[1]	2022
E2	Management of Implicit Requirements Data in Large SRS Documents: Taxonomy and Techniques[27]	2022
E3	Data-Driven Requirements Elicitation: A Systematic Literature Review[22]	2021
E4	A Big Data Conceptual Model to Improve Quality of Business Analytics[2]	2020

Tabela 2.3 (continuação):

ID	Estudo	Ano
E5	REBD: a conceptual framework for Big Data requirements[3]	2020
E6	Requirements Engineering Practices and Challenges in the Context of Big Data Software Development Projects: Early Insights from a Case Study[12]	2020
E7	Systematic Mapping Study of Non-Functional Requirements in Big Data System[28]	2020
E8	BiDaML in Practice: Collaborative Modeling of Big Data Analytics Application Requirements[4]	2020
E9	A Validation Study of a Requirements Engineering Artefact Model for Big Data Software Development Projects[5]	2019
E10	Perspectives of Information Requirements Analysis in Big Data Projects[6]	2019
E11	State of Requirements Engineering Research in the Context of Big Data Applications[17]	2018

Tabela 2.4: Tabela para pontuação dos estudos selecionados.

ID	QQ.1	QQ.2	QQ.3	Total
E1[1]	1.0	1.0	1.0	3.0
E2[27]	1.0	1.0	1.0	3.0
E3[22]	1.0	1.0	1.0	3.0
E4[2]	1.0	1.0	1.0	3.0
E5[3]	1.0	1.0	1.0	3.0
E6[12]	1.0	1.0	1.0	3.0
E7[28]	1.0	1.0	1.0	3.0

Tabela 2.4 (continuação):

ID	QQ.1	QQ.2	QQ.3	Total
E8[4]	1.0	1.0	1.0	3.0
E9[5]	1.0	1.0	1.0	3.0
E10[6]	0.5	1.0	1.0	2.5
E11[17]	1.0	1.0	1.0	3.0

2.2 Descrição dos Estudos Selecionados

A seção a seguir apresentará os estudos selecionados na literatura, após a realização da SLR, propostos para resolver questões RQ.1, RQ.2 e RQ.3.

2.2.1 E1 - Holistic data-driven requirements elicitation in the big data era

O E1[1] propõe um *framework*, um metamodelo conceitual e um método, para aquisição, análise e agregação contínua e automatizada de fontes digitais heterogêneas que visa apoiar a elicitação e gerenciamento de requisitos orientados por dados. O modelo mencionado foi construído para possibilitar interativa e automaticamente a realização de coleta, análise e agregação de dados, além do mapeamento dos requisitos de maneira semiautomática com o objetivo de gerar os artefatos dos requisitos a serem utilizados no sistema de análise de dados.

Conforme apresentado na Figura 2.2 o metamodelo proposto distingue: uma classificação de fontes de dados (entidades da cor amarela), os elementos para processar os dados (entidades da cor cinza), os elementos para agregação de dados (entidades da cor azeitona) e mapeamento adicional para requisitos e uma conceituação do artefato de requisitos e elementos relacionados (entidades de cor branca).

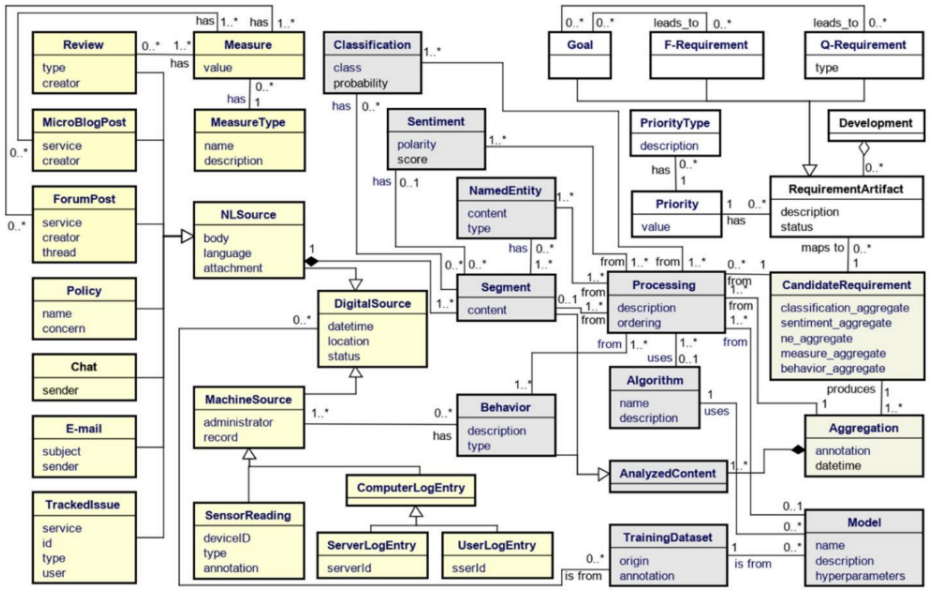


Figura 2.2: Metamodelo para processar e agregar dados de fontes digitais e mapeá-los para artefatos de requisitos novos ou existentes. A cor amarela representa diferentes fontes, cinza para processamento de dados, azeitona para agregação e mapeamento e branca para requisitos proposto por E1[1].

O processo de coletar dados digitais e mapear os dados aos requisitos, com base na conceituação definida pelos autores, e conforme a Figura 2.3, possui quatro principais atividades no processo que são:

- A coleta de dados de diferentes fontes digitais;
- A análise por diferentes meios de processamento de dados;
- A agregação de dados analisados de uma ou mais fontes em requisitos candidatos;
- O mapeamento dos requisitos candidatos para artefatos de requisitos novos e existentes.

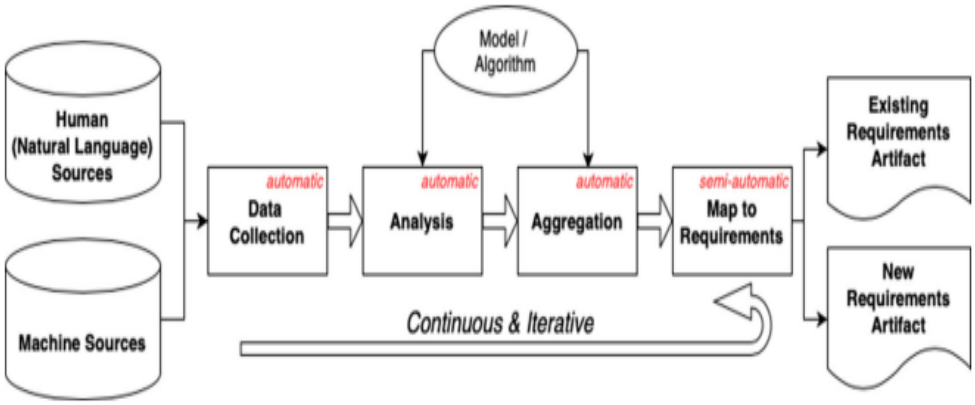


Figura 2.3: Processo de Gerenciamento dos dados propostos por E1[1].

As entradas para a coleta de dados são várias fontes de dados potencialmente heterogêneas, geradas por humanos e máquinas. A saída do processo pode ser um novo artefato de requisitos ou uma mudança em um artefato de requisitos existente. O processo é automatizado para coleta e processamento de dados e, portanto, ocorre continuamente, enquanto alguma intervenção manual, particularmente nas fases posteriores, pode ser necessária. Descrição de cada uma das atividades:

- **Coleta de dados** - A primeira tarefa do processo é identificar quais fontes de dados são relevantes e que devem ser consideradas para a elicitação de requisitos. O metamodelo proposto orienta a classificação das fontes entre seus diferentes tipos:

  – As fontes geradas por humanos (*NLSource*), onde os dados estão principalmente na forma de linguagem natural, incluindo *E-mail*, *Review*, *MicroblogPost*, *ForumPost*, *Chat*, etc.

  – As fontes geradas por máquina podem ser dados de sensores (*SensorReading*) ou *logs* de computador de vários tipos: *ServerLogEntry* ou *UserLogEntry*.

  Cada um desses tipos de origem contém alguns metadados.

- **Análise** - O objetivo desta tarefa é fornecer ferramentas combinadas ou desenvolvidas de processamento de dados de uma forma bruta para uma forma estruturada, de modo que informações ou comportamentos relevantes podem ser identificados. A análise totalmente automatizada é fundamental porque, para a maioria das fontes, os dados são gerados em volumes e velocidades tão altos que não é viável analisá-lo manualmente, resultando na perda de dados potencialmente valiosos. Com base na classificação de tipos de fonte de dados diferentes tarefas são necessárias para processar e analisar os dados coletados.

- **Agregação** - Nesta etapa do processo, as semelhanças nos dados obtidas de diferentes fontes digitais são detectadas e agrupados em uma Agregação que é identificada por uma anotação (numérica ou simbólica). Uma Agregação pode ser composta de dados segmentados de comentários e postagens de microblog, bem como o comportamento obtido dos *logs* do usuário, todos relativos à mesma experiência, como falta de uma funcionalidade específica. Quanto mais fontes são usadas e quanto mais eles diferem, mais difícil é encontrar e agregar aqueles que dizem respeito ao mesmo requisito; no entanto, isso pode ser facilitado por meio da automação e do uso de técnicas baseadas em *Machine Learning* (ML).
- **Mapear para requisitos** - Uma vez que um novo requisito candidato é instanciado, podem ser criados os artefatos do tipo *Goal*, *FunctionalRequirement* ou *QualityRequirement*, etc. A primeira tarefa é entender corretamente se as intenções estão inteiramente relacionadas ao conteúdo do requisito candidato, que em alguns casos também exigem a revisão do conteúdo de *AnalyzedContent*, ou seja, Segmento (para *NLSource*) e Comportamento (para *MachineSource*).

A tarefa é então decidir se o requisito candidato pertence a um artefato de requisitos existentes: se o conteúdo do requisito candidato é avaliado como totalmente novo ou relacionado a alguma mudança de algum artefato de requisitos existentes. Então um objetivo ou um conteúdo de requisito funcional ou de qualidade pode ser criado.

Isso depende do atributo *classification_aggregate* aplicado na etapa de análise usando Classificação. Todo o passo precisa ser apoiado e revisado pelo engenheiro de requisitos, enquanto algum processamento pelo suporte de modelos de ML e algoritmos podem aumentar a automação dessas tarefas. Finalmente, a abordagem de RE em uso influenciará como o engenheiro de requisito ou equipe de desenvolvimento finalizará o conteúdo do artefato de requisitos, incluindo a discussão possivelmente necessária da decomposição, por exemplo, de um objetivo (ou seja, semelhante ao épico no abordagem ágil) em direção a um ou mais requisitos funcionais e/ou de qualidade.

2.2.2 E2 - Management of Implicit Requirements Data in Large SRS Documents: Taxonomy and Techniques

O E2[27] apresentou a importância da identificação dos requisitos implícitos (IMR), também chamados de requisitos ocultos, ausentes, vagos, ambíguos ou derivados em documentos de Especificações de Requisitos de Software de BD, para evitar falhas na concepção do projeto, obter qualidade dos dados, veracidade e relevância na recuperação de informações ocultas, descoberta de conhecimento nas fontes de dados heterogêneas, incluindo textos, imagens, tabelas e outros infográficos.

Os autores mencionam sobre o impacto dos dados de requisitos ocultos na qualidade e no desenvolvimento de software publicada pelo Projeto *Naming the Pain in Requirements Engineering(NaPiRE)*[7], responsável por congregar e publicar semestralmente pesquisas globais sobre problemas práticos em RE. A falta de experiência e conhecimento de práticas de elicitação de RE, utilização inadequada de técnicas de gerenciamento e integridade de dados, e a indisponibilidade de ferramentas de RE, foram apontadas como prováveis causas dos requisitos ocultos na referida pesquisa.

Apresenta uma taxonomia de termos, definições e exemplos associados aos IMR, abrangendo principalmente as categorias de segurança (identificação e autenticação, disponibilidade, responsabilidade e privacidade); acessibilidade aos usuários com deficiência (perceptível, operável, compreensível e robusto); manutenibilidade ou a capacidade de ser adaptado ou modificado (analisabilidade, mutabilidade, estabilidade e rastreabilidade), sustentabilidade (econômica, técnica, social, individual e ambiental) e usabilidade, os relacionados à interface do usuário (recursos de *feedback*, recursos de desfazer, recursos de cancelamento, validação de formulário ou campo, requisitos do assistente, recursos de experiência do usuário, idiomas diferentes e recursos de alerta).Detalham e comparam o uso das ferramentas dos estudos objetos da SLR com as respectivas técnicas informadas para identificação de IMR, quais sejam:

- **IPT** - *Implicit Priming Test*: utiliza ontologia, semântica e práticas de garantia da qualidade;
- **IMR** - *Architectural Framework*: utiliza raciocínio baseado em analogia;
- **The PROMIRAR Tool**: utiliza raciocínio baseado em analogia;
- **Using Templates for IMR Data Detection: Templates** – utiliza modelos de artefatos para comparação;
- **InfoVis: IMR Data Visualization** & **PLN - Processamento de Linguagem Natural**: utiliza PLN associada com técnicas similares de semântica;
- **Machine Learning Classification techniques**: utiliza ML Supervisionado, Semi-supervisionado e sem supervisão. Informa que o uso de ML geralmente atinge mais de 70% de precisão ao identificar e classificar os requisitos não funcionais.

Os autores sugerem, por meio da ontologia, explorar se padrões como RDF (*Resource Description Format*) e OWL (*Web Ontology Language*) seriam úteis em tarefas de especificação de IMR, possibilitando o compartilhamento de dados globalmente, assim como

[7] `http://napire.org/#/home`

a criação de bases de conhecimento de domínio específico, tais como na área de finanças, de saúde.

Como abordagem futura, sugere também que técnicas de ciência de dados, como mineração de regras de associação, poderiam ser utilizadas na descoberta da relação entre a fonte real do erro que causa o defeito na fase de RE, contribuindo com a gestão de IMR e na qualidade das especificações de requisitos.

Destaque para limitação da pesquisa que atende muito mais ao modelo de desenvolvimento em cascata "Waterfall", não abrangendo a identificação de requisitos ocultos de histórias de usuários.

2.2.3 E3 - Data-Driven Requirements Elicitation: A Systematic Literature Review

O E3[22] tratou de uma SLR do estudo E1[1] sobre a elicitação automática e contínua de requisitos baseada em dados de fontes digitais dinâmicas não coletados intencionalmente, com grande volume, para fins de elicitação de requisitos de BD. Estes tipos de fontes de dados possibilitam a descoberta de dados relevantes para novos requisitos do sistema; permite capturar os requisitos atualizados do usuário em tempo real, criando melhorias ou novas oportunidades de negócio, são legíveis por máquina, facilitando a engenharia de requisitos automatizada, contínua e escalável. As referidas fontes de dados digitais são caracterizadas pelos seguintes tipos:

- **Fonte de dados de origem humana** representa os registros digitalizados de experiências humanas em mídias, por exemplo: Avaliações *online* (avaliações públicas *online* de clientes que adquiriram produtos ou serviços); Microblogs (um tipo de blog publicado em sites de mídia social, onde os usuários postam mensagens em diferentes formatos como textos curtos, áudio, vídeo e imagens); Fóruns de discussões *online* (sites que permitem a postagem de mensagens entre as pessoas para troca de conhecimento); Repositórios de Software (plataformas para compartilhamento de pacotes de software ou códigos-fonte, que contêm principalmente três elementos: tronco, ramificações e *tags*; inclui-se também os relatórios detalhados de *bugs* ou reclamações escritas na forma de textos livres); Listas de Discussão (é um tipo de Fórum de discussão, onde mensagens de *e-mail* são encaminhadas pelos assinantes e compartilhadas numa lista de discussão).
- **Fontes de dados mediadas por processos** se referem aos registros do monitoramento de processos e eventos de negócios, por exemplo, transações comerciais, registros bancários, pagamentos com cartão de crédito.

- **Fontes de dados geradas por máquina** são os registros de sensores de máquinas que são usadas para medir eventos e situações físicas, por exemplo, leituras de sensores de pressão barométrica e ambiental, saídas de dispositivos médicos, dados de imagens de satélite e dados de localização, leituras de chip RFID (Radio-Frequency Identification) e saídas de GPS (Global Positioning System).

Com base nos resultados da SLR, a maioria dos estudos concentrou-se nas fontes dos dados de origem humana (na forma de Avaliações *online*, Microblogs, Fóruns de discussões *online*, Repositórios de Software e Listas de Discussão), poucos trabalhos consideraram as fontes de dados mediadas por processos e geradas por máquinas. Segundo os autores, isto é devido ao volume de fontes de dados em linguagem natural, disponível e acessível publicamente, utilizada pelos usuários para expressar suas preferências e necessidades, facilitando a obtenção dos requisitos de software.

A grande maioria dos estudos coletaram os dados dinâmicos externos à organização. Os autores destacaram a importância dos engenheiros de requisitos identificarem os requisitos originários destas potenciais fontes externas que facilitam a evolução ou desenvolva novas oportunidades de negócio.

Quanto às técnicas para elicitação automatizada de requisitos, utilizaram-se algoritmos categorizados: em aprendizado de máquina (ML), classificação baseada em regras, abordagem orientada a modelos, modelagem de tópicos e agrupamento tradicional. A maioria dos estudos utilizou Processamento de Linguagem Natural (PLN) como técnica para processamento de dados; técnica de ML para classificação e agrupamento, e poucos estudos conseguiram obter requisitos de alto nível e com aplicação no mundo real. Como conclusão, a elicitação automatizada de requisitos, apresentada nos estudos objeto da SLR, propõe identificar e classificar as informações relacionadas a requisitos ou apenas obter a identificação de recursos de requisitos. Faltam métodos para apoiar a elicitação de requisitos de fontes de dados heterogêneas.

2.2.4 E4 - A Big Data Conceptual Model to Improve Quality of Business Analytics

O E4[2] apresentou o modelo IRIS, um modelo conceitual de BD, baseado em uma ontologia com conceitos como problemas de negócios, soluções, orientada aos objetivos da organização, para identificar os tipos e fontes de dados externos apropriados, racionalizando sua seleção, com base na característica de "variedade" de BD, com foco na modelagem de BD e na sua qualidade.

O IRIS utiliza o modelo Extended Entity-Relationship (EER) para representar cada conceito como uma meta, podendo ser dividida em submetas específicas com os respec-

tivos relacionamentos (representada no modelo como uma entidade "Relacionamento de Refinamento"), agregadas à entidade raiz denominada "Business Concept" com relacionamentos pai/filho. A ideia é criar um modelo de *Big Data Virtual* para modelar se o Problema e a Solução, tratado como uma visão do "Business Concept", contribuem positivamente ou negativamente para atingir uma meta. A entidade "Insight" é representada como um objetivo a atingir, que inclui "Problema" e "Solução", que são validados pelos resultados de *Big Data Queries* ou KPI (*Key Performance Indicator*).

Para avaliar a qualidade da modelagem de dados, que está fortemente relacionada à qualidade dos dados que podem ser tratadas num nível conceitual, o IRIS propõe aplicar três atributos de qualidade, considerando a Variedade dos dados de BD: "Relevância", isto é se os dados e relacionamentos entre os dados são relevantes, descartando os que não são; "Abrangência" por considerar a variedade de diferentes tipos e fontes de dados; e "Prioridade Relativa" para determinar a relação entre a quantidade limitada de recursos e volume de dados ou a velocidade de processamento.

Existem muitas dimensões relacionadas aos dados em BD, tais como estruturadas/não estruturadas ou descritivas/preditivas. A título de exemplo, o estudo utilizou três dimensões: Dimensão Interna (dado no data center da organização)/ Externa à organização (dados no site de rede social, por exemplo); Dimensão *Offline*/*online*; e Dimensão Adequada (dados comuns de primeira ordem, por exemplo histórico de vendas) /Analítica (dados de segunda ordem, isto é os obtidos por meio de análises dos dados comuns , por exemplo tendência de vendas). Para ajudar a capturar e utilizar dados de várias fontes e tipos é efetuado um cruzamento entre essas dimensões, representadas assim por oito entidades no modelo de BD do IRIS.

Quanto à Relevância, considera Estruturalmente Relevante quanto mais ligação houver entre um elemento do modelo de dados e algum problema ou solução. Também se considera Relevância Semântica quanto mais positiva for a contribuição de um modelo de dados (elemento) para a validação de um problema ou solução. Quanto à Prioridade, deve refletir as prioridades que os dados se destinam no tratamento dos Problemas e Soluções.

As três dimensões organizacionais a seguir auxiliam estruturar a variedade, o volume e alta velocidade dos dados: Dimensão Classificação/ Instanciação (serve para relacionar instâncias a classes); Dimensão Generalização/ Especialização (relaciona dados por meio de relacionamentos de subclasse ou superclasse); e Dimensão Agregação/ Decomposição (permite associar dados de diferentes classes). O objetivo é integrar entidades de dados existentes com as novas entidades externas de BD. É apropriado utilizar essas três dimensões para explorar e relacionar dados na mesma ou em diferentes dimensões de abrangência de BD.

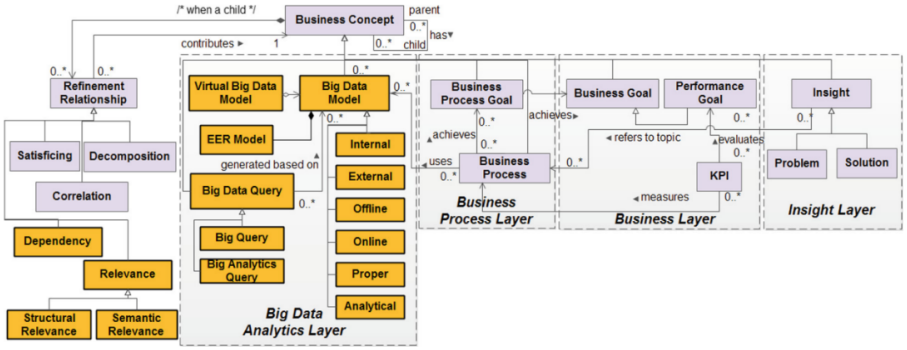


Figura 2.4: Modelo da ferramenta IRIS proposto por E4[2].

2.2.5 E5 - REBD: a conceptual framework for Big Data requirements

O E5[3] apresentou um modelo que busca organizar o processo de RE na produção de um sistema de BD com a finalidade de conciliar e integrar os processos e técnicas tradicionais de RE com as atividades centradas em dados, permitindo extrair, nas diversas e volumosas informações, os padrões ocultos de dados valiosos para a organização. Além disso, o REBD destaca a necessidade de atuação conjunta, em ambiente colaborativo, do engenheiro de requisitos de RE e o cientista de dados de BD. Portanto, o REBD propõe os seguintes processos [20]:

- **Elicitação de Requisitos** - a atuação conjunta se inicia logo nesta etapa, permitindo ao cientista de dados de BD identificar, nos requisitos do usuário, os dados que podem ser usados para gerar valores de negócios e, posteriormente, aprimorar os requisitos do usuário. Considerada como uma atividade crítica onde a comunicação é primordial para se descobrir e reunir as necessidades e desejos dos *stakeholders* transformados em requisitos funcionais e não funcionais, sugere-se utilizar múltiplas técnicas para sua obtenção como, por exemplo, entrevistas, prototipagem, análise de domínio, etc. O produto de trabalho deste processo é a: Lista de Requisitos;
- **Aquisição de Dados** - é efetuada pelo cientista de dados de BD, que tem como produto de trabalho dados filtrados usados para descoberta de conhecimento, por meio de coleta dos diversos tipos de dados ou diferentes métodos de aquisição. A Análise do que deve ser descartado e mantido envolve alto custo de conversão dos formatos dos dados, além das características particulares do 5V's de BD. Exemplos de métodos de aquisição de dados: método de coleta de log do sistema, método de coleta de dados não estruturados da rede, dentre outros;

- **Análise de Requisitos** - consiste em definir os limites do sistema e o seu ambiente de interação, resolver os conflitos entre os requisitos, manter sua clareza, consistência, completude. Esse processo é realizado em paralelo com o processo de encontrar valores de negócios e aquisição de dados pelo engenheiro de requisitos de RE. Produto de Trabalho: Especificações técnicas e modelos de requisitos;
- **Análise de Dados** - os dados adquiridos são analisados na busca e na descoberta de conhecimento por meio de múltiplos e variados processos e etapas analíticas efetuadas pelo cientista de dados de BD, que também é responsável pela seleção das técnicas de análise de dados com o objetivo de acelerar e evitar falhas no processo decisório. Exemplo destas técnicas: Aprendizado de máquina, mineração de dados, análise de texto, análise preditiva, processamento de linguagem natural, mineração de processos e ferramentas como redes inteligentes. O produto de trabalho deste processo expressa os valores extraídos dos dados e modelos de análise de BD;
- **Consolidação de Casos de Uso** - permite fusão entre os modelos obtidos pelo cientista de dados de BD e os de casos de uso tradicional elaborado pelos engenheiros de requisitos de RE. Como produto de trabalho, este processo apresenta um novo artefato denominado "Diagrama de Caso de Uso Acionável", que permite documentar os valores de negócios dos dados, como estes são obtidos, os algoritmos utilizados, os métodos de análise, os meios que se utilizam para acionar estes valores (inteligência acionável, conhecimento), bem como as suas funções e suas interações com outros componentes de software. O produto de trabalho deste processo é integração dos modelos de requisitos e de análise de BD;
- **Modelagem de Requisitos** - diferente da RE tradicional, a negociação dos requisitos para BD é efetuada após a consolidação dos modelos do engenheiro de requisitos RE e do cientista de dados BD. Os modelos de Requisitos consolidados auxiliam documentar os requisitos durante a etapa de especificação. O produto de trabalho deste processo é "Diagrama de Caso de Uso Acionável" finalizado;
- **Validação de Requisitos** - O engenheiro de requisitos RE busca verificar se os requisitos retratam o que realmente os usuários desejam, se não existem conflitos de requisitos, bem como se o custo e prazo serão atendidos. Exemplos de técnicas que podem ser utilizadas nesta validação: Análise manual sistemática dos requisitos, geração de casos de teste, ferramentas de avaliação comparativa de produtos ou algumas das técnicas de elicitação de requisitos. O respectivo produto de trabalho é Requisitos validados;

- **Especificação de Requisitos** - os requisitos do usuário e do software são documentados utilizando o "Diagrama de Caso de Uso Acionável" para especificar e documentar os valores dos dados, isto é, os 5V's das características de BD, consideradas como requisitos de qualidade, tais como: desempenho do sistema, confiabilidade, disponibilidade. Assim, para se garantir que as característica dos 5V's foram consideradas durante o processo de RE de BD, faz parte da documentação de requisitos, a matriz dos atributos de qualidade, onde se relaciona uma característica de BD com um atributo de qualidade. O produto de trabalho é o Relatório de Especificação de Requisitos.

Conforme acima detalhado o REBD propõe integrar os processos referentes a RE (Elicitação de requisitos, Análise de requisitos, Modelagem de Requisitos, Validação de requisitos e Especificação de Requisitos), com alguns passos a mais para a construção de um projeto de BD (Aquisição de dados, Análise de dados e descoberta de valor e Consolidação de caso de uso), além de permitir documentar e descobrir os valores dos dados.

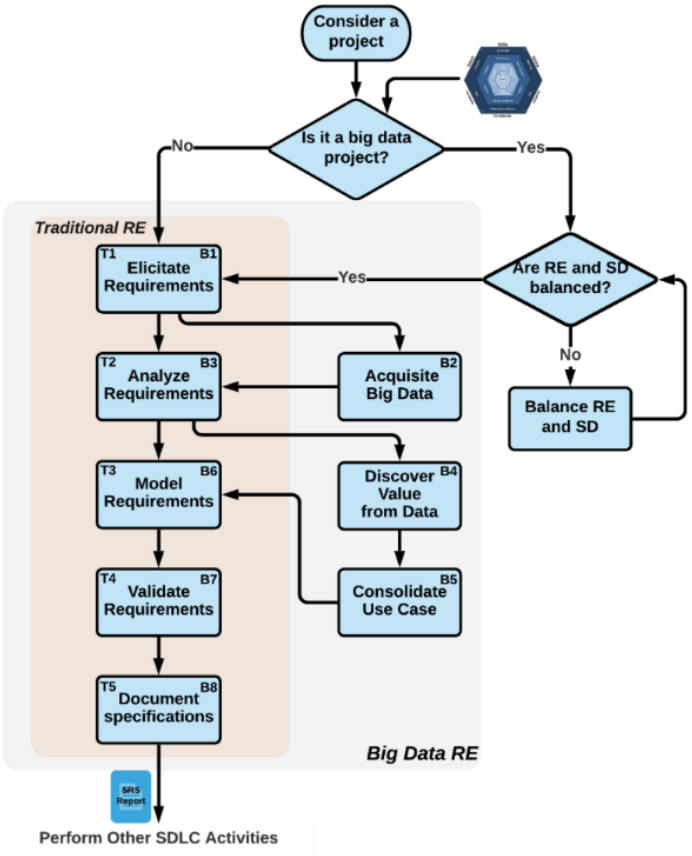


Figura 2.5: Modelo REBD proposto por E5[3].

2.2.6 E6 - Requirements Engineering Practices and Challenges in the Context of Big Data Software Development Projects: Early Insights from a Case Study

O E6[12] apresentou um estudo de caso exploratório sobre um projeto de desenvolvimento de software BD na área de petróleo e gás, para investigar as práticas de RE para elicitar, especificar, analisar e priorizar os requisitos do sistema; as fontes de identificação dos requisitos de BD e sua proporção em relação ao total dos requisitos do projeto; a importância das características e tecnologias de BD na definição dos requisitos de qualidade e na arquitetura do sistema.

A equipe do projeto foi constituída por desenvolvedores, engenheiros de dados, arquitetos de software, analista de negócio (de requisitos). Utilizou-se uma metodologia mista de desenvolvimento: *Sprints* (método ágil) para o planejamento do projeto e para o desenvolvimento se adotou um sequenciamento lógico de atividades, com características de processo progressivo e estruturado, mas que permitia a implementação e a integração de funcionalidades de forma incremental e evolutiva. Quanto ao processo de RE, adotaram-se as atividades padrão de Elicitação, Especificação e Modelagem, Análise e Priorização.

Na Elicitação foram efetuadas reuniões frequentes com os *stakeholders* e representantes dos clientes, além de análise de documentos contratuais. Como suporte na identificação de requisitos, os analistas de negócios elaboraram um modelo de requisitos representando os vários componentes do sistema e suas interações com outros sistemas, fontes de dados e usuários.

Na fase de Análise, coube aos analistas de negócios examinarem se os requisitos documentados refletiam as expectativas do usuário, realizando inspeções manuais e visuais, apresentando-os aos gerentes de projeto, que mantinha toda a equipe atualizada, para que as alterações fossem refletidas nos requisitos associados e nos trabalhos em andamento.

Na Especificação e Modelagem, utilizou-se casos de uso (UML) para descrição textual e alguns requisitos do projeto foram documentados de forma gráfica, assim como para a modelagem de fluxos de dados, modelo de dados e modelos conceituais e arquitetura, foi utilizando a ferramenta Visual Paradigm[8] (ferramenta UML CASE que suporta UML, SysML e *Business Process Modeling Notation*). Para os requisitos de interface de usuário e camada de apresentação foram utilizadas ferramenta de *mock-up* na criação de protótipo.

Na priorização, são identificados os requisitos relevantes e importantes, conforme critérios estabelecidos pelo projeto. No caso do projeto em estudo, foram priorizados os requisitos de acordo com as perspectivas dos usuários em detrimento às características específicas de BD ou atributos de qualidade relacionados ao sistema.

[8] `https://www.visual-paradigm.com/`

A equipe do projeto utilizou o Google Drawings[9] para desenhar a arquitetura do sistema, diante da ausência de ferramenta específica de modelagem e padrões industriais para *Big Data Systems*, necessitando efetuar adaptações para retratar serviços e fluxo de dados. Algumas das tecnologias de BD foram logo identificadas na fase inicial de elicitação, durante a definição do projeto em razão do volume e variedade dos dados, custo, benefício e facilidade de manutenção. Quanto às características de BD, não foram explicitamente descritas nos casos de uso, mas implicitamente compreendidas. Notações foram inseridas, por exemplo, no protótipo para representar que os dados seriam analisados em tempo real.

O estudo apontou os seguintes desafios: ausência de práticas e ferramentas apropriadas para seleção de requisitos de tecnologia de BD e na qualidade estrutural da aplicação; ausência de ferramentas de modelagem específicas para os requisitos de BD; faltam técnicas de especificação de BD para documentar os requisitos do sistema; ausência de padrões e especificações industriais para arquitetura e requisitos de BD e, por último, falta conhecimento adequado sobre arquiteturas e tecnologias de BD.

2.2.7 E7 - Systematic Mapping Study of Non-Functional Requirements in Big Data System

O E7[28] apresentou investigação na literatura sobre os atributos de qualidade mais adequados na definição da arquitetura de BD, quais sejam: eficiência de desempenho, adequação funcional, confiabilidade, segurança, usabilidade e escalabilidade. Destaca que a arquitetura de software de BD é mais complexa, devido ao uso intensivo de dados, necessitando analisar todos os requisitos não funcionais, que garantam a comunicação e coordenação entre os componentes, conectores e restrições arquiteturais. Como os requisitos não funcionais afetam a arquitetura de software, torna-se necessário identificar estes requisitos antes de se projetar o produto.

Neste trabalho foram selecionados 14 artigos, de 2012 a 2019, que tratam de requisitos não funcionais necessários aos sistemas de BD. Com base nestes requisitos não funcionais, mais de 40 (quarenta) atributos de qualidade diferentes foram relacionados a sistemas BD, de acordo com os termos utilizados nos próprios artigos. Em seguida, foi efetuada correspondência dos termos dos atributos de qualidade com as sub-características definidas ISO/IEC 25010:2011[10]. Para realizar a agregação, o atributo de qualidade bastava ter pelo uma sub-característica correspondente a característica da ISO/IEC 25010:2011.

Como resultado, 100% dos estudos selecionados discutem eficiência de desempenho, traduzindo sua importância para sistemas com uso intensivo de dados; 79% apresen-

[9] `https://chrome.google.com/webstore/detail/google-drawings/mkaakpdehdafacodkgkpghoibnmamcme`
[10] `https://blog.onedaytesting.com.br/iso-iec-25010/`

tam adequação funcional, confiabilidade e segurança; 71% abordam a usabilidade. Além disso, o estudo propõe incluir a característica escalabilidade, mesmo não fazendo parte da ISO/IEC 25010:2011, considerando a flexibilidade, retenção de dados, paralelismo, cobertura de dados e consistência, como sub-características correspondentes, já que 86% dos estudos de pesquisa abordam este atributo, também considerado importante para BD diante do alto volume de armazenamento, processamento de diferentes tipos de dados em tempo real e necessidade de infraestrutura distribuída.

2.2.8 E8 - BiDaML in Practice: Collaborative Modeling of Big Data Analytics Application Requirements

O E8[4] apresentou a *Big Data Analytics Modeling Languages* (BiDaML) é uma ferramenta colaborativa, composta por um conjunto de linguagens visuais de um domínio específico, que utiliza vários tipos de diagramas em diferentes níveis de abstração para dar suporte a soluções de software de BD. Visa apoiar todas as partes interessadas, tais como: especialistas de domínio, analistas/gerentes de negócio que não tem formação em ciência de dados e programação; analistas de dados, cientistas de dados e engenheiros de software que não possuem conhecimento de domínio (negócio); os cientistas de dados que não possuem experiência em engenharia de software. A proposta é possibilitar uma linguagem compreensível entre equipes bem diversas, fornecer suporte na evolução de soluções de software de BD, viabilizar o reuso de soluções existentes.

BiDaML possibilita uma apresentação de alto nível das etapas para capturar, representar e comunicar a análise e projeto de requisitos de negócios, pré-processamento de dados, processo de análise de dados de alto nível, implantação de soluções e visualização de dados. Em uma visão geral, apresenta cinco diagramas com diferentes níveis de abstração cobrindo todo o ciclo de desenvolvimento de software de análise de dados, desde requisito de alto nível e definição do problema até a implantação do produto final, quais sejam:

- Diagrama de *Brainstorming* é definido para cada projeto. Apresenta uma visão geral de alto nível das tarefas e subtarefas no design da solução. Possibilita utilizar a técnica de *brainstorming* interativo para identificar os requisitos, metodologias analíticas e tarefas específicas. Contém ícone para visualizar o problema definido, suas respectivas tarefas associadas, a hierarquia de subtarefas para cada tarefa e as informações dos subsistemas utilizados ou produzidos. Os seguintes agrupamentos de atribuição são definidos na ferramenta, com base na terminologia de construção de um sistema de IA (Inteligência Artificial): Domínio e atividades relacionadas aos negócios (*BusinessOps*); atividades relacionadas a dados (*DataOps*); inteligência

artificial e atividades relacionadas a ML (*AIOps*); e atividades de desenvolvimento e implantação (*DevOps*);

- Diagrama de Processo utiliza anotação BPMN (*Business Process Modeling and Notation*) adaptada para especificar os processos de análise de BD, que servem como suporte ao gerenciamento de processos de negócio, de acordo com o detalhamento apropriado às pessoas envolvidas da organização (diferentes "lanes") ou entre organizações (diferentes "pools"). Também são definidas diferentes camadas para tarefas distintas de negócios (*BusinessOps*), técnicas (*DataOps* e *AIOps*) e tarefas operacionais (*DevOps* e as relacionadas a aplicativos);
- Diagrama de Técnica apresentam o detalhamento particular de baixo nível para diferentes tarefas e subtarefas de análise de BD, ampliando o diagrama de *brainstorming*. Contém as técnicas utilizadas ou planejadas em cada tarefa para resolver cada sub-tarefa e os respectivos resultados (problema ou sucesso);
- Diagrama de Dados documentam os dados, os itens de dados estruturados e semiestruturados envolvidos em diferentes etapas do projeto de análise de dados, o processo de coleta e artefatos que são obtidos em cada um dos diagramas acima descritos, bem como as saídas associadas às diferentes tarefas;
- Diagrama de implantação apresenta os detalhes da implantação, os artefatos de software, os componentes, plataformas, serviços e estruturas de nuvem distribuída.

BiDaML possui um recurso do tipo *drag and drop* que consegue gerar uma documentação a partir do diagrama de *brainstorming*, de técnicas, de dados, de processos ou de implantação selecionado, um relatório de atividades e um template para códigos de acesso a API ou código em python disponíveis para os cientista de dados utilizarem.

O BiDaML foi aplicado e avaliado em três projetos reais de software de BD: um site de previsão de preços de imóveis, outro sobre a integração de aplicativo de BD em plataforma de controle de tráfego rodoviário e o último sobre a obtenção mais precisa de análise radiológica em hospital. É possível acessar a versão web, cadastrar-se e utilizar a ferramenta[11].

[11] `https://bidaml.web.app/`

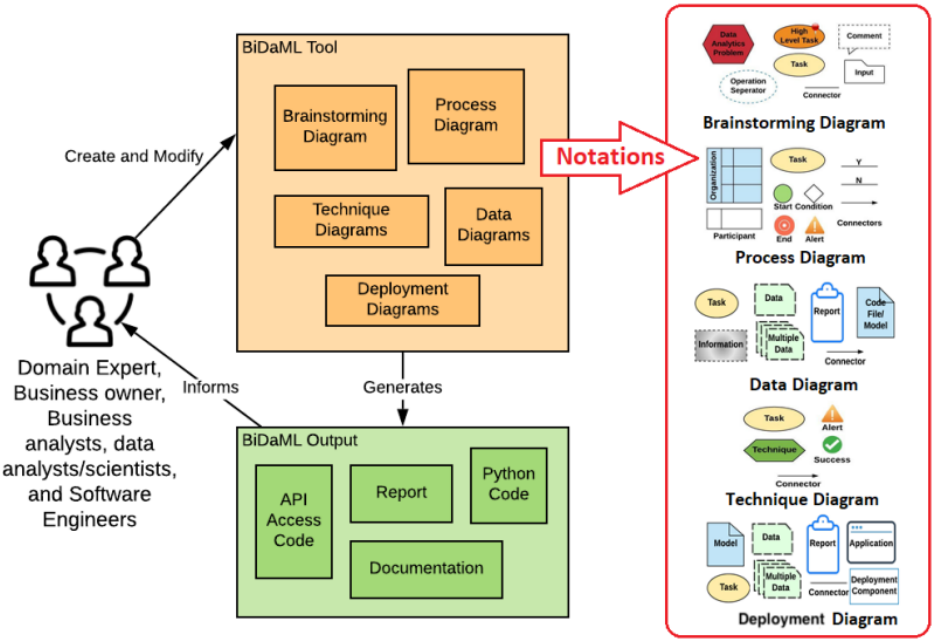


Figura 2.6: Modelo da ferramenta BiDaML proposto por E8[4].

2.2.9 E9 - A Validation Study of a Requirements Engineering Artefact Model for Big Data Software Development Projects

O E9[5] apresentou uma versão validada do Artefato de Modelagem BD-REAM de acordo com a Figura 2.7, efetuada por dez profissionais terceirizados de diversos projetos de desenvolvimento de software de BD. O modelo é orientado a artefatos, com o objetivo fornecer suporte ao levantamento e especificação de requisitos, definição de processos específicos de RE em BD, obter uma visão comum e rastreabilidade que ligam os artefatos. O BD-REAM é composto por três tipos de elementos: artefato (Entidade), associação entre dois artefatos (Relacionamento) e cardinalidade.

O BD-REAM contém 43 (quarenta e três) Entidades (artefatos) agrupadas com base no Modelo de Referência de Engenharia de Requisitos (REM) de Geisberger [29], nas três categorias a seguir:

1)Necessidades de Negócios, “Business Case” (retângulos na cor rosa), congrega 7 (sete) artefatos de necessidades de negócios , que especificam requisitos estratégicos e de clientes, incluindo produtos e objetivos de negócios [29].

2)Especificação de Requisitos ,“Software Requirements Specifications”(retângulos na cor verde), congrega 15 (quinze) artefatos de especificação de requisitos, que contém requisitos funcionais e não funcionais, analisados e modelados, com base nas perspectivas do usuário, e derivados e justificados pelas necessidades do negócio [29].

3)Especificação de Sistemas, “Systems Requirements Specifications” (retângulos na cor azul), congrega 21 (vinte e um) artefatos de especificação de sistemas, que compõem a definição do funcionamento sistema, sua integração, restrições, comportamentos e ambiente.

As 17 (dezessete) Entidades específicas de BD (retângulos com bordas grossas) são:

- Big Data Scenarios;
- Big Data Software Requirements Specifications , que é composta por: Data Processing Requirements Specification, (que contém Data Processing Requirements) e Data Consumer Requirements Specifications (que contém Data Consumer Requirements);
- Data Requirements Specifications (que contém Data Requirements e Data Modeling and Linking details);
- Data Source Requirements Specifications, (que contém Data Source Requirements);
- Technological Requirements Specifications (que contém Data Collection Technological Requirements, Data Storage Technological Requirements, Data Processing Technological Requirements, Data Visualization Technological Requirements, Data Management Technological Requirements).

Exemplos de Entidades:

- Requisitos de infraestrutura (o sistema deve suportar grande armazenamento de dados distribuídos);
- Requisitos de fontes de dados (o sistema deve suportar alta capacidade de transmissão de dados entre fontes de dados e clusters de computação);
- Requisitos de processamento de dados (o sistema deve suportar análises em lote e em tempo real);
- Requisitos de consumidor de dados: o sistema deve suportar diversos formatos de arquivo de saída para visualização.

Exemplos de Relacionamentos:

- “is-derived-from”: quando um ou mais artefatos podem ser derivados de outro artefato (por exemplo, requisitos de qualidade são derivados de cenários de Big Data);
- “Contains”: quando um artefato contém informações sobre outro artefato (por exemplo, Especificação de Requisitos Funcionais contém Requisitos Funcionais);

- “Assist-in”: quando um ou mais artefatos auxiliam na criação de um ou mais outros artefatos (por exemplo, requisitos do sistema na criação da arquitetura do sistema).

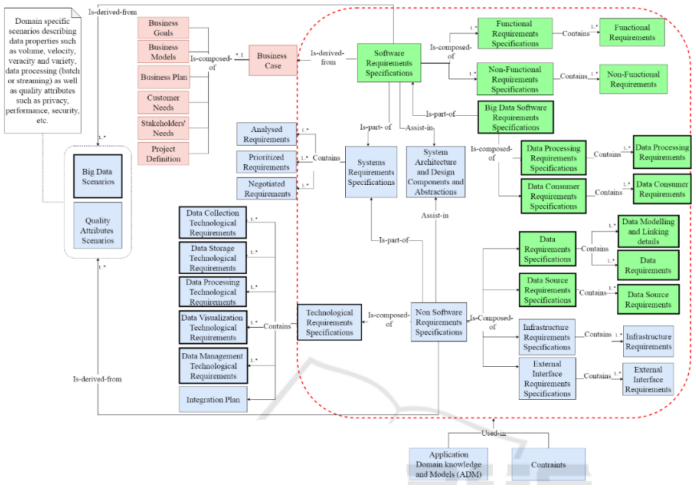


Figura 2.7: Artefato de Modelagem BD-REAM proposto por E9[5].

2.2.10 E10 - Perspectives of Information Requirements Analysis in Big Data Projects

O E10[6] apresentou uma avaliação dos métodos para análise de requisitos em projetos de BD existentes na literatura com a possibilidade de automatizar as fases de engenharia de requisitos, apontando a ausência de estudos de como obter “requisitos de informação” em projetos de BD, processando os dados de forma automática ou semiautomática.

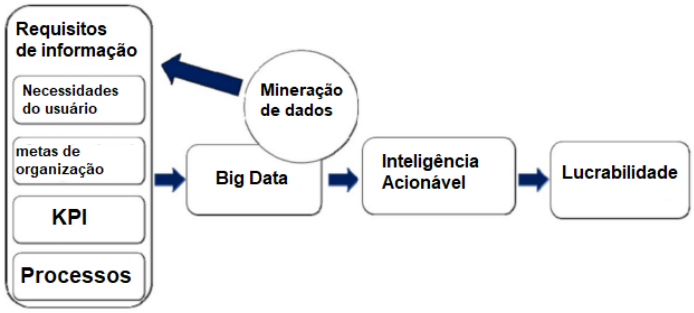


Figura 2.8: Uma visão esquemática da definição de requisitos de informação e elicitação de BD segundo o E10[6].

De acordo com a Figura 2.8, os autores propõem a obtenção destes requisitos previamente ao início do projeto de BD, conforme é feito em soluções de inteligência de negócios, considerando: 1. analisar os processos e metas organizacionais; especificar as necessidades dos usuários, utilizando, por exemplo, histórias de usuários; e 3. alinhar estes requisitos

com as KPI (*Key Performance Indicator*), ou seja conforme os indicadores chaves de desempenho da organização, definido processos de medição apropriados.

Com o carregamento dos dados é possível obter novos requisitos de informação, utilizando técnica de mineração de dados, como, por exemplo, padrões de rastreamento, classificação, regras de associação ou agrupamento, detecção de "*outliers*"[12], bem como técnicas de PLN. O estudo considera carregar os dados disponíveis da organização no HDFS (*Hadoop Distributed Fill System*) como uma prática comum e que pode haver informações ocultas e não óbvias nos dados detectados por processos automatizados.

A proposta visa apoiar as atividades diárias da organização, promovendo o uso intensivo de dados, indicando se estão alinhados aos objetivos e estratégias de negócios. Os autores concluem que a utilização de PLN na obtenção de requisitos não declarados é um método adequado para serem integradas na elicitação de requisitos e seu pós processamento. Propõem, citando outros autores, que os requisitos de informação sejam obtidos de forma iterativa durante o desenvolvimento do projeto, conforme os métodos usuais de RE, e que, durante a análise de dados, quando todos os dados são coletados, novos relacionamentos e valores podem ser explorados e integrados à solução de BD.

2.2.11 E11 - State of Requirements Engineering Research in the Context of Big Data Applications

O E11[17] apresentou uma análise dos estudos acadêmicos sobre RE no contexto de aplicações de software envolvendo BD, buscando identificar: as fases aplicadas do processo de RE, visando avaliar a cobertura do processo de RE; os tipos de requisitos, para verificar se a ênfase está na funcionalidade, na qualidade, nos dados ou no domínio de aplicação; os desafios de pesquisa de RE e propostas de soluções.

Dos 14 (quatorze) estudos selecionados de 2013 a 2017, não foram encontrados detalhes sobre a aplicabilidade dos requisitos de BD para um domínio específico. A maioria abordou as fases de análise ou especificação de RE. Elicitação, modelagem e validação somente foram identificadas em um estudo cada. Não foram encontrados estudos abordando as fases de negociação, priorização e gerenciamento de requisitos em aplicações de BD.

Apenas um estudo apresentou um conjunto de requisitos genéricos para aplicação de BD, com base em descrições de caso de uso (UML), obtidas de diferentes domínios de aplicação. Abordou-se sobre a importância dos requisitos funcionais para o entendimento do projeto, mencionando requisitos funcionais genéricos que qualquer aplicativo de BD deve conter. Quanto aos requisitos de qualidade, destaque para privacidade e segurança; depois desempenho e disponibilidade; por último, escalabilidade, consistência, elasticidade

[12]dados discrepantes, pontos fora do comum, anomalias, valores atípicos

e baixa latência. Quanto aos requisitos de dados, apenas dois estudos apresentaram a necessidade de se obter o tipo certo de dados, considerar as propriedades de dados na fase de elicitação e especificar os requisitos de dados quanto ao seu tamanho, tipos de dados, formatos de arquivos, taxa de crescimento em repouso ou em movimento. Dois modelos diferentes foram apresentados num estudo para apoiar a definição de requisitos de dados: um para a fonte de dados e outro relacionando os problemas de negócios com os dados. Foi proposto um *framework* de especificação de requisitos para coleta de BD.

Como conclusão são apontados como desafios:

1. A necessidade de abordar as características "V" de BD, em conjunto com os atributos de qualidade, na definição, análise e especificação dos requisitos funcionais e de qualidade;
2. Separação de requisitos para infraestruturas, ferramentas e técnicas analíticas e aplicativos de usuário final;
3. Definição de novos métodos, ferramentas, processos e metodologias de RE para aplicações de BD.

2.3 Síntese deste Capítulo

Neste capítulo foi descrito o referencial teórico para realizar a pesquisa quantitativa. Na Seção 2.1 foram descritos os passos da SLR conjuntamente com os detalhes principais de cada fase. Além disso, na seção 2.2 foram descritos os 11 (onze) estudos selecionados e suas propostas. O relacionamento dos estudos encontrados com as questões de pesquisa é disposto no Capítulo 3.

Capítulo 3

Resultados das Questões de Pesquisa

Este Capítulo apresentará a análise dos estudos selecionados e detalhados no item 2.2, visando responder as questões de pesquisa RQ.1, RQ.2 e RQ.3. Foi necessário apresentar a definição dos termos "abordagem", "método", "técnica" e "ferramenta", com o objetivo de nivelar conceitos, mesmo que se tenha procurado considerar os próprios apresentados nos estudos.

Segundo [7], "abordagem é um arranjo sistemático, geralmente em etapas, de ideias ou ações destinadas a lidar com um problema ou situação"; "técnica é uma maneira de fazer algo ou um método prático aplicado a alguma tarefa específica"; "ferramenta se refere a um implemento, como um software ou um artefato". Portanto, de acordo com a definição de "abordagem" e "técnica" acima, o presente estudo considerou "método" como uma forma organizada e sistemática de atingir um determinado objetivo.

3.1 RQ.1: Quais as abordagens de RE no contexto de BD existentes na literatura?

Para responder a RQ.1, a Tabela 3.1 apresenta as abordagens de RE e as respectivas conclusões dos estudos selecionados, com o maior quantitativo direcionado ao processo de RE para BD contendo 5 (cinco) estudos: um deles com foco em RE orientada aos objetivos de negócio como no E4[2], outra com processo de projetos de *Big Data Analytics* como no E8[4] e as demais envolvendo proposta de integração entre processos de RE com BD como em E5[3], adaptações num projeto real como em E6[12] e proposta de suporte ao processo de RE em E9[5]. Como segundo critério de classificação, 4 (quatro) estudos apresentam a utilização de métodos automáticos na identificação dos requisitos implícitos, sejam por dados de BD como em E1[1] e E3[22], por documentos de RE em E2[27] ou por ambos em E10[6]. Os 2 (dois) estudos restantes apresentam pesquisa na literatura sobre os atributos

de qualidade mais adequados na definição da arquitetura de BD como em E7[28] e lacunas de RE nos projetos de BD conforme descrito em E11[17].

Tabela 3.1: Abordagens identificadas na SLR.

ID	**Abordagem de RE**	**Fase de RE**	**Conclusão do Estudo**
E1[1]	Elicitação automática de Requisitos orientada por dados em BD	Elicitação, Gerenciamento	Definição de um Metamodelo conceitual para apoiar a elicitação e o gerenciamento de requisitos orientados por dados com ML e PLN
E2[27]	Identificação e gerenciamento dos requisitos não funcionais implícitos (IMR) em documentos de especificação de RE em BD	Identificação (Elicitação), Gerenciamento	Taxonomia para IMR. O uso de ML geralmente atinge mais de 70% de precisão ao identificar e classificar os requisitos não funcionais. Pesquisa direcionada para o desenvolvimento "Waterfall"
E3[22]	Elicitação automática de Requisitos de dados heterogêneos de BD	Elicitação	Faltam métodos para apoiar a elicitação de requisitos de fontes de dados heterogêneas, para facilitar a evolução ou permitir novas oportunidades de negócio. SLR do E1[1]
E4[2]	RE orientada a objetivos de negócios para avaliar a qualidade da modelagem de BD	Modelagem	Modelagem conceitual para modelos de BD, com base nos objetivos de negócios e na qualidade dos dados de BD
E5[3]	Processo planejado de RE para projetos de BD	Elicitação, Análise, Modelagem, Validação, Especificação	Integração de processo de RE com o de análise de BD com a definição de novo Artefato "Diagrama de Caso de Uso Acionável"

Tabela 3.1 (continuação):

ID	Abordagem de RE	Fase de RE	Conclusão do Estudo
E6[12]	Processo RE para projetos de BD - Estudo de caso real na área de Petróleo e Gás	Elicitação, Especificação e Modelagem, Análise e Priorização	Necessidade de técnicas e ferramentas adequadas de RE na identificação de requisitos específicos de BD para evitar retrabalho e problemas com a arquitetura
E7[28]	Identificação dos requisitos não funcionais para obtenção dos atributos de qualidade na arquitetura de BD	Não consta	Atributos de qualidade obtidos por SMS: Eficiência de desempenho, Adequação funcional, Confiabilidade, Segurança, Usabilidade e Escalabilidade
E8[4]	Engenharia de Requisitos de Análise de Dados - Capturar, representar e comunicar a análise de requisitos em *Big Data Analytics*	Identificar, capturar (Elicitação), Especificação, Análise e Modelagem	Ferramenta colaborativa *Big Data Analytics Modeling Languages* (BiDaML)
E9[5]	RE baseada em artefatos, contendo a validação de proposta de modelo por profissionais de BD	Levantamento (Elicitação), Especificação	Artefato de Modelagem - BD-REAM para suporte ao levantamento e especificação de requisitos, definição de processos específicos de RE em BD, obter uma visão comum e rastreabilidade que ligam os artefatos constantes no modelo
E10[6]	Método para obtenção de requisitos não declarados em projetos de BD	Elicitação	Utilização de algoritmos de mineração de dados e tecnologias (PLN) na obtenção de requisitos não declarados

Tabela 3.1 (continuação):

ID	Abordagem de RE	Fase de RE	Conclusão do Estudo
E11[17]	Análise dos estudos acadêmicos sobre RE no contexto de aplicações de software envolvendo BD	Elicitação, Análise, Especificação, Modelagem, Validação	SLR apontou necessidade de novos métodos, ferramentas, processos e metodologias de RE para aplicações de BD

Observa-se que as abordagens dos estudos apresentam soluções de RE para 3 (três) tipos de desenvolvimento de aplicações de BD:

1. Projetos com processo definido que utilizam BD para aprimorar funcionalidades e os serviços prestados aos usuários finais em E5[3] e E9[5];
2. Os projetos destinados à análise de requisitos de *Big Data Analytics*, seja cobrindo todo o processo de análise em E8[4], seja avaliando a qualidade da modelagem de BD com base nos objetivos de negócio e na característica "variedade" dos dados em E4[2];
3. Solução orientada por dados de fontes digitais dinâmicas não intencionais, conforme proposto em E1[1], por meio da elicitação automática e contínua de requisitos, que utiliza BD como fonte de requisitos, independentemente do tipo de aplicação envolvendo BD.

Quanto aos demais estudos, que são revisões da literatura, segue o detalhamento das sugestões propostas:

- O E2[27] apresenta sugestões para obtenção automática de dados das especificações de requisitos de BD em projetos de desenvolvimento de software que usam o modelo de desenvolvimento "Waterfall", não abrangendo histórias de usuários;
- O E3[22] se refere a SLR referenciada no E1[1] sobre a elicitação automática de requisitos baseada em dados;
- O E6[12] descreve projeto real de BD na área de petróleo e gás, que necessitou de adaptação para execução das atividades de requisitos, diante da especificidade de BD, em razão de ausência de práticas, técnicas, ferramentas, padrões e tecnologias envolvidas para o desenvolvimento do projeto;
- O E7[28] por meio de SMS apresenta seleção de atributos de qualidade (requisitos não funcionais) mais adequados na arquitetura de software de BD, quais sejam: eficiência de desempenho, adequação funcional, confiabilidade, segurança, usabilidade

e escalabilidade, para que seja possível identificá-los antes do desenvolvimento do produto;

- O E10[6] propõem que os requisitos de informação sejam obtidos de forma iterativa durante o desenvolvimento do projeto, conforme os métodos usuais de RE, e que, durante a análise de dados, algoritmos de mineração de dados e tecnologias (PLN) sejam utilizados para novas descobertas e integrados à solução de BD;
- O E11[17] com apontamento de ausência de métodos, práticas e ferramentas de RE para BD, apresentou objetivos bem alinhados aos da presente pesquisa, o que possibilitou motivação para verificar se houve redução das lacunas identificadas nos artigos selecionados de 2013 a 2017.

Para analisar as fases ou atividades de abrangência de RE dos estudos selecionados, consideraram-se os termos "identificar", conforme apresentado em E2[27] e E8[4], "capturar" também constante em E8[4] e "levantamento" em E9[5] como "Elicitação" de requisitos. O gráfico apresentado na Figura 3.1 mostra uma visão desta abrangência. Somente o E7[28] não menciona atividade, processo ou fase de RE. Observa-se que a maioria dos estudos (nove) abrange a "Elicitação" de requisitos, reforçando a sua importância e criticidade também em aplicações de BD dos estudos selecionados; na segunda posição com 5 (cinco) estudos cada, encontra-se a fase de "Especificação" e de "Modelagem"; depois com 4 (quatro) a de "Análise"; com 3 (três) e a de "Gerenciamento"; com 2 (dois) a de "Validação" e somente em um estudo se menciona a atividade de "Priorização".

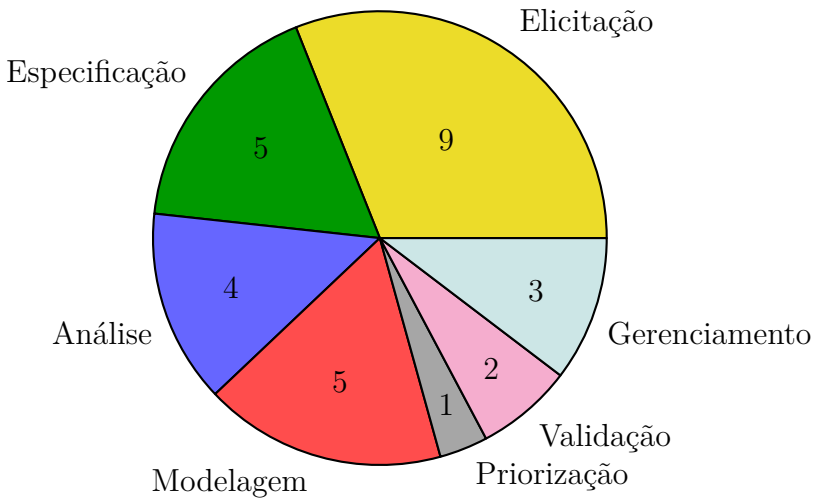


Figura 3.1: Distribuição das fases encontradas.

3.2 RQ.2: Quais são os métodos, técnicas e ferramentas de RE no contexto de Big Data existentes na literatura?

Esta sessão irá apresentar o relato dos estudos selecionados que propõem métodos, técnicas e ferramentas e suas descrições. Assim, com a presente pesquisa foi possível identificar 5 (cinco) propostas de ferramentas de RE, constantes nos estudos E1[1], E4[2], E5[3], E8[4] e E9[5], com os respectivos métodos e/ou técnicas utilizados, bem como se foram aplicadas no mercado, conforme apresentado na Tabela 3.2, com o objetivo de responder a questão RQ.2.

No E5[3], além de detalhar e apresentar fluxograma de um processo de RE para BD, também menciona, sem entrar no mérito de como viabilizar, a mineração de processo como método eficiente para a elicitar, priorizar e validar requisitos de BD usando logs de execução, descoberta de processos e técnicas de conformidade. Da mesma forma, E2[27], sem apresentar como fazer, sugere que técnicas de ciência de dados, como mineração de regras de associação, poderiam ser usadas na descoberta da relação entre a fonte real do erro que causa o defeito na fase de RE, contribuindo com a gestão de requisitos não funcionais implícitos e na qualidade das especificações de requisitos. Já E10[6], também sem entrar no mérito conforme detalhado em E1[1], propõe que os requisitos de informação de BD sejam obtidos de forma iterativa durante o desenvolvimento do projeto, de acordo os métodos usuais de RE, e que, durante a coleta e análise de dados com PLN, sejam descobertos novos requisitos não declarados, incorporando-os à solução de BD.

A SLR de E3[22] apurou que as técnicas para elicitação automatizada de requisitos em sua pesquisa utilizaram algoritmos categorizados em aprendizado de máquina (ML), classificação baseada em regras, abordagem orientada a modelos, modelagem de tópicos e agrupamento tradicional. A maioria dos estudos utilizou Processamento de Linguagem Natural (PLN) como técnica para processamento de dados, técnica de ML para classificação e agrupamento, e poucos estudos conseguiram obter requisitos de alto nível e com aplicação no mundo real.

Na Elicitação de requisitos no caso do projeto real em E6[12], foram efetuadas reuniões com os envolvidos, análise em documentos contratuais e definição de um modelo de requisitos contendo os componentes do sistema, suas interações externas, fontes de dados e usuários. Na Especificação e Modelagem, utilizaram-se casos de uso (UML) para descrição textual e alguns requisitos do projeto foram documentados de forma gráfica, assim como para a modelagem de fluxos de dados, modelo de dados e modelos conceituais e arquitetura, utilizando a ferramenta Visual Paradigm. Para os requisitos de interface de usuário

e camada de apresentação se utilizou ferramenta de mock-up na criação de protótipo. O Google Drawings foi utilizado para desenhar a arquitetura do sistema, com adaptações para retratar serviços e fluxo de dados. Entretanto, constatou-se que 35% (trinta e cinco por cento) dos requisitos de BD foram identificados nas fases de design, arquitetura, codificação, gerando retrabalho e prejudicando a escalabilidade da arquitetura.

A SLR do E11[17], publicada em 2018, apontou a necessidade de abordar as características "V" de BD, em conjunto com os atributos de qualidade, na definição, análise e especificação dos requisitos funcionais e de qualidade, a separação de requisitos para infraestruturas, ferramentas e técnicas analíticas e aplicativos de usuário final e definição de novos métodos, ferramentas, processos e metodologias de RE para aplicações de BD. São os mesmos autores de E9[5] com a proposta de modelo de artefato de RE BD-REAM, publicada em 2019, e o primeiro autor do E6[12] sobre o caso do projeto real de 2020.

Tabela 3.2: Tabela para destaque dos estudos que constam algum dos três resultados da revisão.

ID	**Ferramenta**	**Método/Técnica**	**Aplicação no mercado**
E1[1]	*Framework*, na forma de um metamodelo conceitual	Modelo, com o processo para automatizar a elicitação e o gerenciamento de requisitos orientados por dados, com técnicas de ML e PLN	Aplicativo de video game do mercado
E4[2]	Modelo conceitual - IRIS	Modelo Extended Entity-Relationship (EER), com base em conceitos de negócios, atributos de qualidade e análise de BD	Estudo empírico sobre remessa de produtos da empresa Zara Inc.
E5[3]	Diagrama de caso de uso acionável	UML- Unified Modeling Language	Não houve
E8[4]	Sistema *Big Data Analytics Modeling Languages* (BiDaML)	Linguagem visual, anotação BPMN, geração de código de programação	Projetos reais de software de BD na área de finanças, transporte e saúde

Tabela 3.2 (continuação):

ID	Ferramenta	Método/Técnica	Aplicação no mercado
E9[5]	Modelagem de artefato de RE BD-REAM	Modelo de Referência de Engenharia de Requisitos (REM)[29]	Não houve

Observa-se que dos 5 (cinco) estudos que apresentam ferramentas de RE para BD, 3 (três) utilizam modelos com entidades, relacionamentos e cardinalidades (E1[1], E4[2] e E9[5]), contendo os seguintes aspectos:

- E1[1] detalha um metamodelo conceitual para estruturar e projetar o processo de elicitação e gerenciamento de requisitos por dados de BD de fontes heterogêneas, geradas por humanos e máquinas. O processamento dos dados é automatizado na coleta, classificação, análise, agregação e mapeamento dos requisitos candidatos a novos ou alteração dos existentes, com a possibilidade de intervenção manual ou semiautomática na geração de novos artefatos de requisitos ou mudança dos pré-existentes. Utilizam-se técnicas e algorítimos de processamento de dados, classificação, agregação, análise de sentimento, comportamento dos usuários, sistemas e máquinas, com PLN e ML.

  - **Vantagens**:

    1. Compartilhar e compreender conceitos comuns entre as equipes envolvidas;
    2. Ampliar as fontes de requisitos, incluindo grandes e diversas bases de usuários;
    3. Automação da análise dos dados com PLN e ML, inviável de se tratar manualmente;
    4. Apoiar a manutenção e evolução do software, atualizando os requisitos com os novos dados gerados;
    5. A classificação de tipos de dados digitais permite ter visão da quantidade e importância dos dados de diversas fontes, direcionando a alocação e utilização de recursos de forma mais eficaz;
    6. As entidades relacionadas ao processamento de dados no metamodelo fornecem base para documentar as técnicas e os modelos preditivos utilizados;
    7. Elementos de agregação e mapeamento de requisitos possibilita automação de tarefas para o engenheiro de requisitos sem eliminar sua criatividade;
    8. Automação da análise de sentimentos proporciona economia de alocação de recursos humanos na realização da atividade antes manual, o aumento

do volume de dados analisados, a diminuição da subjetividade e rapidez na melhoria lançamento de novos produtos.

  - **Desvantagens**:
    1. Ampliar o uso de técnicas para derivar requisitos ou informações de dados gerados por máquina;
    2. Etapas insuficientes relacionadas ao mapeamento automático de requisitos em linguagem natural para Caso de Uso;
    3. O *framework* foi validado de forma parcial, pois o estudo de caso não envolveu todas as classes previstas no Metamodelo.

- Um modelo de *Big Data Virtual* com base nos objetivos de negócio é proposto em E4[2], denominado IRIS, para avaliar a qualidade dos dados e, consequentemente, a modelagem de BD. Utiliza o modelo *Extended Entity Relationship* (EER) para representar cada conceito como meta ou submetas, visando analisar se o "problema" e a "solução", conforme os resultados de *Big Data Queries* ou KPI (*Key Performance Indicator*), contribuem positivamente ou não para atingir a meta ou submetas relacionadas. Considera três atributos de qualidade na característica de "Variedade" dos dados de BD: "Relevância" (se os dados e relacionamentos entre os dados são relevantes, descartando os que não são); "Abrangência" (considerar a variedade de diferentes tipos e fontes de dados) e "Prioridade Relativa" para determinar a relação entre a quantidade limitada de recursos e volume de dados ou a velocidade de processamento. Além disso, o modelo contém dimensões relativas aos dados (interna/externa à organização; *offline*/*online*; primeira - dados comum de primeira ordem - ou analítica, obtida por meio de análise dos dados comuns) e dimensões organizacionais que auxiliam estruturar a variedade, o volume e alta velocidade dos dados (Classificação/Instanciação, Generalização/Especialização, Agregação/ Decomposição).

  - **Vantagens**:
    1. Auxilia o rastreamento do processo de modelagem de BD conceitual;
    2. Possibilita explorar e selecionar alternativas em problemas, soluções, análise de negócios e modelos de BD;
    3. Possibilita identificar os tipos e fontes de dados externos apropriados, racionalizando sua seleção.
  - **Desvantagens**:
    1. Ausência de validação real envolvendo profissionais de TI;
    2. Considerar mais atributos de qualidade de BD, não só a Variedade;

3. Automatizar a mensuração dos critérios de qualidade aplicados para a análise dos dados e do modelo.

- O E9[5] apresenta um modelo orientado a artefatos de RE, denominado BD-REAM, com o objetivo de identificar o conhecimento de domínio da aplicação e as restrições de projeto para definir o escopo dos requisitos de cada cenário identificado. Contém 43 (quarenta e três) Entidades (artefatos) agrupadas com base no Modelo de Referência de Engenharia de Requisitos (REM)[29], categorizadas em: 7 (sete) artefatos de necessidades de negócios, que especificam requisitos estratégicos e de clientes, incluindo produtos e objetivos de negócios; 15 (quinze) artefatos de especificação de requisitos, que contém requisitos funcionais e não funcionais, analisados e modelados, com base nas perspectivas do usuário, e derivados e justificados pelas necessidades do negócio; 21 (vinte e um) artefatos de especificação de sistemas, que compõem a definição do funcionamento sistema, sua integração, restrições, comportamentos e ambiente. Do total das 43 (quarenta e três) Entidades, 17 (dezessete) são específicas de BD.

  – **Vantagens**:

    1. Criação de uma visão comum de RE em projetos de BD;
    2. Definição de processos específicos de RE;
    3. Auxilia no levantamento e especificação de requisitos;
    4. Ferramentas de rastreabilidade que ligam os artefatos.

  – **Desvantagens**:

    1. Necessidade de abranger novos domínios de aplicação, como IoT (Internet das Coisas);
    2. Ausência de aplicação real em projetos de BD para avaliar, adaptabilidade e generalização do modelo;
    3. Análise do custo da adoção do modelo de artefato em projetos industriais.

- Uma adaptação de Caso de Uso em UML, “Diagrama de Caso de Uso Acionável”, é um novo artefato apresentado como produto de trabalho da atividade “Consolidação de Casos de Uso” no processo de RE proposto em E5[3], que permite documentar os valores de negócios dos dados, como estes são obtidos, os algoritmos utilizados, os métodos de análise, os meios que se utilizam para acionar estes valores (inteligência acionável, conhecimento), bem como as suas funções e suas interações com outros componentes de software.

  – **Vantagens**:

1. Integração de processos conhecidos;
2. Atuação conjunta, em ambiente colaborativo, entre o engenheiro de requisitos de RE e o cientista de dados de BD;
3. Documentar e descobrir os valores dos dados.

– **Desvantagens**:

1. Modelo conceitual sem aplicação real;
2. Ausência de atividades de gerenciamento de requisitos, especialmente de qualidade de BD.

- Uma ferramenta colaborativa é proposta em E8[4], composta por um sistema denominado BiDaML, constituída por um conjunto de linguagens visuais de um domínio específico, que utiliza 5 (cinco) tipos de diagramas em diferentes níveis de abstração, cobrindo todo o ciclo de desenvolvimento de software de análise de dados de BD, desde requisito de alto nível e definição do problema até a implantação do produto final, quais sejam: Diagrama de *Brainstorming*, Diagrama de Processo, Diagrama de Técnica, Diagrama de Dados e Diagrama de implantação. Possibilita gerar uma documentação a partir de qualquer diagrama selecionado, um relatório de atividades e um template para códigos de acesso a API ou código em python disponíveis para os cientistas de dados utilizarem. É possível acessar a versão web, cadastrar-se e utilizar a ferramenta BiDaML[4].

  – **Vantagens**:

  1. Criar ambiente colaborativo visual, uma linguagem comum entre usuários e equipes de TI, a variedade de perspectivas, tarefas e interações;
  2. Comunicar e documentar o processo e o resultado de análise e o projeto de requisitos de negócios, o processamento dos dados com as técnicas utilizadas e a implantação;
  3. Fornecer suporte na evolução de soluções de software de BD;
  4. Viabilizar o reuso de soluções existentes, além de compartilhar conhecimento;
  5. Disponibilizar códigos de acesso a API ou código em "python", permitindo visualizar as incorporações ou expansão de algoritmos;
  6. Ferramenta autônoma na Web, conectada às caixas de ferramentas de recomendação de ML de técnicas apropriadas para os dados selecionados.

  – **Desvantagens**:

  1. Necessidade de treinamento e uso de material de instrução para navegação pelo usuário;

2. Não fornecer visualizações diferentes conforme os tipos de usuários;
3. Permitir ocultar ou exibir diferentes componentes dos diagramas conforme preferências dos usuários;
4. Na geração de código, separar as tarefas realizadas por humanos ou ferramentas.

Das 5 (cinco) ferramentas, somente duas apresentaram aplicação real no mercado, conforme descrito em E1[1] e E8[4]. O E4[2] utilizou dados empíricos para validar o modelo IRIS. O E6[12], apesar do modelo BD-REAM ter sido validado por 10 profissionais da área de desenvolvimento de software de BD, o artefato não foi utilizado em projeto real de BD.

Conforme descrito nos estudos que apresentaram proposta de ferramenta de RE para soluções de BD, foi possível relacionar acima suas vantagens e desvantagens. Destaque para os estudos com aplicação real, que possibilitou considerar as próprias opiniões das equipes que utilizaram a ferramenta. Observa-se que a obtenção de "uma linguagem comum", "compartilhamento" e "ambiente colaborativo" entre as equipes, são os pontos positivos apresentados nas respectivas ferramentas de todos os estudos. Quanto aos pontos negativos, da mesma forma, todos os estudos apresentam a necessidade de ampliar, considerar outros ou novos aspectos relevantes de RE, BD ou tecnológico, visando atingir maior abrangência, completude com o uso das referidas propostas.

3.3 RQ.3. Quais técnicas e ferramentas da RE no contexto de BD são utilizadas nas instituições financeiras?

Com o detalhamento dos métodos, técnicas e ferramentas de RE em BD selecionados pela SLR, foi elaborado um *survey* [23], direcionado às equipes de TI, com atuação em soluções de BD em algumas instituições financeiras nacionais de grande porte, utilizadas como amostra para este nicho de mercado.

Em resposta à RQ.3, foi possível constatar, conforme análise dos dados do *survey*, que os métodos, técnicas e ferramentas selecionados na literatura são parcialmente utilizados nas instituições financeiras investigadas. Além disso, obtiveram ampla aceitação pela utilização das ferramentas apontadas nos estudos selecionados, conforme detalhado na análise da Seção 5.2 e sintetizada a seguir.

A maioria dos participantes da pesquisa possui mais de 10 (dez) anos de experiência na área de TI. Afirmaram a existência na instituição de aplicação *online*, que obtém resposta

de *Big Data Analytics* destinadas aos usuários finais. Os "Requisitos Funcionais" e os "Não Funcionais" foram levantados para o desenvolvimento da referida aplicação, utilizando a técnica História de Usuário (Método Ágil) e/ou Caso de Uso (UML).

A pesquisa mostrou o uso inexpressivo de fontes digitais externas à organização para análise de BD. A maioria que afirmou o seu uso, indicaram utilizar as fontes de dados mediados por processos (tais como, transações comerciais, registros bancários, pagamentos com cartão de crédito), bem característicos aos negócios bancários. Por outro lado, houve expressiva concordância quanto à importância de identificar os requisitos originários de diversos tipos de fontes de dados para descobrir novas oportunidades de negócio.

Por fim, os participantes da pesquisa indicaram a complexidade tecnológica do BD e questões relativas ao conhecimento técnico e organizacional como os principais desafios e problemas no contexto de BD.

3.4 Síntese deste Capítulo

Neste Capítulo, foram relacionadas às questões de pesquisa RQ.1 e RQ.2, com os resultados dos 11 (onze) estudos obtidos pela SLR no Capítulo 2. Quanto à RQ.3, o Capítulo 4 contém o detalhamento das configurações do *survey* e o planejamento para análise dos dados obtidos. O Capítulo 5 descreve e analisa os resultados do *survey* utilizados para responder esta questão.

Capítulo 4

Survey

Este Capítulo descreve as configurações do *survey*, em como os dados foram coletados e interpretados para responder à questão de pesquisa: RQ.3. Quais técnicas e ferramentas da RE no contexto de BD são utilizadas nas instituições financeiras?.

4.1 Configuração do *survey*

Para operacionalizar a pesquisa, foi necessário definir: a população-alvo, o procedimento de amostragem, os objetivos do estudo em um conjunto de perguntas, a estratégia de aplicação da pesquisa e coleta dos dados, o desenho do questionário, as abordagens para análise de dados e questões de validade [30].

O questionário se destinou aos profissionais de TI, que trabalham com desenvolvimento de Software ou de Sistema, nas áreas específicas de Engenharia de Requisitos ou na área de Ciência de Dados, de instituições financeiras nacionais de grande porte, que utilizam BD. Portanto, utilizou-se a amostragem não probabilística por conveniência.

As instituições financeiras, que foram selecionadas para a realização da pesquisa, possuem abrangência nacional, tradicionais no uso intensivo de dados como apoio à tomada de decisão, conforme divulgado nos meios de comunicação. Além disso, foi possível o envolvimento de 3 (três) especialistas destas instituições da área de TI, que atuaram como facilitadores e ponto focal para a aplicação da pesquisa. Vale ressaltar a dificuldade de se obter informação sem aprovação prévia da área de governança das organizações, diante da exigência de estudo detalhado do seu conteúdo e uso, sem garantia de prazo para apreciação.

Os especialistas possuem mais de 20 (vinte) anos de experiência de TI, não participaram da pesquisa (não responderam o *survey*), mas atuaram na distribuição do questionário para as áreas específicas das instituições que utilizam BD. Além disso, eles participaram também na validação do survey quanto à formatação, clareza e compreensão do conteúdo

do questionário.Assim foi realizado um piloto com os 3 especialistas da área com o objetivo de realizar uma avaliação prévia do entendimento das questões do survey. Foram apresentadas 24 (vinte e quatro) questões no questionário diagnóstico, as quais foram reestruturadas na versão final da pesquisa, totalizando 22 (vinte e duas) questões.

As questões, no questionário diagnóstico, não foram divididas em seções específicas como apresentada na versão final. A ordem das perguntas também era diferente, entretanto, havia um detalhamento mais específico sobre cada ferramenta abordada, a disponibilização de uma extensa explicação sobre cada modelo, incluindo as respectivas figuras originárias dos estudos, conforme apresentadas no Capítulo 2 (Figura 2.2, Figura 2.4, Figura 2.5, Figura 2.6, Figura 2.7).

Os especialistas sugeriram a separação das questões em 3 (três) sessões, a simplificação do detalhamento das ferramentas e retirada das figuras, além da exclusão de 2 (duas) questões para tentar diminuir o tempo de resposta dos participantes e a profundidade e complexidade do *survey* aplicado. Na Tabela 4.2, são apresentados os estudos que respondem as questões do *survey*.

O questionário foi elaborado utilizando Microsoft Forms[1] devida a organização da apresentação do seu conteúdo, facilidade de acesso para os respondentes, controle online da coleta e processamento das respostas. Neste estudo, nas questões em que era necessário obter a percepção mais pontual dos respondentes com relação aos modelos apresentados foi utilizada a escala de Likert. Segundo Komorita [31], em uma escala Likert, o respondente é apresentado a um conjunto de declarações de atitude em uma escala que varia de concordo totalmente a discordo totalmente.

4.2 Perguntas do *survey*

Foram elaboradas 22 (vinte e duas) questões, sendo 19 (dezenove) de natureza fechada e 3 (três) abertas. O *survey* foi dividido em 3 (três) seções. A primeira seção apresenta perguntas acerca da identificação da atuação profissional do respondente, a segunda apresenta questionamentos acerca de RE em BD e na terceira são expostos às respectivas ferramentas deste contexto obtidas na literatura. As questões são apresentadas na Tabela 4.1.

Tabela 4.1: Perguntas do *survey*.

ID	**Pergunta**	**Tipo**	**Opções**

[1] `https://www.microsoft.com/pt-br/microsoft-365/online-surveys-polls-quizzes`

Tabela 4.1 (continuação):

ID	**Pergunta**	**Tipo**	**Opções**
P01	Área de atuação	Fechada	Ciência de Dados (Big Data); Engenharia de Requisitos; Engenharia de Software ou de Sistemas.
P02	Experiência na área de atuação	Fechada	Menos de 5 anos; De 5 até 10 anos; Mais de 10 anos.
P03	Instituição possui aplicação online que obtém resposta de "Big Data Analytics" destinadas aos usuários finais?	Fechada	Sim; Não; Não sei informar.
P04	Se sim, os "Requisitos Funcionais" e os "Não Funcionais" (Requisitos de Qualidade ou Restrições) foram levantados para o desenvolvimento da referida aplicação?	Fechada	Sim; Não; Não sei informar.
P05	Se sim, qual a técnica utilizada nas especificações dos referidos requisitos:	Fechada	Caso de Uso (UML); História de Usuário (Método Ágil); Outra (com opção de texto livre para digitação).
P06	Utiliza dados de fontes digitais externas, fora da plataforma dessa instituição, na análise de dados de Big Data?	Fechada	Sim; Não; Não sei informar.

Tabela 4.1 (continuação):

ID	Pergunta	Tipo	Opções
P07	Se sim, indique qual ou quais os seguintes tipos de fontes utilizados:	Fechada (Múltipla escolha)	Fonte de dados de origem humana, principalmente na forma de linguagem natural, podendo conter áudio, vídeo e imagens, tais como E-mail, Review, MicroblogPost, ForumPost, Chat, etc; Fontes de dados mediadas por processos, referentes aos registros do monitoramento de processos e eventos de negócios, por exemplo, transações comerciais, registros bancários, pagamentos com cartão de crédito; Fontes geradas por máquina, isto é, dados de sensores (SensorReading) ou logs de computador de vários tipos: ServerLogEntry ou UserLogEntry; Outra (com opção de texto livre para digitação).
P08	É importante identificar os requisitos originários das fontes externas, pois isso facilita a evolução e o desenvolvimento de novas oportunidades de negócio. Você concorda com essa afirmação?	Fechada	Concordo totalmente; Concordo; Neutro; Discordo; Discordo totalmente.

Tabela 4.1 (continuação):

ID	Pergunta	Tipo	Opções
P09	Você utiliza alguma ferramenta para comunicar, documentar o processo e o resultado da análise de dados e o projeto de requisitos de negócios, incluindo como será efetuado o processamento dos dados com as técnicas utilizadas, bem como a implantação de solução em Big Data, disponível para todos os envolvidos?	Fechada	Sim; Não; Não sei informar.
P10	Se sim, qual?	Aberta	
P11	Você utiliza alguma ferramenta para avaliar a qualidade dos dados e, consequentemente, a modelagem de Big Data?	Fechada	Sim; Não; Não sei informar.
P12	Se sim, qual?	Aberta	
P13	A arquitetura de software de Big Data é mais complexa, devido ao uso intensivo de dados. Há necessidade de analisar todos os requisitos não funcionais, que garantam a comunicação e coordenação entre os componentes, conectores e restrições arquiteturais. É necessário identificá-los antes de se projetar o produto. Você concorda com essa afirmação?	Fechada	Concordo totalmente; Concordo; Neutro; Discordo; Discordo totalmente.
P14	Indique a seguir o(s) seguinte(s) atributo(s) de qualidade (requisitos não funcionais) considerado(s) como o(s) mais adequado(s) na definição da arquitetura de software de Big Data:	Fechada (Múltiplas alternativas)	Eficiência de desempenho; Adequação Funcional; Confiabilidade; Segurança; Usabilidade; Escalabilidade; Outra (com opção de texto livre para digitação).

Tabela 4.1 (continuação):

ID	Pergunta	Tipo	Opções
P15	Na sua percepção quais são os principais desafios e problemas no contexto de Big data.	Aberta	
P16	Para se obter os requisitos das fontes de dados externas à organização um framework é proposto para automatizar a sua identificação por meio de PLN (Processamento de Linguagem Natural) e ML (Machine Learning). Na sua visão, a proposta seria útil na sua área de atuação?	Fechada	Sim; Não.
P17	É necessária a atuação conjunta, do engenheiro de requisitos e do cientista de dados no projeto de desenvolvimento com Big data. Você concorda com essa afirmação?	Fechada	Concordo totalmente; Concordo; Neutro; Discordo; Discordo totalmente.
P18	Uma adaptação do Caso de Uso (UML), denominada "Diagrama de Caso de Uso Acionável", com o objetivo de documentar os valores de negócios dos dados, como são obtidos, algoritmos utilizados, métodos de análise, funções e interações com outros componentes de software. Na sua visão, a referida proposta seria útil na sua área de atuação?	Fechada	Sim; Não.

Tabela 4.1 (continuação):

ID	Pergunta	Tipo	Opções
P19	Uma ferramenta web denominada "Big Data Analytics Modeling Languages" (BiDaML) propõe capturar, representar, documentar e comunicar a análise e projeto de requisitos de negócios, pré-processamento de dados, processo de análise de dados de alto nível, implantação de soluções e visualização de dados. Na sua visão, a referida ferramenta seria útil na sua área de atuação?	Fechada	Sim; Não.
P20	A definição de um Big Data Virtual em meta ou submetas, com a finalidade de analisar se o "problema" e a "solução" contribuem positivamente ou não para atingir os objetivos de negócio. Na sua visão, a referida proposta seria útil na sua área de atuação?	Fechada	Sim; Não.
P21	Um modelo propõe criar uma visão comum da Engenharia de Requisitos em projetos de Big Data, auxiliar no levantamento e na especificação de requisitos, além de possibilitar a rastreabilidade que ligam os artefatos. Você concorda com essa proposta?	Fechada	Concordo totalmente; Concordo; Neutro; Discordo; Discordo totalmente.

Tabela 4.1 (continuação):

ID	Pergunta	Tipo	Opções
P22	Na sua visão qual das propostas apresentadas agregaria mais valor na aplicação da Engenharia de Requisitos no desenvolvimento de software envolvendo Big Data nesta instituição?	Fechada	O framework para levantamento de requisitos orientado a dados de fontes externas à organização; O Artefato "Diagrama de Caso de Uso Acionável"; A Ferramenta colaborativa BiDaML; O Big Data Virtual; O modelo orientado a artefatos da Engenharia de Requisitos; Nenhuma das propostas apresentadas.

Tabela 4.2: Assuntos abordados no *survey*.

Seção	Questões	Objetivo	Estudos relacionados
Atuação Profissional	P01 e P02	Identificação da atuação profissional do participante	Não consta
Engenharia de Requisitos em Big Data	P03 a P15	Perguntas relacionadas à RE e BD	P03 a P05: Todos os estudos primários selecionados. P06: E1[1], E3[22] e E4[2]. P07 e P08: E1[1] e E3[22]. P09 e P10: E8[4]. P11 e P12: E4[2]. P13 e P14: E7[28]
Propostas da Engenharia de Requisitos para Big Data	P16 a P22	Perguntas sobre as ferramentas identificadas na literatura	P16: E1[1]. P17: E1[1], E5[3] e E8[4]. P18: E5[3]. P19: E8[4]. P20: E4[2]. P21: E9[5]

4.3 Descrição e Análise dos Resultados do *survey*

Para a análise de dados do *survey* foi utilizado o guia proposto por Molleri et al. [23]. Os autores detalham todo o desenvolvimento de um *survey* conjuntamente com as distinções de análises de dados qualitativos e quantitativos [32]. Os passos para descrição e análise dos dados foram realizados de acordo com as seguintes etapas:

1. Coleta de dados do *survey*;
2. Apresentação e análise dos dados adquiridos;
3. Interpretação dos dados para responder a questão de pesquisa RQ.3.

A etapa 1 se restringe à própria coleta de dados do *survey*, extraídos da ferramenta Microsoft Forms. Na etapa 2 os dados obtidos são apresentados, relacionados e analisados de acordo com o número e tipo de respondentes por questões formuladas no questionário, utilizando tabelas e gráficos. Por fim, na etapa 3, os dados apresentados e analisados são interpretados com o objetivo de responder à questão de pesquisa RQ.3.

4.4 Síntese deste Capítulo

Neste capítulo foi apresentando a configuração do *survey* desenvolvido na presente pesquisa. Na seção 4.1 foi descrito o detalhamento das configurações do *survey*. Já na seção 4.2, foram apresentadas as 22 (vinte e duas) questões do *survey*, assim como as suas características. Por fim, na seção 4.3, foram definidas as etapas para a descrição e a análise dos dados do *survey* que serão detalhados no Capítulo 5.

Capítulo 5

Análise do Resultado do *survey*

Este Capítulo descreve e analisa os resultados do *survey*, bem como responde a RQ.3. Em seguida, são apresentadas as ameaças à validade e limitações deste trabalho.

5.1 Análise dos Dados do *survey*

As 22 (vinte e duas) questões (P01 a P22) do *survey*, conforme apresentado na Tabela 4.1, obtiveram 52 (cinquenta e duas) respostas de profissionais que trabalham com Engenharia de Requisitos, Ciência de Dados ou Engenharia de Software ou de Sistemas. O *survey* foi disponibilizado entre os dias 10 de Maio a 13 de Junho de 2023. A divulgação do questionário foi realizada através do LinkedIn[1], Facebook[2] e Whatsapp[3], destinado às equipes de desenvolvimento de sistemas de BD em instituições financeiras, assim como setores de Engenharia de Requisitos de fábricas de software das respectivas organizações. O tempo médio de resposta foi de treze minutos e dezoito segundos. Nós informamos que estávamos a disposição para sanar qualquer dúvida ou para compartilhar os resultados do *survey*, caso o respondente desejasse, disponibilizando o e-mail institucional para contato.

A Tabela 5.1 apresenta a consolidação das áreas de atuação dos participantes da pesquisa. 57,7% (30/52) dos participantes pertencem à área de Engenharia de Software ou de Sistemas; 23,1% (12/52) são de Ciência de Dados e 19,2% (10/52) dos participantes são de Engenharia de Requisitos.

O tempo de experiência dos respondentes pode ser considerado elevado em sua maioria, sendo que, 61,5% (32/52) afirmaram ter mais de 10 anos de atuação na área, dos quais 38,5% (20/52) são da Engenharia de software ou de sistema, 13,4% (7/52) são de Engenharia de Requisitos e 9,6% (5/52) são de Ciência de Dados. O total de participan-

[1] `https://br.linkedin.com/`
[2] `https://www.facebook.com/`
[3] `https://www.whatsapp.com/?lang=pt_br`

tes que afirmaram ter entre 5 a 10 anos de atuação foi de 17,3% (9/52), sendo que 9,6% (5/52) pertencem a área de Ciência de Dados e 3,85% (2/52) são tanto da Engenharia de software ou de sistema quanto da Engenharia de Requisitos. O restante dos participantes, 21,2% (11/52), são pertencentes ao grupo com menos de 5 anos de experiência, dos quais 15,4% (8/52) são de Engenharia de software ou de sistema, 3,85% (2/52) de Ciência de Dados e somente um 1,95% (1/52) de Engenharia de Requisitos.

Na Tabela 5.1 é possível observar que mais de 60% dos participantes possuem uma vasta experiência na área de atuação, o que nos possibilita inferir que as técnicas e ferramentas da RE no contexto de BD identificadas na literatura podem ser utilizadas nas instituições financeiras.

Tabela 5.1: Respostas às questões P01 e P02.

Área de Atuação	**Menos de 5 anos (%)**	**Entre 5 e 10 anos (%)**	**Mais de 10 anos (%)**
Engenharia de Software ou de Sistemas	15,4%	3,85%	38,5%
Ciência de Dados (Big Data)	3,85%	9,6%	9,6%
Engenharia de Requisitos	1,95%	3,85%	13,4%

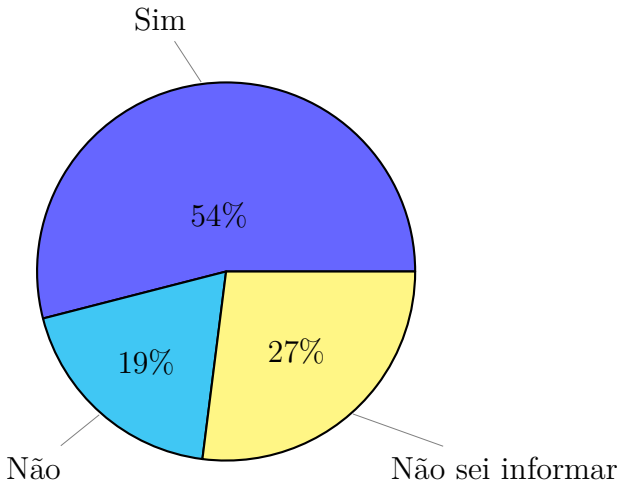
Os resultados das questões P03, P04 e P05, são apresentados na Figura 5.1. Essas questões englobam uma problemática abordada em todos os estudos selecionados quanto ao emprego da RE em BD em projetos com uso intensivo de análise de dados, identificando se os tipos de requisitos funcionais e os de qualidade foram levantados e quais foram as técnicas utilizadas nas especificações desses requisitos.

Dos 52 (cinquenta e dois) respondentes, 53,8% (28/52) afirmaram que a instituição possui aplicação *online* que obtém resposta de Big Data Analytics destinadas aos usuários finais (P03 da Figura 5.1), sendo 30,7% (16 /52) de Engenharia de software e 11,5% (6/52) tanto de Ciência de Dados como de Engenharia de Requisitos. Para o restante dos participantes, 19,2% (10/52) informaram não possuir este tipo de aplicação e 26,9% (14/52) não souberam informar.

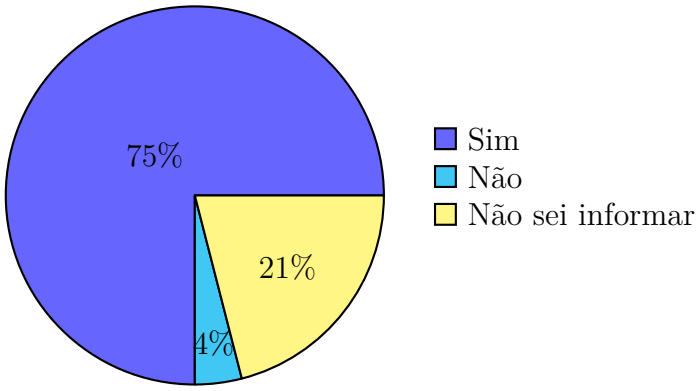
Os 53,8% (28/52) respondentes que marcaram "sim" na P03 da Figura 5.1, foram questionados se os "Requisitos Funcionais" e os "Não Funcionais" (Requisitos de Qualidade ou Restrições) foram levantados para o desenvolvimento da referida aplicação, conforme descrito na P04 da Figura 5.1. Já 40,3% (21/52) informaram "sim" para a P04, sendo

que 19,2% (10/52) da área de Engenharia de software ou de sistema, 11,5% (6/52) de Engenharia requisitos e 9,6% (5/52) de Ciência de Dados. Apenas 1,9% (1/52) marcou "não" e os restantes 11,5% (6/52) indicaram não saber informar.

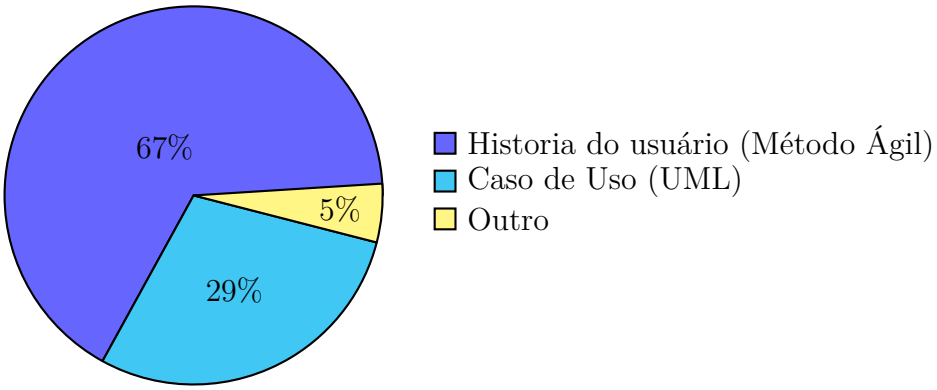
Os 40,3% (21/52) respondentes da P04 da Figura 5.1 foram solicitados a indicar na P05 qual a técnica utilizada nas especificações dos requisitos, 26,9% (14/52) marcaram História de Usuário (Método Ágil), 11,5% (6/52) marcaram Caso de Uso (UML) e apenas 1,9% (1/52) dos respondentes marcaram a opção com texto livre e indicou as técnicas "UC e HU" (P05 da Figura 5.1).



(a) P03



(b) P04



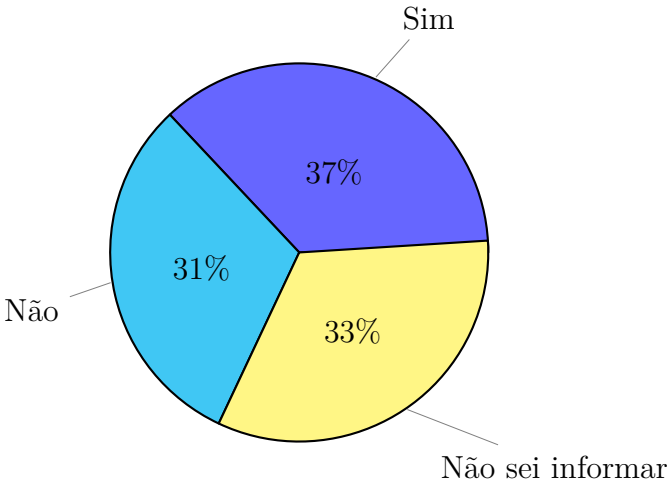
(c) P05

Figura 5.1: Respostas das questões P03, P04 e P05.

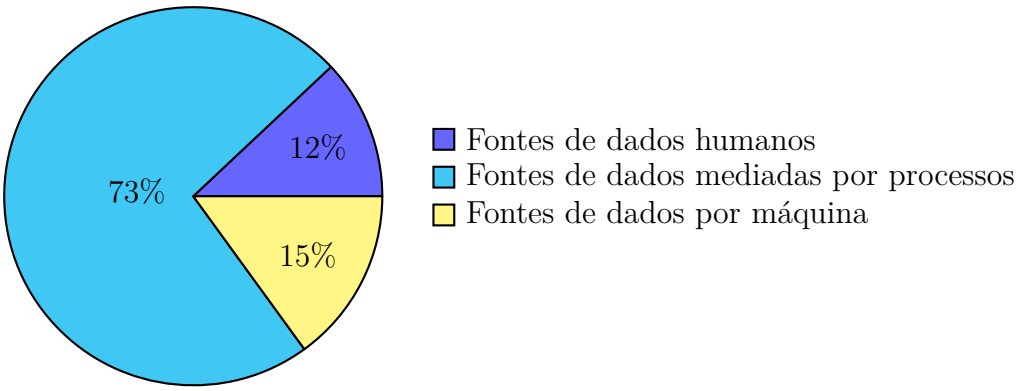
Com relação à questão P06, 36,5% (19/52) dos respondentes informaram que a instituição utiliza fontes digitais externas na análise de BD, 30,7% (16/52) marcaram não utilizar tais fontes, enquanto que 32,7% (17/52) não souberam informar (P06 da Figura 5.2).

Os 36,5% (19/52) respondentes que marcaram “sim” na P06 indicaram que utilizam

fontes de dados por processos, conforme opção constante na P07, o que reforça a característica da instituição ao ter como prática a análise e monitoramento de negócios bancários. Por outro lado, os tipos de fontes de dados de origem humana e fontes de dados geradas por máquina, constante como demais opções na P07, obtiveram 5,7% (3/52) e 7,6% (4/52) marcações respectivamente, conforme apresentado na P07 da Figura 5.2.



(a) P06



(b) P07

Figura 5.2: Respostas das questões P06 e P07.

Houve uma concordância muito alta com relação à questão P08, na qual 90,4% (47/52), conforme apresentado na P08 Figura 5.3, dos participantes afirmaram positivamente a importância de identificar requisitos de fontes externas para facilitar a evolução e desenvolvimento de novas oportunidades de negócio. Outrossim, é possível observar na P13 da Figura 5.3, a qual apresentou questionamento sobre a identificação de requisitos não funcionais antes de projetar o produto, que 80,8% (42/52) dos participantes concordaram com essa afirmação. Além disso, 92,3% (48/52) dos participantes concordaram que a participação de um engenheiro de requisitos é necessária em um projeto de BD, conforme

apresentado na P17 da Figura 5.3. 76,9% (40/52) dos participantes do *survey* concordaram com a afirmação apresentada na P21 sobre o modelo orientado a artefatos, o que demonstra sua importância para as instituições financeiras, conforme apresentado na P21 da Figura 5.3.

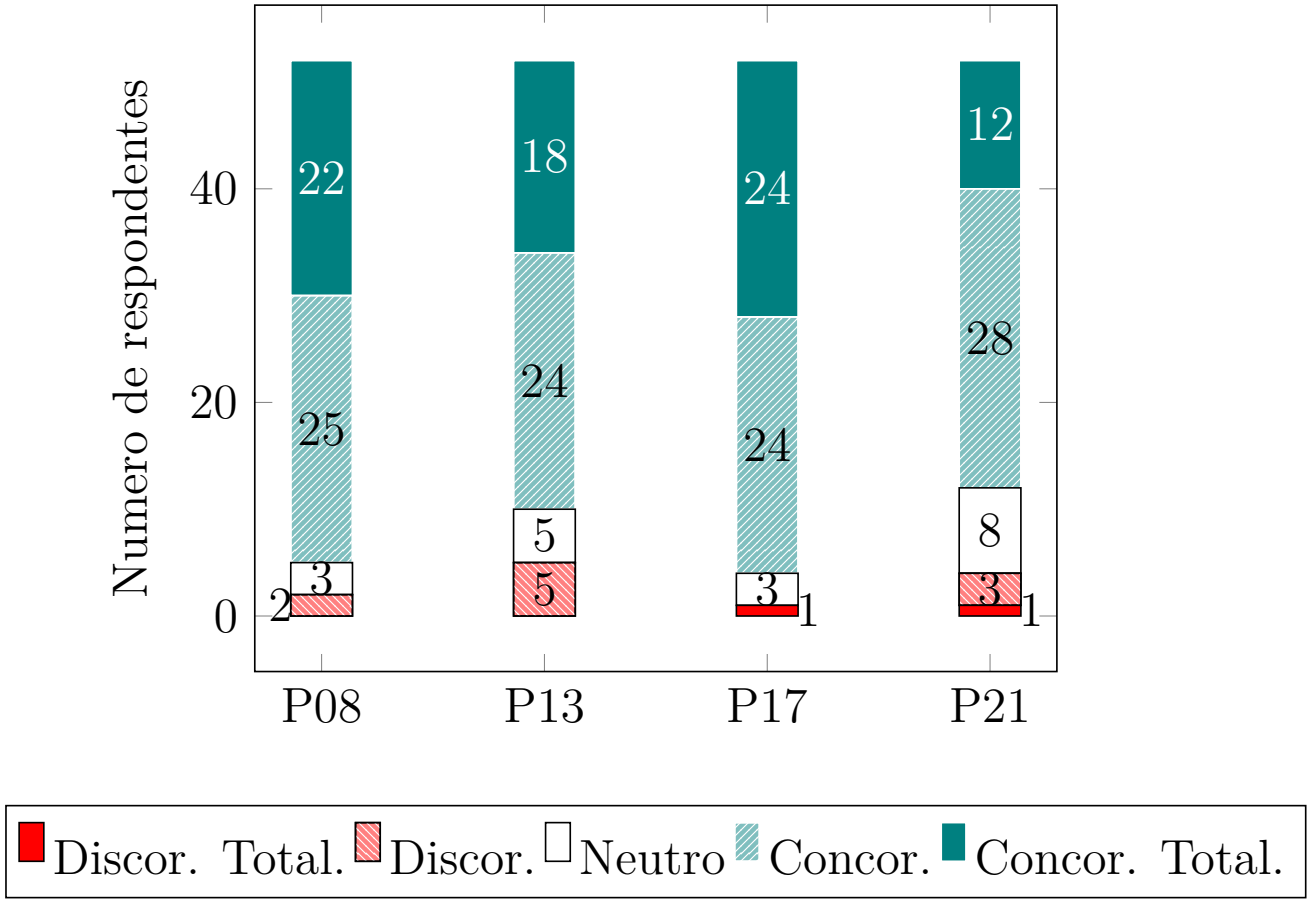


Figura 5.3: Resultado das questões P8, P13, P17 e P21.

As perguntas P09 e P10 apresentaram questionamentos relacionados à presença de ferramentas que podem auxiliar no desenvolvimento de sistemas envolvendo RE e BD. A maioria dos participantes 42,3% (22/52) afirmou não utilizar ferramenta neste contexto para apoiar as suas atividades diárias, outros 28,8% (15/52) não souberam informar e os restantes 28,8% (15/52) afirmaram que utilizam as seguintes ferramentas:

- Engenharia de Software ou de Sistemas:
  - PowerCenter [33];
  - Ferramentas da Rational, RTC e RDNG[4];

[4]https://www.ibm.com/docs/pt-br/engineering-lifecycle-management-suite/lifecycle-management/6.0.2?topic=acica-links-across-project-areas-in-configurations

  - Vários sistemas.

- Ciência de Dados:
  - Ferramenta interna;
  - PowerBI [34];
  - Método ágil, através das histórias de usuários, tarefas e entregas.

- Engenharia de Requisitos:
  - Mediawiki[5];
  - RTC e RQM;
  - RDNG;
  - Jira [35].

As perguntas P11 e P12 questionaram sobre a utilização de ferramentas ligadas à qualidade dos dados e a modelagem de BD. Apenas 15,3% (8/52) dos respondentes afirmaram que utilizam uma ferramenta para essa atividade. A maioria dos participantes 46,1% (24/52) responderam "não", e 38,4% (20/52) não souberam informar. Foram obtidas as seguintes 13,4% (7/52) respostas na P12 quanto às ferramentas indicadas na P11:

- Engenharia de Software ou de Sistemas:
  - Informatica Data Quality[6];
  - Vários Sistemas;

- Ciência de Dados:
  - ERWIN[7] e outras internas;
  - Power Designer [36];
  - Spark [37], nifi[8], elastic search [38];
  - Excel [39], PowerBI [34].

- Engenharia de Requisitos:
  - Pentaho [40];

[5] https://www.mediawiki.org/wiki/MediaWiki
[6] https://www.informatica.com/br/products/data-quality/informatica-data-quality.html
[7] https://www.erwin.com/br-pt/
[8] https://nifi.apache.org/

Em relação à pergunta P14, conforme apresentado na Figura 5.4, foram disponibilizados os 6 (seis) atributos descritos no estudo primário E7[28] de acordo com a ISO/IEC 25010:2011 e uma opção alternativa (outra) para os respondentes indicarem os atributos de qualidade mais adequados na definição da arquitetura de software de BD, conforme a sua percepção. A opção mais escolhida pelos respondentes foi o atributo "Confiabilidade", com 78,8% (41/52) das possíveis marcações, seguido de "Segurança" com 67,3% (35/52), "Eficiência de desempenho" e "Escalabilidade" ambos com 57,7% (30/52), e "Usabilidade" com 40,3% (21/52). Diferente do resultado do E7[28], que apresentou os atributos "Adequação funcional" com 100% das marcações totais, nesta pesquisa, este atributo obteve somente 28,8% (15/52) de marcações. Sobre o campo livre foram informados os seguintes atributos: "Integridade", "Utilidade", e "Privacidade/Controle de Acesso".

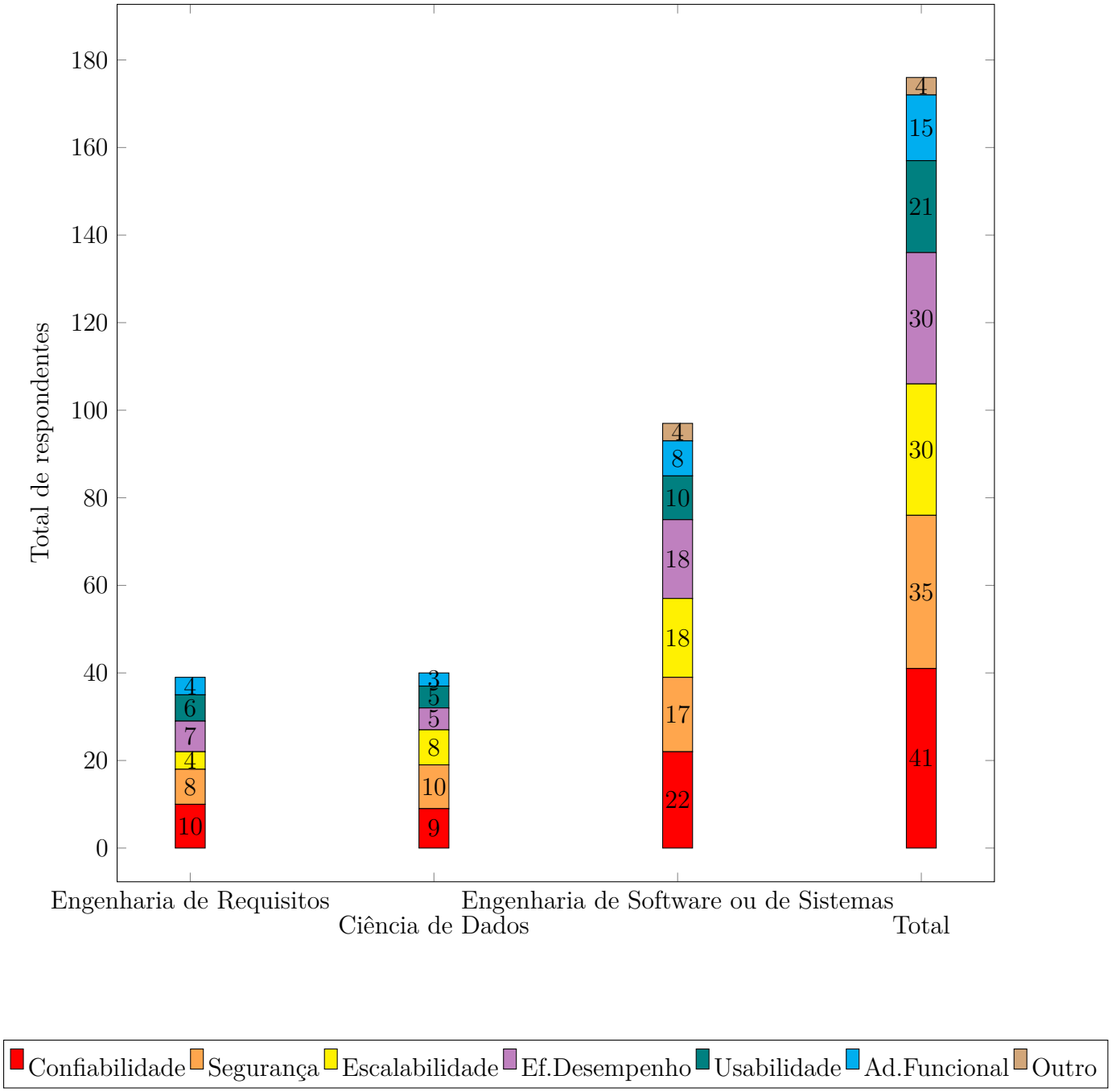


Figura 5.4: Resultado da questão P14.

A pergunta P15 é uma questão aberta e solicitava aos participantes relatarem sua percepção sobre os principais desafios e problemas no contexto de BD. Os participantes informaram os seguintes desafios e problemas, apresentados por área:

- Engenharia de Software ou de Sistemas:
  - *Tempo de resposta*;
  - *Excesso de dados que serão descartados*;

- *Dificuldade na obtenção dos requisitos funcionais para que a modelagem seja adequada, muitas vezes os gestores querem ter um volume absurdo de informações e nem sabem o que farão com elas*;
- *Escalabilidade, Ética, vieses sociais*;
- *Estruturar dados e mudar cultura de análise de dados*;
- *Desempenho, armazenamento (especialmente se for em Nuvem), controle de acesso, privacidade, integridade dos dados, padronização dos dados coletados de diversas fontes*;
- *Confiabilidade de dados externos*;
- *Identificar e unificar as informações das várias fontes de dados tornando útil o conjunto de informações*;
- *Identificar filtros e padrões nos dados que tenham algum significado, representam alguma informação para o negócio da empresa ou área de interesse. Um engenheiro inexperiente pode facilmente interpretar ou manipular os dados erroneamente e apresentar resultados que não necessariamente refletem os dados.*;
- *Dar a devida importância ao Big data pela empresa em relação ao desenvolvimento e uso do big data na concorrência com a evolução de funcionalidades do sistema. Normalmente o Big Data fica em prioridade baixa por ser usado normalmente para gerar dados gerenciais e as funcionalidades negociais são priorizadas*;
- *Privacidade dos dados e a falta de conhecimento na área*;
- *Reunir as informações necessárias e definição do escopo*;
- *Preço*;
- *Encontrar bons profissionais para atuar no levantamento de requisitos, e desafios para extração dos dados*;
- *A falta de conhecimento dos técnicos*;
- *Maior desafio é mostrar para tomadores de decisão que big data não é uma solução que vai resolver todo e qualquer problema.*

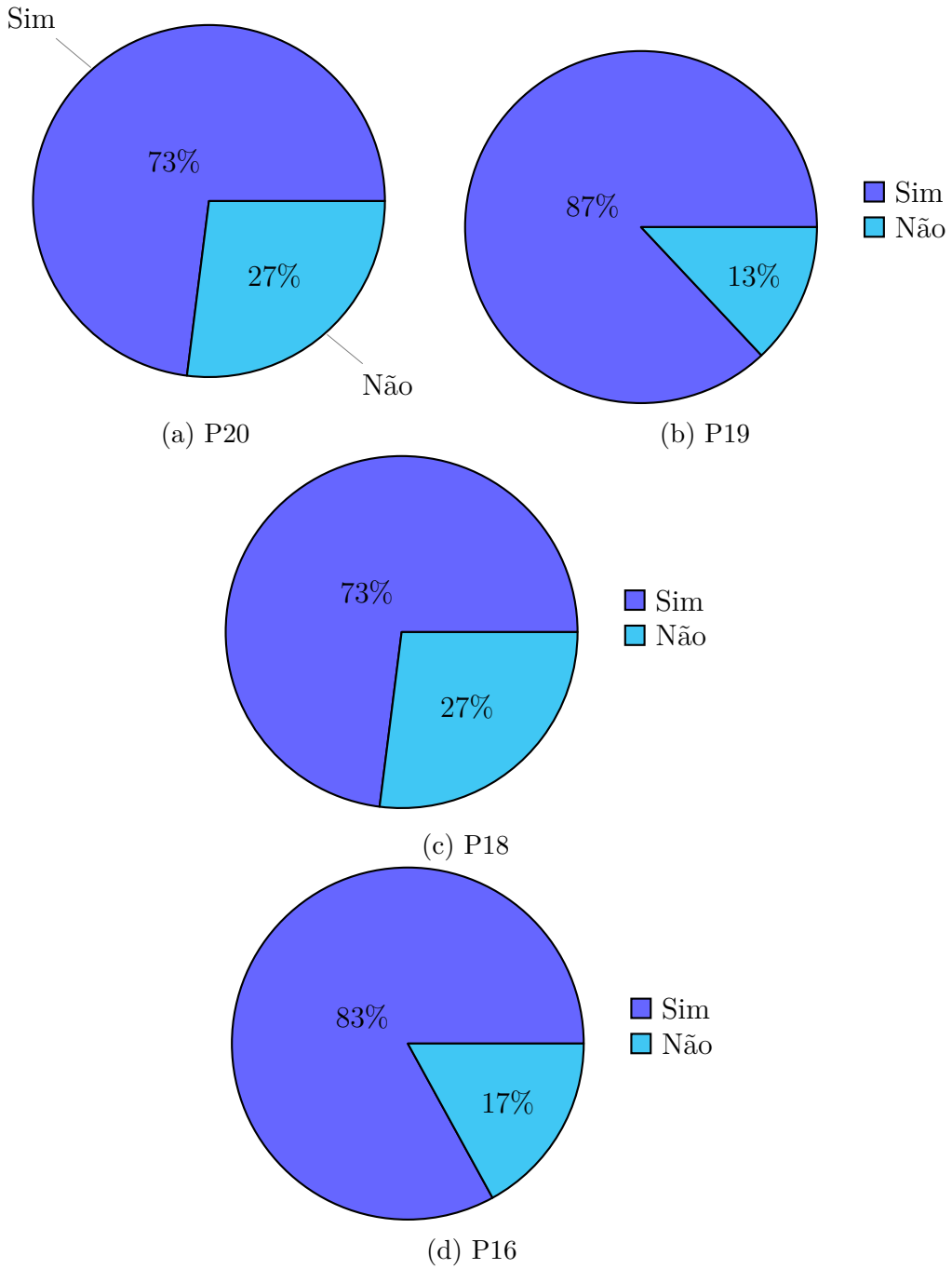
- Ciência de Dados:
  - *O uso ótimo dos recursos pelos usuários*;
  - *As pessoas entenderem de fato o que é o Big Data*;
  - *Falta de maturidade da organização acerca do tema*;

  - *Granularidade*;
  - *Integração das fontes*;
  - *Falta de conhecimento e não entendimento da finalidade da tecnologia*;
  - *Fontes duvidosas*;
  - *Qualificação dos dados*;
  - *Padronizar processos e documentar.*

- Engenharia de Requisitos:
  - *Ter fontes de dados estruturados*;
  - *Vejo que a implementação do Big Data ainda é um desafio*;
  - *Confiabilidade*;
  - *Qualidade e volume dos dados.*

Os 29 (vinte e nove) respondentes que mencionaram os principais desafios e problemas no contexto de BD na questão P15, apresentaram dois tipos de respostas classificadas como referentes à complexidade tecnológica do BD, e as relativas ao conhecimento técnico e organizacional dos envolvidos.

Na seção 3 do questionário foram apresentadas as principais propostas dos estudos selecionados pela SLR, utilizando perguntas em escala likert ou de sim ou não. A ferramenta BiDaML do E8[4], apresentada pela questão P19 (Figura 5.5), foi a mais aceita pelos respondentes da área de Engenharia de software ou de sistema com 86,6% (26/30) (50% da amostra geral), seguido do *framework* para automatização de requisitos descrito na questão P16 com 83,3% (25/30) (48% da amostra geral) de aceitação. Nas questões P18 e P20 que apresentaram o "Diagrama de Caso de Uso Acionável" e o modelo de *Big Data Virtual*, respectivamente, obtiveram o mesmo percentual de aceitação sendo ele de 73,3% (22/30) (42,3% da amostra geral), conforme apresentado nos gráficos das questões P18 e P20 (Figura 5.5).



(a) P20 (b) P19

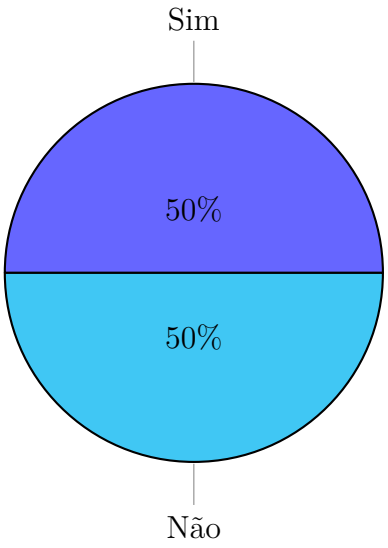
(c) P18

(d) P16

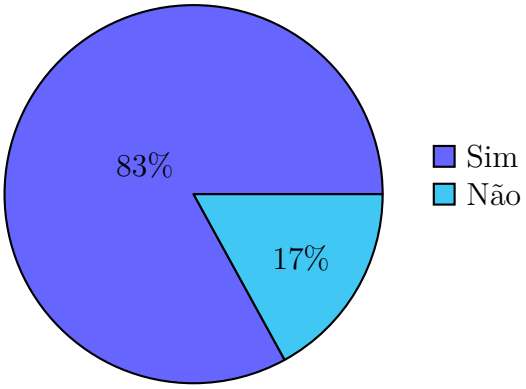
Figura 5.5: Resultado das questões P16, P18, P19 e P20 dos participantes de Engenharia de Software ou de Sistemas.

Os resultados referentes aos profissionais da área de Ciência de Dados foram diferentes em relação aos engenheiros de software ou de sistemas, a aceitação da ferramenta BiDaML (P19 da Figura 5.6) foi a maior com 83,3% (10/12) (19,2% da amostra geral) , seguido do *framework* para automatização de requisitos (P16 da Figura 5.6) com 66,6% (8/12) (15,3% da amostra geral) de aceitação. Na questão P20 (Figura 5.6), que apresenta o *Big Data Virtual* houve uma aceitação de 50% (6/12) (11,5% da amostra geral), contudo,

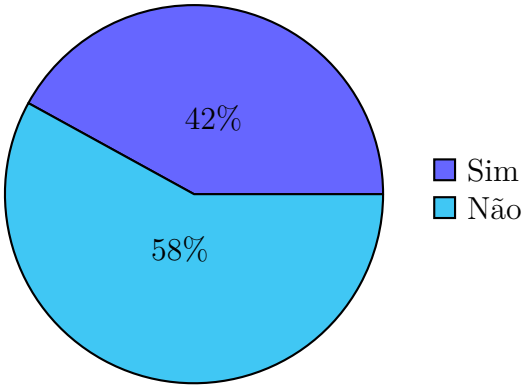
na questão P18 (Figura 5.6) sobre “Diagrama de Caso de Uso Acionável” houve uma aceitação de apenas 41,6% (5/12) (9,6% da amostra geral).



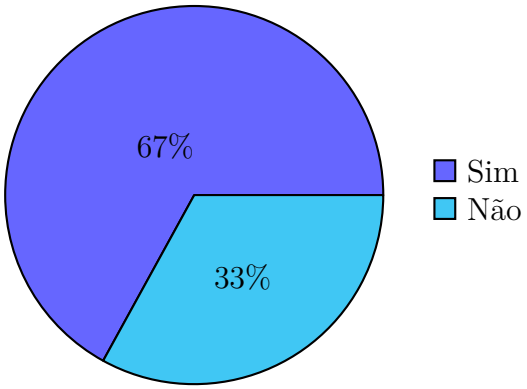
(a) P20



(b) P19



(c) P18



(d) P16

Figura 5.6: Resultados das questões P16, P18, P19 e P20 dos participantes de Ciência de Dados.

Entretanto, os resultados referentes a Engenharia de Requisitos, obtiveram uma grande variação. O mais aceito foi o modelo de *Big Data Virtual* da P20 (Figura 5.7) com 100% (10/10) (19,2% da amostra geral), seguido pela aceitação do *framework* para automatização de requisitos da P16 (Figura 5.7) e da ferramenta BiDaML P19 (Figura 5.7) em que ambos os casos foi de 80% (8/10) (15,3% da amostra geral), o "Diagrama de Caso de Uso Acionável" obteve uma aceitação de apenas 50% (5/10) (9,6% da amostra geral), conforme apresentado na questão P18 da Figura 5.7.

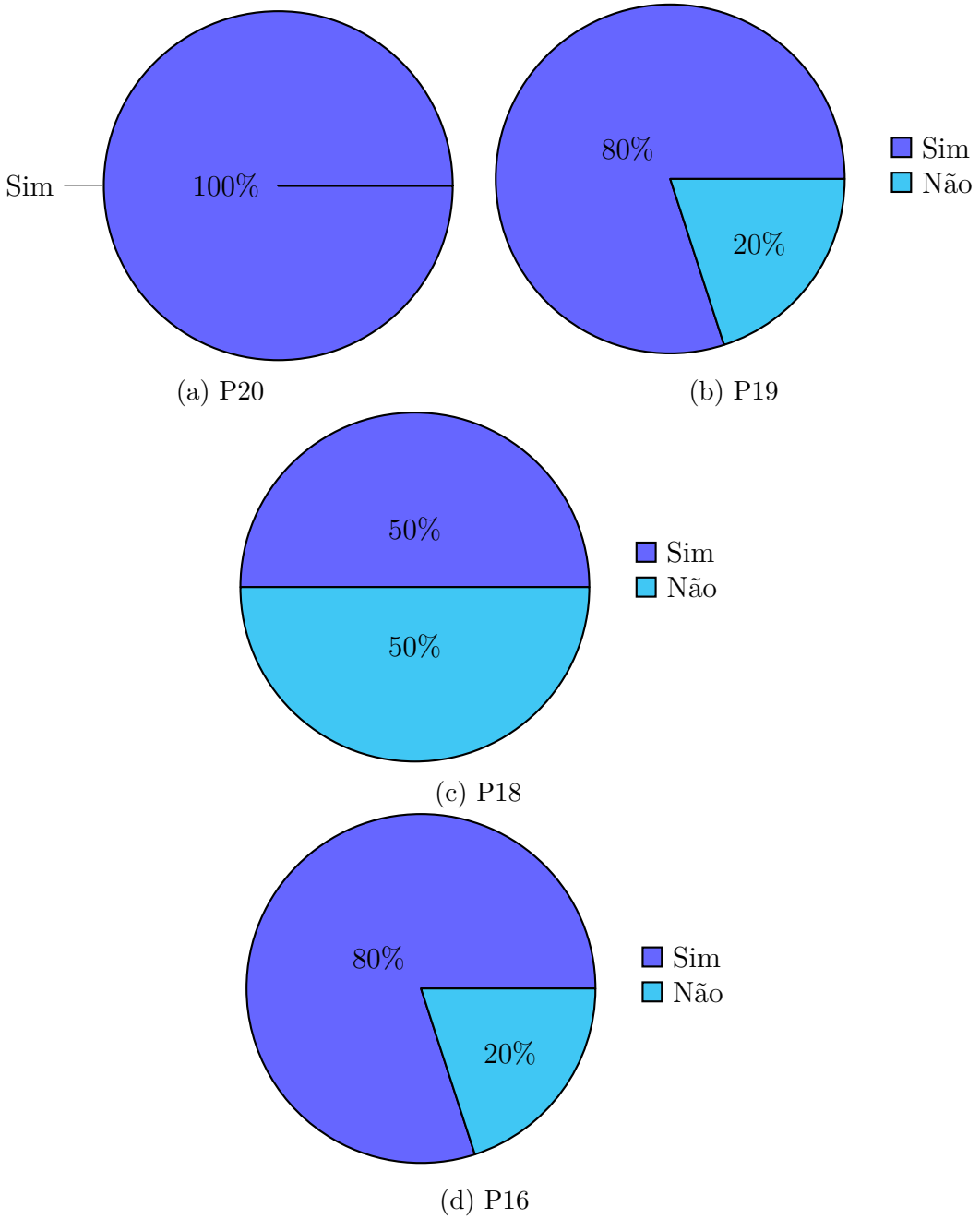


Figura 5.7: Resultados das questões P16, P18, P19 e P20 dos participantes de Engenharia de Requisitos.

Em relação à pergunta P22, a proposta mais selecionada pelo total dos participantes foi a ferramenta BiDaML identificada na questão P19 (Figura 5.8), tanto para os Engenheiros de Software ou de Sistemas com 36,6% (11/30) (21,1% da amostra geral), quanto para os Cientistas de Dados com 33,3% (4/12) (7,7% da amostra geral) das escolhas, no entanto, os Engenheiros de Requisitos determinaram o *framework* para automatização de requisitos identificado na questão P16 como o melhor dentre as opções, com 40% (4/10) (7,7% da amostra geral), de acordo com o apresentado na Figura 5.8.

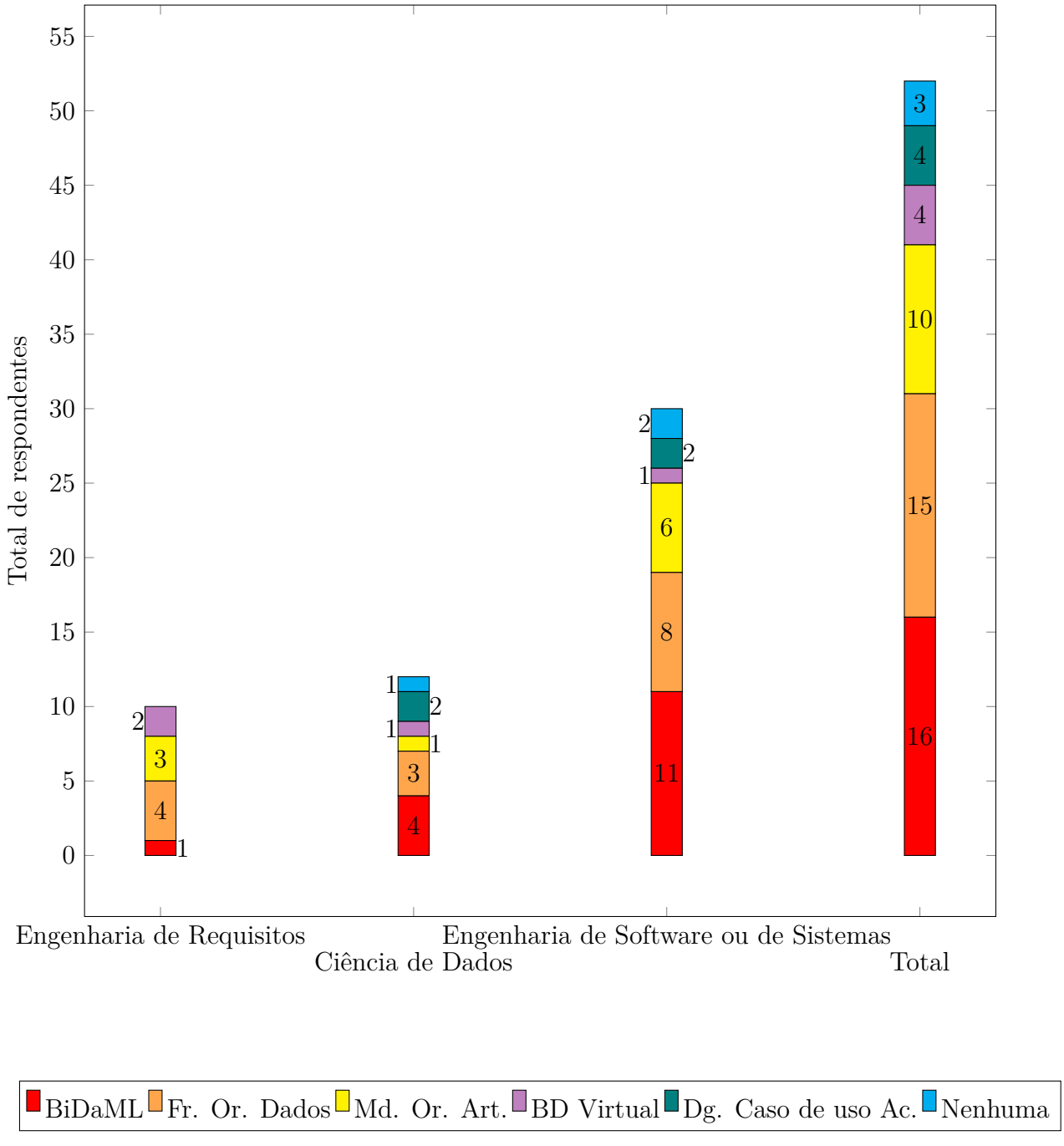


Figura 5.8: Resultado da questão P22.

Ao analisarmos o total das escolhas direcionadas a P22 na Figura 5.8, identificamos que uma ferramenta para a organização do processo é relevante para a realidade dos respondentes, bem como um *framework* para o levantamento de requisitos orientado a dados de fontes externas, ambas as propostas procuram facilitar o processo de RE em BD e garantir a sua validez.

5.2 Discussão dos Resultados em Resposta à RQ.3

Os resultados analisados na seção 4.1 foram utilizados para responder a questão de pesquisa: RQ.3: Quais destes métodos, técnicas e ferramentas de RE em BD selecionados são utilizados total ou parcialmente nas instituições financeiras? conforme descrito a seguir:

Os métodos e técnicas relatados por 40,3% (21/52) dos respondentes do *survey* foram: Caso de uso (UML), História do usuário (Método ágil) e "UC e HU", interpretado como a combinação de "Use Case" e "História de Usuário" de acordo com o apresentado na Figura 5.1. De acordo com os resultados do *survey* e dos estudos primários identificados na SLR, temos:

- E1[1] sugere o *framework* de elicitação automática de requisitos;
- E4[2] propõe o modelo IRIS descrito pelo *Big Data Virtual*;
- E5[3] propõe a integração dos processos RE com os de BD utilizando "Diagrama de Caso de Uso Acionável";
- E7[28] descreve a importância da identificação de requisitos não funcionais antes de projetar o produto;
- E8[4] promove a ferramenta *online* colaborativa BiDaML;
- E9[5] descreve o modelo BD-REAM orientado a artefatos de requisitos.

Podemos inferir que nenhuma das 6 (seis) propostas destacadas especificamente no *survey* são utilizadas em sua totalidade pelos profissionais das instituições financeiras que participaram do *survey*. Entretanto, as ferramentas mencionadas possuíram uma considerável aceitação entre os respondentes da amostra, o que pode sugerir a possibilidade de uso futuro pelas instituições.

Houve mais de 84% de aceitação para o BiDaMl (E8[4]) e para o *framework* de elicitação automática de requisitos (E1[1]), cerca de 78% (41/52). Mais de 76% (40/52) de concordância para o modelo orientado a artefato (E9[5]), para o *Big Data Virtual* (E4[2]) cerca de 73% (38/12) e mais de 61% (32/52) para o "Diagrama de Caso de Uso Acionável" (E5[3]) .

Também foi possível identificar que houve mais de 90% (48/52) de concordância na proposta de atuação conjunta entre o engenheiro de requisitos e do cientista de dados no desenvolvimento de projetos com BD, constante no E1[1],E5[3] e E8[4] de acordo com a Tabela 4.2.

Ademais, os questionamentos acerca do E1[1], o qual apresenta um *framework* para elicitação automática de requisitos, demonstram uma necessidade desta proposta para captar requisitos de fontes externas à organização, sendo elas de natureza humana ou de máquinas. Além de uma aceitação de cerca de 78% (41/52) sobre o conceito, mais de 90% (47/52) dos respondentes concordaram com a importância da identificação dos requisitos originários das fontes externas para evoluir e desenvolver novas oportunidades de negócio. Em contrapartida, cerca de 63% (33/52) dos respondentes do *survey* alegaram que não utilizam ou não sabem se utilizam dados de fontes externas à organização, o que pode ressaltar a importância da aplicação da proposta de E1[1].

Apenas 28,8% (15/52) dos respondentes registraram possuir ferramentas para comunicar, documentar o processo e o resultado da análise de dados, o que agregaria importância com a utilização do BiDaML (E8[4]). Outro caso se refere a ferramenta utilizada para avaliar a qualidade dos dados e, consequentemente, a modelagem de BD, em que apenas 15,3% (8/52) dos participantes alegaram possuir tal funcionalidade, viabilizando também a proposta do E4[2].

Além disso, cerca de 81% (42/52) dos respondentes concordaram com uma das principais inferências do E7[28], o qual propõe a necessidade de analisar e identificar todos os requisitos não funcionais, que garantam a comunicação e a coordenação entre os componentes, conectores e restrições arquiteturais, devido a arquitetura mais complexa de BD. Pode-se concluir que, para essa amostra, requisitos de qualidade tem uma grande importância para a composição de um sistema confiável, seguro e funcional.

Portanto, é importante ressaltar que os estudos selecionados na SLR apresentados no Capítulo 2 não são utilizadas completamente pelas instituições financeiras que participaram do *survey*, contudo, poderiam vir a ser utilizadas, de acordo com a amostra obtida, além de facilitarem o desenvolvimento de projetos no contexto de BD.

5.3 Ameaças a Validade e Limitações do Estudo

Uma das ameaças internas envolve o quantitativo dos métodos, técnicas e ferramentas selecionados na literatura, uma vez que não se pode garantir que todos os estudos primários foram identificados e analisados durante a SLR. Para minimizar essa ameaça, as publicações nas bases selecionadas foram monitoradas até a última semana de conclusão da revisão da literatura.

Outra ameaça de caráter interno está relacionada a possibilidade haver viés dos participantes ao responder às questões da *survey*, ao não apresentar suas reais percepções e preocupação com as regras organizacionais de segurança da informação. A fim de minimizar esta ameaça foi enviado um texto de apoio que retrata todo caráter do processo, e a garantia total da proteção do anonimato do respondente e das instituições investigadas.

Uma das ameaças externas se refere a generalização dos resultados adquiridos por meio da aplicação da *survey*, assim como o número obtido de participantes. Por isso, o questionário foi aplicado em diferentes instituições financeiras com o intuito de diversificar as respostas e descobrir opiniões diversas em relação aos assuntos abordados. Para se obter maior quantitativo de respostas, foi possível expandir a pesquisa para as equipes terceirizadas ligadas às soluções de BD das instituições investigadas. Além disso, o trabalho não pretende generalizar os resultados obtidos com o *survey*, mas apresentar a percepção de profissionais com relação aos modelos estudados.

A principal limitação do trabalho foi a apresentação das ideias centrais das ferramentas constantes na terceira seção do questionário, no qual foram descartadas as figuras e a extensa explicação sobre cada uma delas. O objetivo foi tornar o questionário mais sucinto e direto conforme orientação dos profissionais das instituições financeiras que auxiliaram na validação do conteúdo do *survey*.

Capítulo 6

Conclusão

Diante do volume de investigações em BD, da complexidade tecnológica e sua importância para os negócios das organizações, este trabalho realizou uma revisão sistemática da literatura em dezembro de 2022, com a finalidade de identificar abordagens, métodos, técnicas e ferramentas de RE em BD e investigar, por meio de *survey*, sua utilização total ou parcial em instituições financeiras nacionais de grande porte. Foram selecionados 11 (onze) estudos na literatura, efetuado uma análise comparativa entre eles, apontando seus aspectos comuns, vantagens e desvantagens. O *survey* considerou as propostas específicas de 6 (seis) estudos, abrangendo aspectos gerais dos 5 (cinco) restantes, totalizando 22 (vinte e duas) questões, que obteve 52 (cinquenta e dois) respondentes das áreas de Engenharia de Software ou de Sistema, Ciência de Dados e Engenharia de Requisitos, durante 35 (trinta e cinco) dias em que esteve disponível.

A partir da análise dos dados resultantes do *survey*, pode-se concluir que as propostas de todos os estudos são relevantes para as instituições financeiras investigadas, pois obtiveram mais de 60% de aceitação e mais de 70% de concordância. Destaque para ferramenta web colaborativa abrangendo todo o ciclo de desenvolvimento de projetos de BD, assim como um *framework* para a elicitação automática de requisitos orientado a dados de fontes externas à organização.

Por outro lado, nenhuma das 6 (seis) propostas destacadas especificamente no *survey* são utilizadas em sua totalidade nas instituições analisadas. No entanto, poderiam vir a ser, além de facilitarem o desenvolvimento de projetos de BD. Além disso, 29 (vinte e nove) participantes do *survey* consideraram que os principais desafios e problemas de BD envolvem tanto a qualidade dos profissionais responsáveis quanto a complexidade de lidar com o BD.

Ficou demonstrado pouco uso de ferramentas que podem auxiliar no desenvolvimento de sistemas envolvendo RE e BD. O que pode ter ratificado a considerável aceitação das propostas selecionadas da literatura, assim como a aderência da academia no atendimento

às necessidades do mercado.

Como trabalhos futuros, seria oportuno realizar uma pesquisa de caráter qualitativo, com a finalidade de apresentar as referidas propostas em sua totalidade aos entrevistados, além de diversificar as organizações a serem investigadas e atingir uma maior abrangência nacional.

Referências

[1] Henriksson, Aron e Jelena Zdravkovic: *Holistic data-driven requirements elicitation in the big data era*. Softw. Syst. Model., 21(4):1389–1410, 2022. https://doi.org/10.1007/s10270-021-00926-6. ix, 3, 12, 13, 14, 15, 16, 19, 34, 35, 37, 39, 40, 41, 45, 54, 71, 72

[2] Park, Grace, Lawrence Chung, Haan Johng, Vijayan Sugumaran, Sooyong Park, Liping Zhao e Sam Supakkul: *A big data conceptual model to improve quality of business analytics*. Em Dalpiaz, Fabiano, Jelena Zdravkovic e Pericles Loucopoulos (editores): *Research Challenges in Information Science - 14th International Conference, RCIS 2020, Limassol, Cyprus, September 23-25, 2020, Proceedings*, volume 385 de *Lecture Notes in Business Information Processing*, páginas 20–37. Springer, 2020. https://doi.org/10.1007/978-3-030-50316-1_2. ix, 12, 13, 20, 22, 34, 35, 37, 39, 40, 41, 42, 45, 54, 71, 72

[3] Kourla, Sandhya Rani, Eesha Putti e Mina Maleki: *REBD: A conceptual framework for big data requirements engineering*. CoRR, abs/2006.11195, 2020. https://arxiv.org/abs/2006.11195. ix, 13, 22, 24, 34, 35, 37, 39, 40, 43, 54, 71, 72

[4] Khalajzadeh, Hourieh, Andrew J. Simmons, Tarun Verma, Mohamed Abdelrazek, John C. Grundy, John G. Hosking, Qiang He, Prasanna Ratnakanthan, Adil Zia e Meng Law: *Bidaml in practice: Collaborative modeling of big data analytics application requirements*. Em Ali, Raian, Hermann Kaindl e Leszek A. Maciaszek (editores): *Evaluation of Novel Approaches to Software Engineering - 15th International Conference, ENASE 2020, Prague, Czech Republic, May 5-6, 2020, Revised Selected Papers*, volume 1375 de *Communications in Computer and Information Science*, páginas 106–129. Springer, 2020. https://doi.org/10.1007/978-3-030-70006-5_5. ix, 13, 14, 27, 29, 34, 36, 37, 38, 39, 40, 44, 45, 54, 66, 71, 72

[5] Arruda, Darlan, Nazim H. Madhavji e Ibtehal Noorwali: *A validation study of a requirements engineering artefact model for big data software development projects*. Em Sinderen, Marten van e Leszek A. Maciaszek (editores): *Proceedings of the 14th International Conference on Software Technologies, ICSOFT 2019, Prague, Czech Republic, July 26-28, 2019*, páginas 106–116. SciTePress, 2019. https://doi.org/10.5220/0007927201060116. ix, 13, 14, 29, 31, 34, 36, 37, 38, 39, 40, 41, 43, 54, 71

[6] Kozmina, Natalija, Laila Niedrite e Janis Zemnickis: *Perspectives of information requirements analysis in big data projects*. Em Lupeikiene, Audrone, Olegas Vasilecas e Gintautas Dzemyda (editores): *Databases and Information Systems X - Selected Papers from the Thirteenth International Baltic Conference, DB&IS 2018,*

Trakai, Lithuania, July 1-4, 2018, volume 315 de *Frontiers in Artificial Intelligence and Applications*, páginas 109–124. IOS Press, 2018. https://doi.org/10.3233/978-1-61499-941-6-109. ix, 11, 13, 14, 31, 34, 36, 38, 39

[7] Zowghi, Didar e Chad Coulin: *Requirements elicitation: A survey of techniques, approaches, and tools*. Em *Engineering and managing software requirements*, páginas 19–46. Springer, 2005. 1, 34

[8] Kourla, Sandhya Rani e Eesha Putti: *Importance of process mining for big data requirements engineering*. International Journal of Computer Science & Information Technology (IJCSIT), 12(4), agosto 2020. https://ssrn.com/abstract=3685040. 1, 2, 3

[9] Arruda, Darlan: *Requirements engineering in the context of big data applications*. ACM SIGSOFT Softw. Eng. Notes, 43(1):1–6, 2018. https://doi.org/10.1145/3178315.3178323. 1, 3

[10] Georgiadis, Georgios e Geert Poels: *Towards a privacy impact assessment methodology to support the requirements of the general data protection regulation in a big data analytics context: A systematic literature review*. Computer Law & Security Review, 44:105640, 2022. 1

[11] Georgiadis, Georgios e Geert Poels: *Towards a privacy impact assessment methodology to support the requirements of the general data protection regulation in a big data analytics context: A systematic literature review*. The computer law and security report, 44:105640, 2022, ISSN 0267-3649. 1

[12] Arruda, Darlan e Rodrigo Laigner: *Requirements engineering practices and challenges in the context of big data software development projects: Early insights from a case study*. Em Wu, Xintao, Chris Jermaine, Li Xiong, Xiaohua Hu, Olivera Kotevska, Siyuan Lu, Weija Xu, Srinivas Aluru, Chengxiang Zhai, Eyhab Al-Masri, Zhiyuan Chen e Jeff Saltz (editores): *2020 IEEE International Conference on Big Data (IEEE BigData 2020), Atlanta, GA, USA, December 10-13, 2020*, páginas 2012–2019. IEEE, 2020. https://doi.org/10.1109/BigData50022.2020.9377734. 1, 13, 25, 34, 36, 37, 39, 40, 45

[13] Coda, Felipe A., Rafael M. Salles, Henrique A. Vitoi, Marcosiris A. O. Pessoa, Lucas A. Moscato, Diolino J. Santos Filho, Fabrício Junqueira e Paulo E. Miyagi: *Big data on machine to machine integration's requirement analysis within industry 4.0*. Em Camarinha-Matos, Luis M., Ricardo Almeida e José Oliveira (editores): *Technological Innovation for Industry and Service Systems - 10th IFIP WG 5.5/SOCOLNET Advanced Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS 2019, Costa de Caparica, Portugal, May 8-10, 2019, Proceedings*, volume 553 de *IFIP Advances in Information and Communication Technology*, páginas 247–254. Springer, 2019. https://doi.org/10.1007/978-3-030-17771-3_21. 1

[14] Zaparoli, Wagner: *Engenharia de requisitos: um fundamento na construção de sistemas de informação*. Exacta, 1(0):97–108, 2003, ISSN 1983-9308. https://periodicos.uninove.br/exacta/article/view/522. 1

[15] Aburawi, Yousef e Albaour, Abdulbaset: *Big data: Review paper*. International Journal Of Advance Research And Innovative Ideas In Education, 7, fevereiro 2021. 2

[16] Ishwarappa e J. Anuradha: *A brief introduction on big data 5vs characteristics and hadoop technology*. Procedia Computer Science, 48:319–324, 2015, ISSN 1877-0509. https://www.sciencedirect.com/science/article/pii/S1877050915006973, International Conference on Computer, Communication and Convergence (ICCC 2015). 2

[17] Arruda, Darlan e Nazim H. Madhavji: *State of requirements engineering research in the context of big data applications*. Em Kamsties, Erik, Jennifer Horkoff e Fabiano Dalpiaz (editores): *Requirements Engineering: Foundation for Software Quality - 24th International Working Conference, REFSQ 2018, Utrecht, The Netherlands, March 19-22, 2018, Proceedings*, volume 10753 de *Lecture Notes in Computer Science*, páginas 307–323. Springer, 2018. https://doi.org/10.1007/978-3-319-77243-1_20. 2, 3, 13, 14, 32, 35, 37, 38, 40

[18] Ageed, Zainab Salih, Subhi RM Zeebaree, Mohammed Mohammed Sadeeq, Shakir Fattah Kak, Hazha Saeed Yahia, Mayyadah R Mahmood e Ibrahim Mahmood Ibrahim: *Comprehensive survey of big data mining approaches in cloud systems*. Qubahan Academic Journal, 1(2):29–38, 2021. 2

[19] ABES: *Abes apresenta tendências para o mercado brasileiro de software em 2022 | abes*, maio 2022. https://abes.com.br/abes-apresenta-tendencias-para-o-mercado-brasileiro-de-software-em-2022/\sharp:~:text=%E2%80%9CA%20tend%C3%AAncia%2C%20para%202022%2C, acesso em 2022-09-04. 3

[20] Altarturi, Hamza Hussein, Keng Yap Ng, Mohd Izuan Hafez Ninggal, Azree Shahrel Ahmad Nazri e Abdul Azim Abd Ghani: *A requirement engineering model for big data software*. Em *2017 IEEE Conference on Big Data and Analytics (ICBDA)*, páginas 111–117. IEEE, 2017. 3, 22

[21] Corallo, Angelo, Anna Maria Crespino, Mariangela Lazoi e Marianna Lezzi: *Model-based big data analytics-as-a-service framework in smart manufacturing: A case study*. Robotics Comput. Integr. Manuf., 76:102331, 2022. https://doi.org/10.1016/j.rcim.2022.102331. 3

[22] Lim, Sachiko, Aron Henriksson e Jelena Zdravkovic: *Data-driven requirements elicitation: A systematic literature review*. SN Comput. Sci., 2(1):16, 2021. https://doi.org/10.1007/s42979-020-00416-4. 3, 12, 13, 19, 34, 35, 37, 39, 54

[23] Molléri, Jefferson Seide, Kai Petersen e Emilia Mendes: *Survey guidelines in software engineering: An annotated review*. Em *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM 2016, Ciudad Real, Spain, September 8-9, 2016*, páginas 58:1–58:6. ACM, 2016. https://doi.org/10.1145/2961111.2962619. 5, 45, 55

[24] Keele, Staffs *et al.*: *Guidelines for performing systematic literature reviews in software engineering*. Relatório Técnico, Technical report, ver. 2.3 ebse technical report. ebse, 2007. 6, 9

[25] Petticrew, Mark e Helen Roberts: *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons, 2008. 6

[26] Scannavino, Katia Romero Felizardo, Elisa Yumi Nakagawa, Sandra Camargo Pinto Ferraz Fabbri e Fabiano Cutigi Ferrari: *Revisão Sistemática da Literatura em Engenharia de Software: teoria e prática*. Elsevier, 2017. 7

[27] Dave, Dev, Angelica Celestino, Aparna S. Varde e Vaibhav K. Anu: *Management of implicit requirements data in large SRS documents: Taxonomy and techniques*. SIGMOD Rec., 51(2):18–29, 2022. `https://doi.org/10.1145/3552490.3552494`. 12, 13, 17, 34, 35, 37, 38, 39

[28] Rahman, Md. Saifur e Hassan Reza: *Systematic mapping study of non-functional requirements in big data system*. Em *2020 IEEE International Conference on Electro Information Technology, EIT 2020, Chicago, IL, USA, July 31 - August 1, 2020*, páginas 25–31. IEEE, 2020. `https://doi.org/10.1109/EIT48999.2020.9208288`. 13, 26, 35, 36, 37, 38, 54, 63, 71, 72

[29] Geisberger, Eva, Juergen Kazmeier, Daniel Paulish *et al.*: *Requirements engineering reference model (rem)*. 2006. 29, 41, 43

[30] Punter, Teade, Marcus Ciolkowski, Bernd G. Freimut e Isabel John: *Conducting online surveys in software engineering*. Em *2003 International Symposium on Empirical Software Engineering (ISESE 2003), 30 September - 1 October 2003. Rome, Italy*, páginas 80–88. IEEE Computer Society, 2003. `https://doi.org/10.1109/ISESE.2003.1237967`. 47

[31] Komorita, Samuel S: *Attitude content, intensity, and the neutral point on a likert scale*. The Journal of social psychology, 61(2):327–334, 1963. 48

[32] Linaker, Johan, Sardar Muhammad Sulaman, Martin Höst e Rafael Maiani de Mello: *Guidelines for conducting surveys in software engineering v. 1.1*. Lund University, 50, 2015. 55

[33] Gupta, Abhishek: *A complete reference for informatica power center etl tool*. International Journal of Trend in Scientific Research and Development, 3:1063–1070, 2019. 61

[34] Ferrari, Alberto e Marco Russo: *Introducing Microsoft Power BI*. Microsoft Press, 2016. 62

[35] Li, Patrick: *Jira Software Essentials: Plan, track, and release great applications with Jira Software*. Packt Publishing Ltd, 2018. 62

[36] Capetillo, By Ricardo: *Power designer*. 2008. 62

[37] Salloum, Salman, Ruslan Dautov, Xiaojun Chen, Patrick Xiaogang Peng e Joshua Zhexue Huang: *Big data analytics on apache spark*. International Journal of Data Science and Analytics, 1:145–164, 2016. 62

[38] Gormley, Clinton e Zachary Tong: *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. " O'Reilly Media, Inc.", 2015. 62

[39] Berk, Kenneth N e Patrick Carey: *Data Analysis with Microsoft Excel*. Duxbury Press Pacific Grove, CA, 2000. 62

[40] Roldán, María Carina: *Pentaho 3.2 Data Integration: Beginner's Guide*. Packt Publishing Ltd, 2010. 62