



Pós-Graduação em Ciência da Computação

Dácio Alves Florêncio

Um Projeto Experimental Para a Avaliação de Modelos de Dados de Sistemas de BD NoSQL



Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
<http://cin.ufpe.br/~posgraduacao>

Recife
2019

Dácio Alves Florêncio

**Um Projeto Experimental Para a Avaliação de Modelos de Dados de Sistemas de
BD NoSQL**

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Banco de Dados
Orientadora: Valéria Cesário Times
Coorientador: Sérgio Castelo Branco Soares

Recife
2019

Catálogo na fonte
Bibliotecária Mariana de Souza Alves CRB4-2106

F632p Florêncio, Dácio Alves.

Um Projeto Experimental Para a Avaliação de Modelos de Dados de Sistemas de BD NoSQL/ Dácio Alves Florêncio. – 2019.

105 f.: il., fig., tab.

Orientadora: Valéria Cesário Times.

Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da computação. Recife, 2019.

Inclui referências e apêndices.

1. Banco de Dados. 2. Modelos conceituais. 3. NoSQL. 4. Método Investigativo. I. Times, Valéria Cesário (orientadora). II. Título.

025.04

CDD (22. ed.)

UFPE-MEI 2019-157

Dácio Alves Florêncio

“Um Projeto Experimental Para a Avaliação de Modelos de Dados de Sistemas de BD NoSQL”

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 12 de setembro de 2019.

Orientadora: Profa. Dra. Valéria Cesário Times

BANCA EXAMINADORA

Prof. Dr. Sergio Castelo Branco Soares
Centro de Informática/UFPE
(Co-orientador)

Prof. Dr. Bruno Falcão de Souza Cartaxo
Instituto Federal de Pernambuco / Campus Recife

Profa. Dra. Marizete Silva Santos
Unidade Acadêmica de Educação à Distância e Tecnologia/UFRPE

AGRADECIMENTOS

Gostaria de agradecer a Prof^a Dr^a. Valéria Times por toda a paciência que teve durante todo o mestrado e toda a dedicação que a mesma ofertou em nossas reuniões, sugerindo ideias, artigos para ler. Também ao Prof^o Dr. Sérgio Soares por toda a disponibilidade e ajuda na elaboração do desenho utilizado.

Um agradecimento especial à prof^a Elyda pela disponibilidade para participar das atividades executadas no experimento, e todos os voluntários que cederam algumas horas para participarem dos testes realizados. Um agradecimento à amiga Wanessa que me auxiliou nas dúvidas estatísticas que tive, e toda a disponibilidade que esta teve para sentar e esclarecer as minhas dúvidas. Eu agradeço também aos meus amigos Livya e Clayton pelo companherismo. Também agradeço a Rick por ser essa pessoa tão incrível e ter acreditado em mim mesmo quando eu mesmo não acreditava e obrigado por todo o suporte e ajuda. A todos os meus colegas que me deram de alguma forma uma palavra de apoio, deixo aqui os meus agradecimentos aos professores que me inspiraram.

Finalmente, meus agradecimentos ao meus pais e irmã por sempre me apoiarem, sem o apoio dos mesmos não teria chegado aqui.

RESUMO

Atualmente o uso de sistemas NoSQL se tornaram populares em serviços da Internet, consequentemente criando uma necessidade de se ter um melhor entendimento e compreensão de como os dados são organizados nesses sistemas. O uso de modelos de dados conceituais é o mais indicado para isso, pois, esses modelos podem ajudar usuários a entender como objetos de domínios de aplicação são organizados e como objetos contidos nesses domínios se relacionam sem apresentar detalhes de implementação da aplicação. Contudo, não existem muitos modelos de dados conceituais na literatura que possam representar domínios NoSQL mesmo essa área sendo muito importante para a modelagem de dados. Esse trabalho mostra um estudo experimental que investigou dois modelos conceituais: NoAM e COMN, modelos propostos para sistemas NoSQL. Quatro características desses modelos foram observadas: compreensão, usabilidade, representatividade e qualidade dos modelos. Essa observação foi possível graças ao método investigativo desenvolvido, onde foi definido todo o plano experimental que foi executado para obter as características estudadas. Os achados mostram que o COMN obteve melhores resultados quando observadas duas das características investigadas: compreensão e representatividade. Porém, quando os dados foram submetidos as hipóteses apenas a hipótese relacionada com a compreensão mostrou que houve diferença entre os modelos. Já a segunda métrica que mostrou bons resultados foi a relacionada com a qualidade dos modelos.

Palavras-chaves: Modelos conceituais. NoSQL. Método Investigativo.

ABSTRACT

Nowadays the use of NoSQL systems became popular on most of the Internet services and thus there is a growing need to create a better understanding and comprehension of how the data will be organized in these systems. The use of the conceptual data models are the most indicated to do this because these models can help users to understand how the domains are organized and understanding how objects are related to each other in these domains without presenting implementation details of NoSQL applications. However, there are no many conceptual models in the literature to be used to represent a domain of a NoSQL system even this area being one of the most important area of data modeling. This work shows an experimental study that has investigated two Conceptual Models: NoAM and COMN, which have been proposed to represent NoSQL system's data. Four characteristics of these models were observed in this study: comprehension, usability representativity, and quality of these models. This observation was possible because of the investigative method proposed which has defined all the experimental plan to obtain values for all properties studied. The finds shows that the COMN model has better results for two characteristics observed: comprehension and representativity but when the data obtained was evaluated only the data related with comprehension showed that a difference occurred between the data models. The other characteristic that showed that exists a difference in use the models studied was the quality.

Keywords: Conceptual Models. NoSQL. Investigative Method.

LISTA DE FIGURAS

Figura 1 – Exemplo de um Esquema de Dados	22
Figura 2 – Exemplo de um Esquema de Dados	23
Figura 3 – Exemplo de um Esquema de Dados do Tipo Chave-valor	23
Figura 4 – Exemplo de um Esquema de Dados Orientado à Colunas	24
Figura 5 – Notação do GDBS	33
Figura 6 – Notação do Modelo conceitual proposto	34
Figura 7 – Notação do Modelo ER Enriquecido	35
Figura 8 – Notação NoAM	41
Figura 9 – Notação COMN	42
Figura 10 – Notação COMN Relacionamentos	43
Figura 11 – Resultado Kappa para o NoAM	80
Figura 12 – Resultado Kappa para o COMN	81
Figura 13 – Domínio hospitalar representado no NoAM	102
Figura 14 – Domínio de hospitalar representado no COMN	103
Figura 15 – Comércio eletrônico representado no NoAM	104
Figura 16 – Comércio eletrônico representado no COMN	105

LISTA DE TABELAS

Tabela 1 – Desenho Experimental	58
Tabela 2 – Medianas do COMN e NoAM para o Grupo Presencial em Relação ao Questionário de Compreensão.	71
Tabela 3 – Medianas do COMN e NoAM para o Grupo Remoto em Relação ao Questionário de Compreensão.	72
Tabela 4 – Resultados de usabilidade do grupo presencial.	74
Tabela 5 – Resultados de representatividade do grupo presencial.	76
Tabela 6 – Resultados de usabilidade do grupo remoto.	77
Tabela 7 – Resultados de representatividade do grupo remoto.	78
Tabela 8 – Resultados das medianas da completude de cada modelo.	82
Tabela 9 – Resultados das modas da completude de cada modelo.	82
Tabela 10 – Resultados das modas da corretude de cada modelo.	84
Tabela 11 – Resultados das medianas da corretude de cada modelo.	84

SUMÁRIO

1	INTRODUÇÃO	12
1.1	CONTEXTUALIZAÇÃO	12
1.2	PROBLEMA E QUESTÃO DE PESQUISA	13
1.3	MOTIVAÇÃO	14
1.4	OBJETIVOS	15
1.5	ESTRUTURA DA DISSERTAÇÃO	15
2	REFERENCIAL TEÓRICO	17
2.1	MODELAGEM DE DADOS	17
2.1.1	Níveis na modelagem de dados	17
2.1.2	Análise de modelos conceituais	18
2.2	SISTEMAS DE BD NOSQL	19
2.2.1	Definições básicas	20
2.2.2	Características Principais	21
2.2.3	Modelos de dados de sistemas NoSQL	21
2.3	PLANEJAMENTO DE ESTUDOS EXPERIMENTAIS	25
2.3.1	Características de um experimento	25
2.3.2	Definição dos objetivos da experimentação	26
2.3.2.1	Questionários	26
2.3.3	Execução do experimento	28
2.3.4	Análise dos dados	28
3	TRABALHOS CORRELATOS	30
3.1	MODELOS CONCEITUAIS PARA SISTEMAS NOSQL	30
3.1.1	Modelos de dados baseados no modelo ER (Entidade Relacionamento)	30
3.1.2	Modelos para aplicações híbridas	35
3.1.3	Modelos conceituais baseados em UML	37
3.1.4	Modelagem lógica para representação de dados agregados	41
3.1.5	COMN (Modelagem de Notação Objeto Conceitual)	42
3.1.6	GOOSSDM (Modelo de dados semiestruturado orientado a grafos)	44
3.2	AVALIAÇÃO DE MODELOS DE DADOS POR ESTUDOS EMPÍRICOS	45
3.2.1	Análise do diagrama ER (Entidade Relacionamento)	45
3.2.2	Análise de modelos de dados conceituais	50
3.2.3	Investigação em modelos de dados conceituais	53
3.3	CONSIDERAÇÕES DO CAPÍTULO	56

4	DESENHO EXPERIMENTAL PROPOSTO PARA A AVALIAÇÃO DE MODELOS DE DADOS DE SISTEMAS DE BD NOSQL	58
4.1	DESENHO DO PLANO	58
4.2	DEFINIÇÃO DOS OBJETIVOS DO EXPERIMENTO	59
4.3	DEFINIÇÃO DAS HIPÓTESES	61
4.4	TRATAMENTOS	64
4.5	UNIDADES EXPERIMENTAIS	64
4.6	SUJEITOS EXPERIMENTAIS	64
4.7	VARIÁVEIS	65
4.8	FORMULAÇÃO DOS QUESTIONÁRIOS	65
4.9	AMEAÇAS A VALIDADE	67
4.9.1	Validade de conclusão	68
4.9.2	Validade interna	68
4.9.3	Validade de construção	69
4.9.4	Validade externa	69
5	RESULTADOS	70
5.1	ANÁLISE DOS DADOS OBTIDOS DO QUESTIONÁRIO QUE OBSERVOU A COMPREENSÃO DOS PARTICIPANTES	70
5.1.1	Resultados do grupo presencial ao questionário de compreensão dos modelos de dados	70
5.1.2	Resultados do grupo remoto ao questionário de compreensão dos modelos de dados	71
5.1.3	Teste de hipótese CMD	72
5.2	ANÁLISE DOS DADOS OBTIDOS A PARTIR DO QUESTIONÁRIO APLICADO À USABILIDADE E EXPRESSIVIDADE DOS MODELOS	73
5.2.1	Resultados do grupo presencial ao questionário de usabilidade e expressividade dos modelos de dados	73
5.2.1.1	Investigação da Usabilidade	73
5.2.1.2	Investigação da Representatividade	74
5.2.2	Resultados do grupo remoto ao questionário de usabilidade e expressividade	77
5.2.2.1	Análise da Usabilidade	77
5.2.2.2	Análise da Representatividade	77
5.2.3	Testes de hipótese para UMD e RMD	78
5.3	ANÁLISE DOS DIAGRAMAS MODELADOS	79
5.3.1	Análise da avaliação dos especialistas sobre a completude	82
5.3.2	Análise da avaliação dos especialistas sobre a corretude	83
5.3.3	Teste de hipótese para QMD	85

6	CONCLUSÃO	86
6.1	DISCUSSÃO	86
6.2	CONTRIBUIÇÕES	86
6.3	TRABALHOS FUTUROS	87
6.4	CONCLUSÕES DO TRABALHO	88
	REFERÊNCIAS	89
	APÊNDICE A – SCRIPT DO EXPERIMENTO	95
	APÊNDICE B – QUESTIONÁRIO DE PERFIL	97
	APÊNDICE C – QUESTIONÁRIO DE COMPREENSÃO DOS MODELOS DE DADOS	98
	APÊNDICE D – QUESTIONÁRIO DE USABILIDADE E EXPRESSIVIDADE DOS MODELOS DE DADOS	99
	APÊNDICE E – QUESTIONÁRIO À RESPEITO DA QUALIDADE DOS MODELOS DE DADOS	101
	APÊNDICE F – ESQUEMAS ABORDADOS	102

1 INTRODUÇÃO

Este capítulo relata as principais motivações para a realização deste trabalho, exhibe sua justificativa, apresenta a questão de pesquisa, lista os objetivos de pesquisa almejados e, finalmente, mostra como está estruturado o restante da presente dissertação.

1.1 CONTEXTUALIZAÇÃO

A construção de um modelo conceitual de dados começa através de observações e hipóteses realizadas em uma situação. A partir desta observação são construídos princípios ou axiomas que descrevem a situação observada, permitindo inclusive a possibilidade de uma representação simbólica (BRODIE, 1984). Essas representações simbólicas, por sua vez, utilizam-se de regras pré-determinadas para descrever um conjunto de objetos (SHAW, 1984), onde essa modelagem é feita para especificar todas as propriedades de uma aplicação. Isso inclui modelar: comunicações, interface homem-máquina, o sistema e o domínio de aplicação (BRODIE, 1984). Nesta dissertação apenas serão explicados modelos de dados conceituais desenvolvidos para sistemas de banco de dados (BD).

No contexto de BD, a modelagem de dados é o passo inicial para o desenvolvimento de um BD, pois, essa modelagem facilita o entendimento do projeto de banco de dados através da definição de características chaves que previnem erros de programação e operação (VERA et al., 2015), (CORONEL; MORRIS, 2016). A modelagem de dados é realizada por meio de mecanismos de abstrações que são utilizados para representar detalhes importantes do domínio e para modelar as interações dos objetos deste domínio. Assim, a modelagem de dados auxilia na compreensão das complexidades de um domínio do mundo real (CORONEL; MORRIS, 2016).

O projeto de banco de dados se divide em três níveis: conceitual, lógico e físico. Cada um desses níveis faz uso de metodologias para representar uma aplicação de banco de dados em seus diferentes níveis. O nível conceitual é o que apresenta um maior nível de abstração, pois, neste nível são omitidos detalhes de implementação (ELMASRI; NAVATHE, 2010), (HASHEM; RANC, 2015).

Ao longo dos anos, novos modelos de dados conceituais foram propostos, cada um desses construído para se adequar a uma necessidade. Uma das formas mais difundidas para a construção de projetos conceituais de banco de dados foi através do uso do modelo ER (Entidade Relacionamento), proposto por Chen em 1976 (LUKYANENKO; SAMUEL, 2017). Contudo, segundo Hashem e Ranc (2015), o surgimento da *Web 2.0* trouxe mudanças na produção dos dados gerados, já que os dados passaram pela transformação de dados estruturados para dados semiestruturados ou não estruturados, tais como imagens, textos e vídeos (MCAFEE et al., 2012). De acordo com McCreary e Kelly (2014), o uso do ER e de

ferramentas de modelagem de dados tradicionais tornaram-se inadequadas para representar esses dados, pois, a proposta inicial dessas ferramentas eram representar sistemas de dados estruturados e graças às redes sociais e outros serviços on line, houve uma mudança na estrutura dos dados e na representação de como esses se relacionam (KIM; YU, 2015), (QUE et al., 2016). À vista disto, os modelos de dados e as ferramentas de modelagem até então criados, não conseguiam expressar de forma eficiente os domínios da *Web 2.0* (LI et al., 2016), (AMIRIAN; BASIRI; WINSTANLEY, 2013).

Assim, (VERA et al., 2015), (ZEČEVIĆ et al., 2018), (VILLARI et al., 2016), (HILLS, 2016), propuseram novos modelos de dados para representar domínios da *Web 2.0*. Todavia, na literatura estudada, não foram encontrados estudos que façam uso de um método experimental para avaliar os modelos propostos. Segundo Wohlin et al. (2012), a avaliação experimental é algo natural na ciência, pois quando uma nova técnica ou abordagem é criada, há a necessidade de se saber se a mesma é efetiva para o problema para o qual esta foi proposta.

1.2 PROBLEMA E QUESTÃO DE PESQUISA

A principal função de um modelo de dados é auxiliar a compreensão de um domínio e prover construtores de modelagem de fácil entendimento (GENERO et al., 2001), (HUSSAIN, 2014). A avaliação de um modelo de dados se faz necessária, pois, por meio desta pode-se descobrir se determinado modelo ou técnica de modelagem é efetiva para o domínio para o qual esta foi proposta. Ademais, por meio da avaliação de esquemas construídos para uma aplicação e a partir do uso de construtores de modelos conceituais de dados, pode-se observar se os esquemas auxiliam os usuários a visualizarem todas as informações expressas neste esquema de dados, uma característica muito importante para modelos de dados (GENERO et al., 2001). Além disso, é importante observar se os construtores e as regras propostas pelos modelos são de fácil uso e entendimento, ou seja, intuitivos. Assim, representar de forma clara e correta um domínio é o objetivo de qualquer modelo conceitual de dados (MEHMOOD; CHERFI; COMYN-WATTIAU, 2009), (HUSSAIN, 2014).

Este trabalho propõe um projeto experimental para a avaliação de modelos de dados conceituais para sistemas de BD NoSQL, e para isso, as seguintes perguntas de pesquisa foram elencadas:

- Como medir a qualidade dos modelos de dados conceituais analisados no que diz respeito a corretude e completude dos esquemas modelados a partir de um modelo de dados?
- Os modelos de dados conceituais investigados podem ser ditos modelos de fácil uso e representativos no que diz respeito a visão dos usuários?

Para responder as questões de pesquisa dessa dissertação, foi proposto um estudo experimental, onde dois modelos de dados desenvolvidos para sistemas NoSQL foram investigados. Os modelos investigados foram: o COMN (Modelagem de Notação Objeto Conceitual) proposto em Hills (2016) e o NoaM (Modelo Abstrato NoSQL) proposto em Bugiotti et al. (2014).

1.3 MOTIVAÇÃO

A realização de estudos empíricos é uma das melhores formas de se avaliar o quão promissor ou efetivo é uma metodologia ou processo (SARIS; GALLHOFER, 2014). Embora os resultados de estudos empíricos para avaliação de modelos de dados tenham se mostrado promissores em (GENERO et al., 2001), (HUSSAIN et al., 2006), (POELS et al., 2011), (HUSSAIN, 2014), tal prática ainda não é usual na área de banco de dados.

Não foram encontrados na literatura estudada, trabalhos que dizem respeito a avaliação experimental de modelos de dados desenvolvidos para sistemas NoSQL. Assim, com base nos estudos presentes na Seção 3.2 desta dissertação, onde outros autores propuseram formas de avaliações de modelos conceituais direcionados para outros sistemas de banco de dados, foram definidas quais propriedades seriam avaliadas nesta dissertação:

- Simplicidade do modelo de dados – Segundo Fettke e Loos (2003), esta pode ser medida a partir do fácil manuseio do modelo e do quão claro este consegue expressar um domínio.
- Representatividade do modelo de dados – De acordo com Zheng et al. (2014) esta característica está relacionada ao fato do domínio abordado conseguir ou não ser mapeado pelo modelo avaliado, ou seja, se todas as características do domínio real estão devidamente representadas no modelo de dados investigado.
- Qualidade do modelo de dados – Característica utilizada para analisar se as abstrações envolvidas na construção de um esquema de dados baseado no modelo estudado vão ao encontro da visão do usuário (POELS et al., 2011), isto é, se todas as características do domínio foram corretamente e completamente abordadas no esquema construído. Além disso, esta propriedade visa observar se os esquemas construídos são independentes do paradigma de implementação, pois, modelos de dados conceituais não deve possuir características da plataforma de implementação. Dessa forma, um modelo de dados conceitual que seja classificado como conceitual, e possua dependência da plataforma de implementação quebra uma característica importante definida para modelos de dados conceituais, assim, podendo ser classificado como um modelo de dados com baixa qualidade, por não cumprir essa premissa (POELS et al., 2011).

A forma como cada uma dessas propriedades serão obtidas e analisadas são discutidas no Capítulo 4 desta dissertação. Além disso, no Capítulo 4, é explicado também como cada uma das propriedades foram quantificadas para produzir uma avaliação objetiva dos modelos conceituais estudados neste trabalho.

1.4 OBJETIVOS

O objetivo geral deste trabalho foi analisar o quanto a escolha do modelo de dados interfere na compreensão e no desenvolvimento de projetos conceituais de banco de dados NoSQL.

Para a obtenção do objetivo geral desta dissertação, os seguintes objetivos específicos foram definidos:

- Construir um método de avaliação experimental para modelos conceituais de sistemas NoSQL;
- Definir as propriedades analisadas na avaliação experimental;
- Selecionar formas de mensurar a representatividade, simplicidade e qualidade dos modelos;
- Elaborar um plano experimental;
- Realizar experimento utilizando a metodologia criada no plano experimental;
- Analisar os resultados obtidos a partir do experimento utilizando conceitos estatísticos;
- Compreender o comportamento das características analisadas para os modelos investigados.

1.5 ESTRUTURA DA DISSERTAÇÃO

O restante desta dissertação está organizada como segue:

- Capítulo 2 - Referencial Teórico: Dá embasamento de tópicos importantes para a compreensão do trabalho.
- Capítulo 3 – Trabalhos Correlatos: Apresenta estudos realizados na área de modelagem de dados de sistemas NoSQL e discorre sobre como avaliar modelos de dados.
- Capítulo 4 - Desenho Experimental Proposto para a Avaliação de Modelos de Dados de Sistemas de BD NoSQL: Neste capítulo, é apresentado como o experimento descrito nesta dissertação foi planejado. Assim, nesse capítulo, são abordados como as métricas foram definidas e analisadas, quais os objetivos do experimento, quais

os tratamentos utilizados no estudo, quais hipóteses foram elaboradas, além de qual desenho experimental foi escolhido. Além disso, as ameaças à validade do estudo são também detalhadas neste capítulo.

- Capítulo 5 - Resultados: Detalha o processo executado no experimento. Além dos métodos estatísticos utilizados com os dados obtidos através da condução do experimento proposto, este capítulo aborda os resultados do experimento para as métricas definidas.
- Capítulo 6 - Conclusão: Neste capítulo, são mostradas as conclusões do experimento realizado, suas contribuições e por fim, são discutidos os trabalhos futuros que poderão ser feitos a partir dos resultados alcançados nesta pesquisa.

2 REFERENCIAL TEÓRICO

Este capítulo é dividido da seguinte maneira: Seção 2.1, que aborda conceitos a respeito da modelagem de dados; Seção 2.2, onde são tratados conceitos sob a ótica de sistemas NoSQL, e, por fim, na Seção 2.3, são elencadas as características necessárias para a compreensão de estudos experimentais.

2.1 MODELAGEM DE DADOS

De acordo com Moody e Shanks (1994), a modelagem de dados é uma das fases mais críticas no processo de desenvolvimento de sistemas, pois, a partir dela são construídos modelos representando a estrutura dos dados e suas características, como restrições e relacionamentos (CORONEL; MORRIS, 2016). A criação de modelos de dados tem por objetivo facilitar a compreensão do domínio abordado, seja do ponto de vista de usuários finais ou desenvolvedores (CORONEL; MORRIS, 2016), (KAUR; RANI, 2013). Além disso, Moody e Shanks (1994) enfatizam que, mesmo a modelagem de dados dispende de uma proporção pequena no projeto de desenvolvimento de sistemas de bancos de dados, seu impacto final é maior que o produzido por outras fases, uma vez que o uso de um modelo preexistente torna mais simples o processo de construção de um projeto de banco de dados (CORONEL; MORRIS, 2016).

O uso de modelo de dados deve prover de bases formais fáceis (notacional e semântica) e linguagem natural (BRODIE; MYLOPOULOS; SCHMIDT, 2012), pois, dessa forma, usuários finais podem revisar e verificar qualquer problema modelado e entendê-lo, mesmo sem conhecimento prévio da área de computação. O principal ponto para utilizar modelos de dados é tornar mais simples a interação entre programadores, desenvolvedores e usuários finais (KAUR; RANI, 2013). Entretanto, nem todos os modelos são construídos com o mesmo propósito (KAUR; RANI, 2013), por exemplo, modelos conceituais são desenvolvidos com a finalidade de representar a modelagem de um banco de dados em alto nível, enquanto modelos de implementação ou lógicos são criados com o intuito de representar a forma de gerenciar dados em um sistema específico (CORONEL; MORRIS, 2016).

2.1.1 Níveis na modelagem de dados

De acordo com Elmasri e Navathe (2010), a modelagem de dados no desenvolvimento bancos de dados é dividida em três níveis: conceitual, lógico e físico. Cada um deles é descrito a seguir:

Modelagem conceitual: é a primeira descrição formal da aplicação de banco de dados. Segundo Engels et al. (1992) é a mais importante, por conter uma descrição completa do domínio (FOSTER; GODBOLE, 2016), sendo as demais fases meramente passos de mape-

amento da conceitual (ENGELS et al., 1992). Na fase conceitual, a informação deve possuir um certo grau de abstração para alcançar o máximo de independência de problemas de implementação, e ainda assim, ser compreensiva e representar todo o desenho do banco de dados (FOSTER; GODBOLE, 2016).

A modelagem conceitual de aplicações de um banco de dados consiste em descrever adequadamente partes relevantes do domínio. O papel do desenho conceitual no processo de banco de dados é mostrar todo o processo em alto nível, e que esse seja altamente expressivo. Contudo, os dados representados a partir do modelo conceitual devem ser independentes de *software* e *hardware* (CORONEL; MORRIS, 2016), isto é, devem ser independentes do SGBD (Sistema de Gerenciamento de Banco de Dados) que irá implementar os dados representados pelo modelo conceitual e que o modelo conceitual não dependa do *hardware* usado na sua implementação (FOSTER; GODBOLE, 2016). Do ponto de vista do usuário final, a importância do modelo conceitual é que esse permite a visualização mais próxima de qual dado será armazenado e gerenciado, e de como os dados se relacionam, tornando assim, mais fácil o entendimento do ambiente de dados para o cliente (FOSTER; GODBOLE, 2016), (CORONEL; MORRIS, 2016). Conforme Engels et al. (1992), o modelo conceitual tem várias funções, incluindo: servir de base para auditoria de sistemas, auxiliar na conversão de aplicações de BD, auxiliar na comunicação entre projetista do BD e o usuário e ajudar na avaliação experimental de BD.

Modelagem lógica: esse nível consiste no mapeamento das especificações do nível conceitual para um modelo mais próximo de um paradigma de computação (THALHEIM, 2013). Segundo Elmasri e Navathe (2010), o modelo obtido do mapeamento conceitual para o lógico possui algumas restrições que se aplicam ao SGBD, assim, nessa fase do mapeamento é necessário que se saiba para qual SGBD está sendo realizado o mapeamento (CORONEL; MORRIS, 2016). Isto é, se o mapeamento está sendo realizado para um sistema relacional, NoSQL orientado a documentos, etc.

Modelagem física: é a definição do modelo físico de banco de dados (LONGLEY et al., 2005). Esse por sua vez, é mapeado do modelo lógico. O modelo físico contém conceitos que descrevem os detalhes de como os dados estão armazenados no computador (ELMASRI; NAVATHE, 2010), e esse é dependente do SGBD (LONGLEY et al., 2005). De acordo Elmasri e Navathe (2010), os modelos físicos não são significativos para usuários finais, mas para especialistas em computadores, por possuir uma descrição não muito diferente de como os dados estão organizados dentro do computador.

2.1.2 Análise de modelos conceituais

Quando um modelo conceitual está sendo avaliado, essa avaliação pode ser feita por diferentes óticas: a do usuário e a do desenvolvedor (CHERFI; AKOKA; COMYN-WATTIAU, 2002), (HUSSAIN, 2014). Quando um modelo conceitual é analisado pela ótica do usuário, normalmente se analisa o entendimento dele sobre determinado domínio expresso no

modelo (GENERO; POELS; PIATTINI, 2002). Pois, quanto mais simples forem as regras e conceitos desse modelo, melhor será do ponto de vista do usuário, entender o que está sendo representado (CHERFI; AKOKA; COMYN-WATTIAU, 2002).

A análise feita pela ótica do desenvolvedor é um pouco diferente, pois, além do entendimento do desenvolvedor sobre o modelo conceitual estudado (GENERO; POELS; PIATTINI, 2002), podem ser analisadas a expressividade do modelo para representar diferentes domínios (HUSSAIN et al., 2006) e a usabilidade do modelo (GENERO; POELS; PIATTINI, 2002), (HUSSAIN et al., 2006). Além das características citadas, também pode se observar se os esquemas construídos pelo modelo conseguem expressar de forma completa e correta o domínio (PATIG, 2004). Nesse contexto, as análises feitas sob a visão dos desenvolvedores podem ajudar a tomar melhores decisões durante a fase de modelagem do banco de dados (GENERO et al., 2001), (GENERO; POELS; PIATTINI, 2002).

Na Seção 3.2 desta dissertação são discutidos estudos que realizaram avaliações de modelos conceituais, sendo abordada a forma como essas avaliações foram executadas e como os dados foram obtidos e analisados.

2.2 SISTEMAS DE BD NOSQL

O termo NoSQL surgiu em 1998 em um grupo na *Bay Area*, que realizava encontros periódicos onde eram discutidas questões a respeito dos modelos relacionais e problemas em escalar bancos de dados *open source*, isto é, de código aberto (MCCREARY; KELLY, 2014). Contudo, em 2009, esse termo passou a ser utilizado para se referir a qualquer sistema de banco de dados que armazenasse seus dados sem seguir os padrões dos modelos tradicionais, ou seja, não fazendo uso de SQL como linguagem de consulta e não utilizando diferentes formas de expressar relacionamentos entre os dados (HARRISON, 2015), (VAISH, 2013).

Os sistemas de BD NoSQL normalmente não realizam operações de junções e possuem suporte BASE (*Basically Available, Soft-state, Eventual consistency*), ao invés do ACID (Atomicidade, Consistência, Isolamento, Durabilidade), utilizado pelos sistemas de BD relacionais (GANIEE; BHARGAVA, 2015). O uso do BASE, ao invés do ACID pelos sistemas NoSQL, se deve ao fato de que esses sistemas foram desenvolvidos para se adequar ao novo contexto trazido pela *Web 2.0*, se concentrando em solucionar problemas não funcionais, específicos de aplicações, como: disponibilidade, escalabilidade, agilidade, entre outros (SHIN; HWANG; JUNG, 2017). Dessa forma, esses sistemas utilizam um conjunto de conceitos que permitem o processamento rápido e eficiente de uma quantidade alta de dados (HASEEB; PATTUN, 2017), possibilitando as características citadas acima e impossibilitando o uso do ACID (GANIEE; BHARGAVA, 2015). Por conseguir lidar com dados não estruturados, soluções NoSQL são utilizadas normalmente em aplicações de comércio eletrônico, serviços *Web*, computação móvel, serviços baseados em localização e aplicações de mídias sociais (HASEEB; PATTUN, 2017), (LEAVITT, 2010).

2.2.1 Definições básicas

Segundo Moniruzzaman e Hossain (2013), os sistemas de BD NoSQL são definidos como um grupo de banco de dados distribuídos que não seguem os princípios dos sistemas relacionais. Esses sistemas são altamente distribuídos, uma vez em que focam em características como escalabilidade horizontal e alta disponibilidade (VAISH, 2013). Por essa razão, não são capazes de prover das propriedades ACID, pois, tais características podem acarretar em mau desempenho no processamento de transações ou na execução de consultas distribuídas (TOTH, 2011). O CouchDB e Neo4j são exceções, pois são compatíveis com as propriedades ACID (VAISH, 2013). Por essa razão, grande parte dos sistemas NoSQL utilizam outra abordagem baseada no teorema CAP (Consistência, Disponibilidade e Tolerância à partição) apresentada por Eric Brewer em 2000 (BREWER, 2012). Nesse teorema, é dito que qualquer sistema NoSQL distribuído só pode possuir duas das três propriedades simultaneamente (MCCREARY; KELLY, 2014). Suas idiossincrasias são as seguintes:

- Consistência – descrita quando um dado é modificado e salvo com sucesso na base de dados, garantindo que o dado deverá estar atualizado em todas as partições que o possuírem (FOWLER, 2015). Isto é, quando um item de dado for modificado com sucesso, novas operações de leitura sobre esse item deverão retornar resultados consistentes para todas as partições que o contenham (MCCREARY; KELLY, 2014);
- Disponibilidade – assegura que o sistema fique a maior parte do tempo aberto e acessível para solicitações (LAKE; CROWTHER, 2013), ou seja, clientes devem sempre encontrar pelo menos uma cópia dos dados solicitados (TOTH, 2011);
- Tolerância à partição – é a habilidade que um sistema tem de continuar respondendo a solicitações de clientes mesmo que tenha ocorrido uma falha de comunicação entre a base de dados e as suas partições (MCCREARY; KELLY, 2014). Isto é, caso ocorra qualquer tipo de falha num sistema tolerante à partição, o cliente continuará a realizar operações no mesmo, como se o sistema estivesse efetivo.

Para que se tornasse possível trabalhar em ambientes distribuídos, foi proposto enfraquecer a consistência de dados em prol das outras características. A partir dessa ideia surgiu o conceito BASE, que foi adotado pela maioria dos sistemas NoSQL (TANG; FAN, 2016), como dito anteriormente. As propriedades BASE podem ser sumarizadas da seguinte maneira: uma aplicação funciona basicamente o tempo todo (*basically available*), não precisa ser consistente o tempo todo (*soft-state*), mas em algum momento ela será consistente (*eventual consistency*) (VAISH, 2013). A seguir serão abordadas algumas características que os sistemas NoSQL possuem.

2.2.2 Características Principais

Dentre as características dos sistemas de BD NoSQL pode-se citar: a ausência de um esquema fixo, o não uso da linguagem SQL, a não realização de junções em consultas, entre outras que diferenciam os sistemas NoSQL dos SGBD tradicionais. Isso torna os sistemas de BD NoSQL ideais para lidar com ambientes que necessitem de alta disponibilidade e escalabilidade, como dito em seções anteriores. A seguir, são apresentadas algumas características desses sistemas:

- Os sistemas NoSQL utilizam diferentes formas de representar relacionamentos entre objetos, não fazendo uso de chaves estrangeiras (utilizadas em modelos de dados relacionais). Isto permite a extração de dados usando interfaces simples, eliminando a necessidade de consultas complexas (VAISH, 2013), evitando a necessidade de junções para obter-se determinada informação de um dado (MCCREARY; KELLY, 2014);
- Esquemas menos rígidos: modelos de dados de sistemas NoSQL oferecem uma representação de dados mais flexível (VAISH, 2013), permitindo uma maior liberdade no armazenamento dos dados (FOWLER, 2015), diferentemente do modelo relacional que possui uma estrutura mais rígida. Para organizar dados de sistemas NoSQL não há a necessidade de se pensar muito à frente do que determinada aplicação armazenará, podendo-se adicionar novos campos quando necessários (VAISH, 2013), tornando mais simples o armazenamento de dados (MCCREARY; KELLY, 2014);
- Altamente distribuído: o armazenamento de dados nos sistemas NoSQL é feito em múltiplos processadores, garantindo um alto nível de disponibilidade, uma vez que os dados são armazenados em múltiplos nós da rede (FOWLER, 2015).
- Escalabilidade: por não fazer uso das propriedades ACID, soluções NoSQL conseguem fornecer uma maior escalabilidade, quando comparadas com sistemas tradicionais de BD (VAISH, 2013), pois são capazes de compartilhar sua carga de trabalho entre os nós da rede (KASHYAP et al., 2013). Isso possibilita um aumento de desempenho, além de adição de novos nós quando necessário (LEITE, 2014).

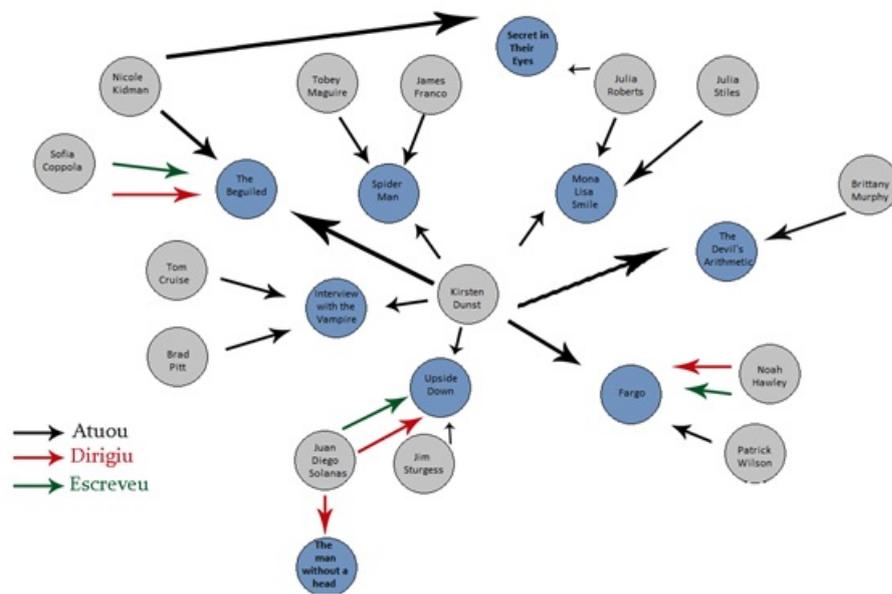
Além das características citadas acima, os modelos de dados NoSQL podem ser divididos em quatro grupos (MCCREARY; KELLY, 2014): orientado a grafos, orientado a documentos, chave-valor e orientado à colunas. Algumas características a respeito de cada um desses grupos são apresentadas a seguir.

2.2.3 Modelos de dados de sistemas NoSQL

Orientado a Grafos: esse modelo possui três componentes básicos: nós, relacionamentos e propriedades, as quais estão presentes nos nós e relacionamentos (HARRISON, 2015). Os nós representam objetos distintos, enquanto os relacionamentos são definidos pelas

conexões entre os objetos (SCHILDGEN; LOTTERMANN; DESSLOCH, 2016), onde, um nó pode conter mais de um *link*, permitindo assim, múltiplos relacionamentos (VAISH, 2013). Na Figura 1, pode-se ver no ponto central, um nó identificado por Kirsten Dunst e *links* que mostram os filmes em que ela atuou. Por modelos orientados a grafos permitirem que um mesmo nó possua n relacionamentos ou *links*, é possível executar diversos tipos de consultas sobre esquemas de dados orientado a grafos modelos. Por exemplo, a seguinte consulta pode ser executada sobre o esquema ilustrado na Figura 1: Quais atores que trabalharam com a Kirsten Dunst, já trabalharam juntos em outro filme?

Figura 1 – Exemplo de um Esquema de Dados



Fonte: O autor (2019)

Devido à possibilidade de se utilizar propriedades de grafos como medidas de centralidade (*pangerank*), busca por nós vizinhos e algoritmos de menor caminho, as consultas nesse tipo de esquema de dados podem ser mais rápidas. Por ser uma abordagem que se baseia em guardar informações úteis de relacionamentos entre dois ou mais nós, modelos orientados a grafos podem retornar bons resultados quando aplicados em redes sociais, sistemas de recomendação, topologias de rede entre sistemas conectados, entre outras aplicações (VAISH, 2013), (TOTH, 2011).

Orientado a Documentos: bancos de dados que utilizam esse modelo permitem a inserção, manipulação e recuperação de dados semiestruturados (VAISH, 2013). Normalmente, o armazenamento dos dados é feito em coleções de documentos semelhantes aos formatos XML, JSON (HARRISON, 2015). Um documento normalmente é um objeto com um identificador único e um conjunto de campos, onde esses campos podem assumir valores do tipo *string*, lista ou documentos aninhados (LIMA; MELLO, 2016).

Figura 2 – Exemplo de um Esquema de Dados

```

Cliente{
  "id":10,
  "nome": "Júlio",
  "endereco": "Alfeneiros, 81"
}

Cliente{
  "id":11,
  "nome":"Fernando",
  "endereco": "Alfeneiros, 90",
  "localização": "80' 24", 36' 24\"",
  "n. cartão": "1234 5431 1992 1993"
}

```

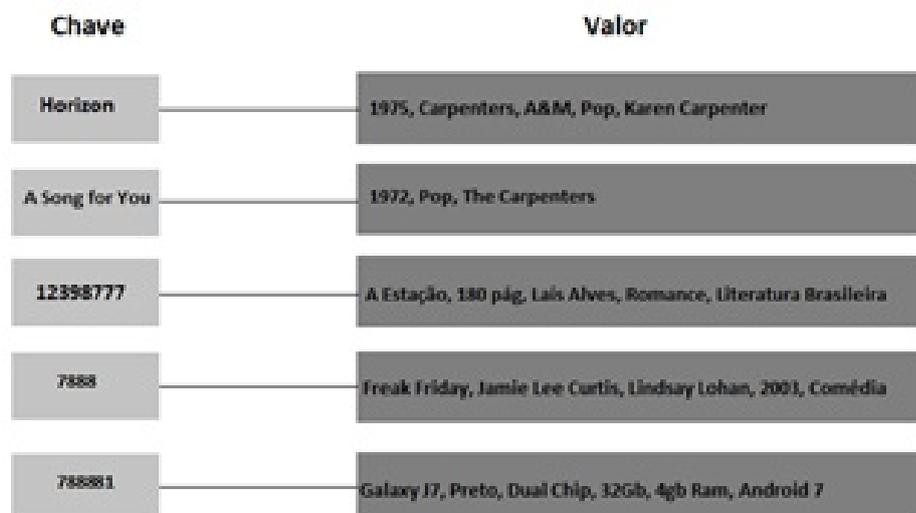
Fonte: O autor (2019)

A Figura 2 apresenta um esquema de dados orientados a documentos, no qual pode-se notar, que documentos do mesmo tipo, nesse caso Cliente, podem possuir atributos distintos, dependendo da necessidade da aplicação. Segundo Harrison (2015), sistemas orientados a documentos podem ser uma opção de meio termo, pois, permitem alguma descrição da formatação sem possuir a rigidez imposta pelos modelos relacionais e sem a total falta de esquema de outros modelos NoSQL (HARRISON, 2015).

Geralmente, modelos orientados a documentos dão suporte secundário a índices e à criação de múltiplos tipos de documentos (objetos) por base de dados (CATTELL, 2011).

Chave-valor: é um sistema de BD de fácil implementação, apresentado com uma *string* simples (a chave) retornando uma grande quantidade de dados arbitrários (o valor) (VAISH, 2013), como pode ser visto na Figura 3.

Figura 3 – Exemplo de um Esquema de Dados do Tipo Chave-valor



Fonte: O autor (2019)

Os sistemas NoSQL do tipo chave-valor não possuem linguagens de consulta, e os dados são armazenados em pares. A combinação entre eles (chave e valor), denota que a chave é associada ao valor, permitindo a visualização de dados como uma tabela *hash* (THURM, 2012). Na maioria dos casos, a interface dos bancos de dados que utilizam o modelo chave-valor seguem dois métodos principais (inserir e recuperar): *Put* (chave, valor), método que insere o registro; e *Get* (chave), método que recupera esse registro (HARRISON, 2015). Como as buscas são apenas realizadas pela chave, ela permite acesso ao dado de forma mais rápida que em outros modelos NoSQL (THURM, 2012). Os sistemas chave-valor não permitem valores atômicos, nem listas aninhadas arbitrárias, conjuntos e mapas (SCHILDGEN; LOTTERMANN; DESSLOCH, 2016). Contudo, tais sistemas proveem um mecanismo de persistência, adicionando funcionalidades como replicação, versionamento, transações e seleção de dados (CATTELL, 2011).

Orientado à Colunas: os dados nesse modelo são formados por uma tríplice (linha, coluna e *timestamp*). Nesta tríplice, linhas e colunas são identificadas por chaves e o *timestamp* é utilizado para diferenciar múltiplas versões de um mesmo dado (LIMA; MELLO, 2016). O modelo orientado à colunas possui algumas características em comum com o modelo objeto-relacional, como o fato de organizar dados em coleções de colunas (MCCREARY; KELLY, 2014). No entanto, as colunas são independentes e armazenam arquivos sem um esquema fixo, ou seja, diferentes linhas podem possuir colunas distintas. Quando há um mesmo tipo de dado usado por um grupo de colunas, essas são agrupadas em uma família de colunas (LIMA; MELLO, 2016), como pode ser visto na Figura 4. Esse tipo de sistema NoSQL permite o armazenamento eficiente em termos de grande volumes de dados (VAISH, 2013) e possibilita escalar uma grande quantidade de dados de forma eficiente (MCCREARY; KELLY, 2014).

Figura 4 – Exemplo de um Esquema de Dados Orientado à Colunas

Nova York		
Manhatan		Queens
Restaurantes	Parques	Museus
404605 73555{nome: Massa, avaliação: 5}	40783091 71234567{nome: Central Park}	400769 737012{nome: MoMA, diretor: Klaus Bresdenbech}

Fonte: Adaptado de Saravanan. Disponível em <https://techieetweak.wordpress.com/2016/01/09/no-sql/>

2.3 PLANEJAMENTO DE ESTUDOS EXPERIMENTAIS

Segundo Wohlin et al. (2012), quando se quer explicar o motivo da ocorrência de determinada situação, a melhor forma de estudar essa situação é por meio de um experimento. Contudo, um experimento deve ser cautelosamente planejado para que seja possível executá-lo da melhor forma (HANNAY; SJOBERG; DYBA, 2007). Assim, deve-se planejar como a situação será observada, que tratamentos serão aplicados, quais métricas serão usadas, como as saídas do experimento serão analisadas e se essas saídas têm algum impacto na situação estudada (SJØBERG et al., 2008). De acordo com (SJØBERG et al., 2008), experimentos são apropriados para investigar diferentes aspectos a respeito de determinado fenômeno. De acordo com Hannay, Sjoberg e Dyba (2007) eles servem para:

- Explorar relacionamentos e descobrir como esses influenciam o contexto que está sendo observado;
- Confirmar uma teoria;
- Avaliar a correteza de determinado modelo ou técnica. Para isso, são realizados testes, a fim de testar se esses têm a resposta esperada.

Nesta seção, serão abordadas algumas características acerca do desenho de um estudo experimental.

2.3.1 Características de um experimento

A principal característica de um experimento é ser capaz de investigar quais situações são verdadeiras, além de fazer uso de certos padrões, métodos e ferramentas que podem ser utilizados para a sua execução (WOHLIN et al., 2012). Um experimento pode ser dividido de duas formas: orientado para seres humanos e para tecnologia (WOHLIN et al., 2012). A diferença entre ambos é que, no orientado a tecnologia, um mesmo teste pode ser aplicado várias vezes para diferentes objetos sem ter os resultados viesados. Enquanto nos estudos envolvendo humanos, um mesmo teste não pode ser aplicado uma segunda vez em outro objeto, pois poderá ter os seus resultados viesados (WOHLIN et al., 2012).

Segundo Juristo e Moreno (2013), a execução de um experimento envolve quatro atividades:

1. Definição dos objetivos da experimentação;
2. O desenho do experimento;
3. A execução do experimento;
4. Análise dos resultados/dados coletados no experimento.

As quatro atividades definidas por Juristo e Moreno (2013), serão explanadas ao decorrer desta seção.

2.3.2 Definição dos objetivos da experimentação

A definição do que o experimento quer observar se constrói a partir de uma teoria que se tem a respeito de determinado assunto e de como uma prática/ou ferramenta influencia o contexto que se quer observar (WOHLIN et al., 2012). Conforme Juristo e Moreno (2013) explicam, uma hipótese é formada a partir de três raciocínios:

1. Dedução: prova-se que algo deve ser;
2. Indução: mostra que o objeto de estudo é realmente operacional;
3. Abdução: sugere que algo deve ser estudado.

A hipótese elaborada tem que estar de acordo com os objetivos do experimento, pois assim, torna-se possível a obtenção de evidências que auxiliam ou não a hipótese levantada (MORETTIN; BUSSAB, 2017). Uma das formas de obter evidências sobre determinado assunto é por meio da aplicação de questionários, pois eles proveem de perguntas sobre a área observada e seus resultados podem corroborar com a hipótese levantada.

2.3.2.1 Questionários

Antes da elaboração de um questionário é importante saber qual metodologia está sendo usada no estudo, se é qualitativa, quantitativa ou híbrida, para assim, construir um questionário que se adeque mais ao desenho do estudo, que pode ser: quasi-experimento, *survey*, correlacional ou qualitativo (CONNOLLY et al., 2012).

Normalmente, antes de se aplicar questionários sobre a abordagem ou técnica analisada, grande parte dos estudos empíricos aplicam questionários de perfil (JURISTO; MORENO, 2013), que é uma técnica comum, pois, a partir dela é possível construir um *background* dos participantes do estudo, como: idade, gênero, experiências anteriores, área de interesse (PAPASTERGIOU, 2009), correlacionando assim, essas informações com o perfil dos participantes do estudo (PAPASTERGIOU, 2009).

Antes de se criar um questionário, é necessário ter certeza do que se deseja observar e qual a melhor forma de formular questões para obter respostas que ajudem a entender determinado fenômeno. As perguntas elaboradas devem ser as mais objetivas possíveis, para que a pessoa que irá responder o questionário, não fique em dúvida sobre o que está sendo perguntado (SARIS; GALLHOFER, 2014). Os questionários podem possuir questões abertas ou fechadas (CONNOLLY et al., 2012).

Questões abertas são mais comuns em estudos qualitativos, ou seja, que não se busca quantificar os dados obtidos, mas apenas entender os relacionamentos entre os dados

(WOHLIN et al., 2012). Contudo, de acordo com Saris e Gallhofer (2014), questões abertas podem ser problemáticas, uma vez que deixam a pessoa livre para responder o questionário, o que possibilita a fuga do tema ou o não fornecimento de dados relevantes para o que está sendo observado.

Enquanto questões abertas são mais comuns em estudos qualitativos em estudos quantitativos são mais comuns questões fechadas, pois essas são mais fáceis de quantificar os resultados (JURISTO; MORENO, 2013). Há diversas formas de se elaborar questionários com respostas fechadas, sendo que as mais comuns são:

- Bateria de respostas: nesse tipo de questionário é feita uma introdução do assunto. A partir da introdução, são feitas afirmações (SARIS; GALLHOFER, 2014), onde a pessoa que está respondendo marcará se concorda ou não como essas afirmações, assumindo valores em escala (variando de concorda completamente para discorda completamente, ou indicando apenas sim ou não) em suas respostas.
- Questões sim e não: segundo Saris e Gallhofer (2014) questionários desenvolvidos dessa forma são mais simples de quantificar, porém, os resultados obtidos podem não ser os mais próximos da realidade, uma vez que determinada questão pode estar em um meio termo.
- Questões em escala: esse formato de questionário permite uma compreensão maior de determinada observação, pois abrange mais opções que apenas questionários do tipo 'sim e não' (SARIS; GALLHOFER, 2014). A escala que será utilizada na construção do questionário pode ser definida pelo seu criador, podendo ser numérica ou de intensidade (JURISTO; MORENO, 2013).

Os questionários aplicados nesta dissertação são discutidos no Capítulo 4. Antes de continuar com a explicação das quatro atividades definidas por Juristo e Moreno (2013), apresentam-se alguns jargões utilizados em estudos experimentais:

- Unidade experimental – são os objetos analisados durante a execução do experimento, esses podem ser chamados de unidades experimentais ou objetos experimentais (JURISTO; MORENO, 2013). Usualmente, experimentos realizados em engenharia de software envolvem a análise de um projeto de desenvolvimento de software, onde sua unidade experimental é caracterizada pela análise feita em todo software produzido, ou em uma parte específica do processo de desenvolvimento (WOHLIN et al., 2012);
- Sujeito experimental – é a pessoa que aplica a técnica ou método na unidade experimental, que pode ser chamada de sujeito ou participante (JURISTO; MORENO, 2013);

- Tratamento – termo empregado para designar as alternativas de um fator em um projeto experimental (JURISTO; MORENO, 2013);
- Parâmetro – qualquer característica (quantitativa ou qualitativa) do projeto de software invariável durante a execução do estudo chama-se parâmetro (JURISTO; MORENO, 2013). Dessa forma, o parâmetro é uma característica que não influencia ou que não se quer observar a sua influência nos resultados do experimento;
- Variável de resposta – é a variável que se quer estudar sua influência/efeito na variável independente. Pode ser chamada de variável dependente, e cujo o seu resultado deve ser quantitativo (JURISTO; MORENO, 2013), (WOHLIN et al., 2012). Normalmente, há apenas uma variável dependente em um experimento, as demais variáveis utilizadas no processo são chamadas variáveis independentes (WOHLIN et al., 2012);
- Replicação – envolve a repetição em condições similares ao estudo que está sendo replicado (WOHLIN et al., 2012). Juristo e Moreno (2013) dividem a replicação de duas formas, a externa e interna. A replicação externa é realizada por outros pesquisadores com outras configurações e diferentes amostras. Essa replicação tem o intuito de obter resultados próximos ao feito no experimento que está sendo replicado (JUDD; SMITH; KIDDER, 1991). Já a replicação interna é a repetição de toda ou alguma parte da unidade experimental realizada no próprio experimento (JURISTO; MORENO, 2013).

2.3.3 Execução do experimento

Esta fase deve ocorrer de acordo com o planejamento elaborado, respeitando o desenho experimental escolhido (WOHLIN et al., 2012). Pois, uma vez que haja mudanças na execução do experimento, elas podem gerar resultados inadequados, impactando nas conclusões do experimento (JURISTO; MORENO, 2013). Ademais, deve-se seguir o planejamento elaborado para que outros pesquisadores possam executar o experimento nas mesmas condições e observar se os resultados são os mesmos. Ao final da execução do experimento, a próxima fase é analisar os dados obtidos.

2.3.4 Análise dos dados

A análise dos resultados busca encontrar relacionamentos entre os dados coletados, que podem ser descritivos, casuais e correlacionais (JURISTO; MORENO, 2013). A análise estatística é utilizada para observar os dados obtidos pelo experimento (JURISTO; MORENO, 2013), e essa é capaz de desenhar conclusões válidas a respeito dos dados experimentais (WOHLIN et al., 2012). A interpretação quantitativa desses dados é feita a partir de três passos:

1. Os dados são caracterizados usando estatística, ou seja, é analisada a tendência principal, ou a dispersão entre outras(SJØBERG et al., 2008);
2. Dados falsos ou anormais são excluídos, reduzindo assim, o conjunto de dados que são considerados (SJØBERG et al., 2008);
3. Os dados são aplicados à hipótese do experimento, sendo avaliados estatisticamente, mostrando o seu nível de significância (WOHLIN et al., 2012).

Por examinar ou analisar os dados de forma estatística, os pesquisadores chegarão a conclusões mais claras sobre os relacionamentos entre os fatores, a influência dos fatores nas variáveis de resposta, além de alternativas que aperfeiçoem os resultados dessas variáveis de resposta (JURISTO; MORENO, 2013). O próximo capítulo apresenta os trabalhos correlatos do estudo descrito nesta dissertação.

3 TRABALHOS CORRELATOS

3.1 MODELOS CONCEITUAIS PARA SISTEMAS NOSQL

Esta seção apresenta propostas de modelos de dados para sistemas NoSQL.

3.1.1 Modelos de dados baseados no modelo ER (Entidade Relacionamento)

Em Chebotko, Kashlev e Lu (2015) é apresentada uma nova metodologia de modelagem de dados destinada ao SGBD Apache Cassandra, que se baseia na conversão do modelo ER para modelos lógicos e físicos. Além disso, foi desenvolvida a ferramenta KDM (Kashlev Data Modeler), que possui um conjunto de princípios, regras e padrões de mapeamento para a conversão do modelo ER em modelos lógicos e físicos, automatizando a construção de projetos de banco de dados para o Cassandra. A conversão do ER para o modelo lógico é definida em tempo de execução, onde são aplicadas cinco regras de mapeamento para a construção de esquemas lógicos:

1. Entidades e relacionamentos – são mapeados para linhas de tabelas.
2. Igualdade entre atributos – uma consulta predicada é mapeada para a chave primária da coluna.
3. Desigualdade nas pesquisas – na definição da chave primária de uma coluna que participa de outra, deve-se acrescentar o atributo chave nessa coluna.
4. Ordenação de atributos – os atributos especificados na consulta são mapeados para a coluna chave agrupando-os com o seu ascendente, ou descendente, que é descrito na consulta.
5. Atributos chave – o atributo chave de cada entidade deve ser mapeado para a chave primária da coluna que representa essa entidade.

Para exemplificar o funcionamento da ferramenta proposta em Chebotko, Kashlev e Lu (2015), para mapeamento de esquemas lógicos e físicos, foi modelado o domínio de uma biblioteca digital, usando a ferramenta proposta. A principal contribuição de Chebotko, Kashlev e Lu (2015) foi a definição de diagramas visuais para modelos de dados lógicos e físicos para o Cassandra, e assim, auxiliando na visualização de como deve ser feita a construção de esquemas de dados para serem implementados no Cassandra.

Em Hamouda e Zainol (2017), os autores propuseram o DODS (Esquema de Dados Orientado a Documentos), que realiza o mapeamento de esquemas conceituais ER para esquemas NoSQL orientados a documentos. A construção do DODS é feita em dois estágios:

1. Mapeamento do modelo ER para especificações DODS, subdividindo-se em duas partes:
 - a) Componentes DODS denotam a descrição de cada componente criado para o mapeamento. Dessa forma, uma coleção representa uma tabela; um documento representa um arquivo; pares chave-valor representam o atributo e o tipo do dado; múltiplos documentos representam um array; o relacionamento embutido, é aquele no qual umq ou mais documento/s estão embutidos em documento, assim, esse tipo de relacionamento é utilizado para representar relacionamentos entre documentos; e o relacionamento por referência representa o relacionamento entre coleções, onde há uma referência de um outro documento em documento específico.
 - b) Mapeamento do modelo ER para o DODS, os recursos do diagrama ER são mapeados para o DODS, incluindo novos construtores e relacionamentos.

O mapeamento ocorre da seguinte maneira:

- Uma entidade é mapeada para uma coleção, a qual pode também representar uma entidade fraca, onde a coleção fraca estará contida na coleção forte.
- Os atributos de cada coleção são delimitados por chaves, onde dentro das chaves há um par com o nome do atributo e seu respectivo valor.
- Quando um atributo é chave, esse deve ter o nome grifado, diferenciando-o dos demais atributos. Caso haja um atributo chave de outra coleção, o nome desse é tracejado, identificando-o como atributo chave, mas de outra coleção.
- Atributos compostos são representados da seguinte forma: o nome do atributo seguido por colchetes, onde dentro desses colchetes são postos os valores que o atributo possui, e os valores são separados por vírgula. O colchete é utilizado para indicar que um único atributo pode conter diversos valores, ou seja, dar a ideia de um *array*.
- A cardinalidade dos relacionamentos é a mesma utilizada no modelo ER (1:1, 1: N e N: M) com a adição de um novo relacionamento: o unário, que é representando pelo '@' após o nome do relacionamento.

A principal contribuição de Hamouda e Zainol (2017) foi possibilitar, através das especificações do DODS, o mapeamento do modelo ER para esquemas lógicos voltados para sistemas NoSQL orientados a documentos.

Em Lima e Mello (2016), o mapeamento de esquemas conceituais EER (Entidade Relacionamento Estendido), para esquemas lógicos de aplicações de sistemas orientados a documentos, é feito em dois passos:

1. Conversão de tipos de hierarquia, composta de três regras:

Regra 1: é gerado um bloco que representa a superclasse e seus atributos;

Regra 2: cria-se blocos de documentos para subclasse, onde os atributos da superclasse são reproduzidos em cada bloco da subclasse;

Regra 3: a superclasse e subclasse são explicitamente representadas por blocos, e a hierarquia dos relacionamentos entre elas se dá através dos blocos das superclasses para a subclasses.

2. Na conversão de tipos de relacionamentos, cada cardinalidade de relacionamento possui uma regra.

Regra 4: os relacionamentos 1:1 são expressos através de um bloco;

Regra 5: nos relacionamentos 1: N é gerado um bloco para cada entidade do relacionamento, onde um bloco é aninhado ao outro bloco, e os atributos do relacionamento são postos no bloco aninhado;

Regra 6: nos relacionamentos N: M, blocos independentes são gerados para os relacionamentos entre as entidades. Além disso, cada bloco possui uma referência para blocos que estão relacionados a ele.

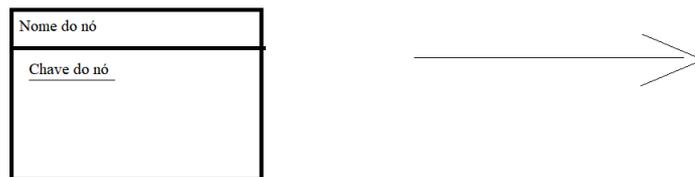
O modelo lógico construído a partir das regras definidas por Lima e Mello (2016) é independente do SGBD utilizado. Por se basear em agregados, ele pode ser usado por qualquer SGBD NoSQL orientado a documentos. A avaliação do modelo ocorreu a partir da construção de um estudo de caso para uma aplicação de comércio eletrônico.

Em Roy-Hubara et al. (2017), é proposto um mapeamento do modelo ER para o GDBS (Esquemas de Banco de Dados para Grafos). Esse mapeamento se dá a partir das seguintes etapas:

1. Mapeamento de entidades em nós: cada entidade é mapeada para um nó. Os atributos dessa entidade se tornam propriedades do nó. Um nó no GDBS é representado por um retângulo dividido em duas partes. A parte superior contém o nome desse nó, já a parte inferior do retângulo contém as propriedades desse nó. A chave do nó é sublinhada para diferenciá-la das demais propriedades, como pode ser visto na Figura 5.
2. Mapeamento de relacionamentos em arestas: cada relacionamento entre as entidades é mapeado para uma aresta, que conecta os respectivos nós. A aresta possui uma direção e um nome (Figura 5), que não precisa ser único, pois o que vai identificar a aresta é sua direção.
3. Cardinalidade: bancos de dados orientados a grafos não possuem restrições no número de nós que participam de uma aresta. Contudo, os autores decidiram restringir

a cardinalidade para alguns casos, como por exemplo: em um sistema de classificação, o mesmo usuário só poderá avaliar um mesmo objeto uma única vez, assim, caso o usuário avalie um determinado filme, esse só poderá ser avaliado uma única vez por esse usuário.

Figura 5 – Notação do GDBS



Fonte: O autor (2019)

Roy-Hubara et al. (2017) exemplificaram como se dá o processo de mapeamento do modelo ER para o GDBS, utilizando um estudo de caso no domínio de uma aplicação de classificação de filmes. O estudo focou na construção desse mapeamento, todavia, faltou ao estudo, analisar os impactos do uso do GDBS no processo de construção de esquemas para sistemas orientados a grafos.

Em Vera et al. (2015), foram apresentadas técnicas de modelagem de dados em diagramas ER, para que esses pudessem ser usados em aplicações de bancos de dados orientados a documentos. Os autores observaram que há duas formas de representar relacionamentos entre dados orientados a documentos: por meio de referências ou por meio de documentos embutidos. Assim, Vera et al. (2015) definiram dois tipos de relacionamentos, como segue:

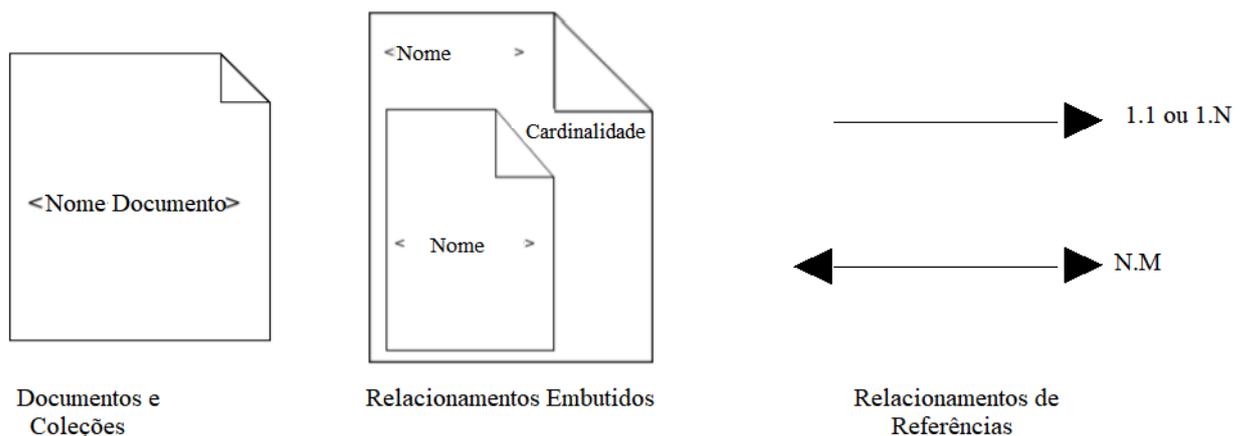
1. Relacionamento de referências: caracterizado pela inclusão de *links* de um documento em outro documento, tornando possível a identificação do relacionamento através dos *links*.
2. Relacionamento de documentos embutidos: definido pela inclusão de dados em um único documento, onde cada dado deve ocupar um campo no documento. Por permitir a inclusão de vários dados em um mesmo atributo, esse relacionamento também possibilita a visualização de um *array*.

Após a definição dos relacionamentos, Vera et al. (2015) utilizaram a mesma cardinalidade de relacionamentos existentes no modelo ER (1: 1, 1: N, N: M). Em modelos

orientados a documentos, os dados são organizados em documentos e coleções de documentos. Dessa forma, os autores propuseram três novos construtores para representar os relacionamentos e organização de dados em modelos orientados a documentos, a notação gráfica de cada um desses novos construtores pode ser vista na Figura 6, onde:

- Documentos e coleções: são graficamente representados por um documento;
- Relacionamento de documentos embutidos: representação gráfica de um documento dentro de outro;
- Relacionamento de referências: representado por uma flecha, onde a ponta da flecha indica o documento que possui a referência. Em casos de relacionamentos N: M, a flecha possui uma ponta em cada extremidade, indicando que cada documento participante do relacionamento possui uma referência do outro documento.

Figura 6 – Notação do Modelo conceitual proposto



Fonte: O autor (2019)

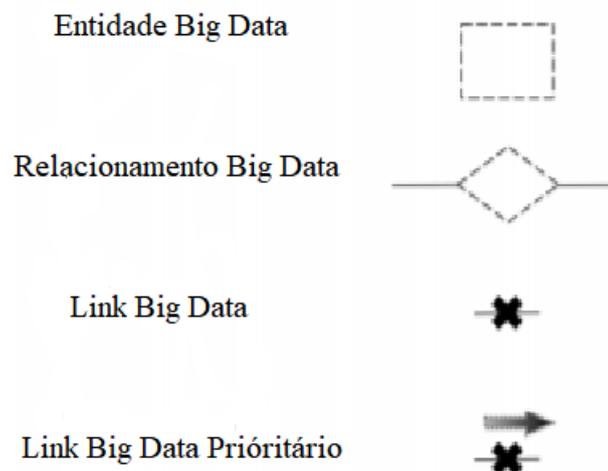
O objetivo de Vera et al. (2015), consiste em prover um diagrama único, isto é, um diagrama que possa ser utilizado na construção de esquemas de dados lógicos para qualquer sistema NoSQL orientado a documentos. O foco do trabalho é que o diagrama proposto seja intuitivo e claro, facilitando a compreensão do projeto do esquema de banco de dados, a fim de fornecer atributos chaves que previnam erros de interpretação, como: modelar um relacionamento entre documentos de forma diferente do necessário para o domínio. Assim, um dos objetivos do uso do modelo de dados proposto por Vera et al. (2015) é evitar que problemas de construção do projeto lógico acarretem em problemas nas próximas fases do projeto de banco de dados. Todavia, mesmo que o foco do trabalho deles tenha sido a construção de um diagrama que facilite a compreensão nos projetos de BD orientado a

documentos, não foi realizado um estudo que investigue se o diagrama proposto realiza ou não o que se pretende.

3.1.2 Modelos para aplicações híbridas

Diferentemente de Chebotko, Kashlev e Lu (2015), Hamouda e Zainol (2017), Lima e Mello (2016), Roy-Hubara et al. (2017), que mapeiam o modelo ER para modelos lógicos NoSQL, Villari et al. (2016) propuseram usar um único modelo para a representação diferenciada de dados *big data* (termo usado na Internet e computação na nuvem para expressar grandes conjuntos de dados, de diferentes tipos (variedade), de petabyte a zettabyte (volume), gerados de forma rápida (velocidade) e dados com baixa consistência ou confiança (veracidade)) e dados convencionais em domínios híbridos, onde esses domínios são aqueles que possuem diferentes tipos de dados, convencionais ou não convencionais. Para representar esses domínios, Villari et al. (2016) sugeriram a extensão do modelo ER, adicionando quatro novos construtores ao modelo, esses podem ser vistos na Figura 7:

Figura 7 – Notação do Modelo ER Enriquecido



Fonte: (VILLARI et al., 2016)

1. Entidade *Big Data*: utilizado para representar dados big data, sendo graficamente representado por um retângulo tracejado.
2. Relacionamento *Big Data*: esse relacionamento deve ser utilizado quando ambas as entidades do relacionamento são *big data*, ou apenas uma das entidades for do tipo big data.

3. *Link Big Data*: esse *link* indica que uma das entidades do relacionamento possui um *link* da outra entidade com que se relaciona.
4. Link Big Data prioritário – utilizado para indicar a necessidade de uma entidade big data ou relacionamento ter mais de um link para sua identificação.

O foco do trabalho de Villari et al. (2016) foi possibilitar a distinção gráfica entre dados *big data* e dados convencionais, tornando possível utilizar um único modelo para representar esquemas de dados em domínios híbridos, isto é, um domínio que possua diferentes tipos de dados. Entretanto, a extensão proposta por Villari et al. (2016) não mostra resultados que justifiquem sua aplicação, e indiquem o que essa tem de diferente de outras abordagens que a torna melhor para ser usada em domínios híbridos.

Em Zečević et al. (2018) foi proposto a utilização da metodologia MDA (Arquitetura Dirigida pelo Modelo), para a criação de esquemas de dados para domínios híbridos. A metodologia MDA, utilizada em Zečević et al. (2018), possui três níveis:

- CIM (Modelo Independente de Computação)- o nível mais abstrato, contém uma descrição de conceitos do domínio;
- PIM (Modelo Independente de Plataforma)- descreve o problema do domínio abordando detalhes de implementação;
- PSM (Modelo Específico de Plataforma)- nesse nível são abordados conceitos da plataforma de implementação.

Dos três níveis, apenas dois foram abordados para a construção do modelo proposto em Zečević et al. (2018), esses foram: o PIM, descrevendo características de implementação do armazenamento de dados, e o PSM, especificando termos da tecnologia de implementação.

Para o **PIM**, foram construídos meta-modelos para os bancos de dados relacionais e NoSQL (chave-valor, documentos e grafos). Cada meta-modelo possui características específicas do sistema escolhido. Para **PSM**, foi feita a integração dos diferentes meta-modelos construídos no PIM, sendo adicionadas restrições e novos elementos referentes às tecnologias de implementações escolhidas. Sendo assim, a junção dos modelos em um único modelo.

Para ilustrar o processo de construção do modelo de dados na abordagem desenvolvida, Zečević et al. (2018) aplicaram a abordagem em um sistema híbrido de Serviço de Informação Baseado em Agricultura de Precisão, sendo esse construído para sistemas relacionais e orientado a documentos. Os resultados obtidos do estudo mostraram que a abordagem proposta pelos autores auxilia no processo de manutenção do sistema, pois, por conter um esquema lógico de dados, torna mais simples identificar que partes do sistema serão afetadas em caso da adição de uma nova característica ou da remoção de uma, por exemplo.

3.1.3 Modelos conceituais baseados em UML

Em Abdelhedi et al. (2016), é proposto um processo intitulado UMLtoNoSQL, que tem como objetivo definir, especificar e automatizar o processo de armazenamento de *big data* em sistemas NoSQL. Esse processo se baseia na metodologia MDA (Arquitetura Dirigida por Modelo). Dos três pontos de vista da metodologia MDA, somente dois foram abordados no trabalho de Abdelhedi et al. (2016): o PIM (Modelo Independente da Plataforma) e o PSM (Modelo Específico da Plataforma).

No PIM, os autores usaram o diagrama de classes UML para descrever *big data*, resultando na proposta do: PIM conceitual e do PIM lógico. O PIM conceitual corresponde a uma tupla (N, C, L) , onde:

- N é o nome do diagrama de classe;
- C é um conjunto de classes, e cada classe c pertencente C é uma tupla $(N, A, IdentO)$, onde:
 1. $c.N$ é o nome da classe;
 2. $c.A$ é o conjunto de atributos dessa classe;
 3. $c.IdentO$ é um atributo especial de c que representa o identificador do objeto.
- L é um conjunto de *links*. Cada *link* l entre n classes com $n \geq 2$ é definido pela tupla (N, Ty, Pr^1) :
 1. $l.N$ é o nome do *link*;
 2. $l.Ty$ é o tipo de *link*: associação, composição ou generalização;
 3. $l.Pr^1$ é um conjunto de n pares, onde cada par possui um *link* da classe c e sua cardinalidade.

Já o PIM lógico descreve a organização dos dados de acordo com características comuns dos sistemas NoSQL orientados à colunas, onde cada banco de dados orientado à coluna consiste em um conjunto de tabelas. Cada tabela possui uma coleção de linhas com tamanhos variados, e cada uma dessas linhas é identificada pela chave dessa linha, que é única. Cada banco de dados possui uma tabela única, chamada T . Esta tabela, possui um conjunto de famílias de colunas, chamada de f , onde f é uma tupla (N, COL, ID) :

- $f.N$ é o nome da família de colunas;
- $f.COL$ é um conjunto de colunas, e cada coluna possui o conjunto (N, T, TS) :
 1. $COL.N$ é o nome da coluna;
 2. $COL.T$ é o tipo da coluna;

3. COL.TS é o *timestamp* dessa coluna;

Contudo, os autores não consideraram o *timestamp* como parâmetro, mas usaram um identificador único para a família de colunas: o f.ID.

A partir da definição do modelo lógico (chamado PIM lógico), os autores mapearam o modelo lógico para o físico (PSM), por meio de um conjunto de regras, listadas a seguir:

1. Pacote para tabela: cada pacote em UML é convertido para uma tabela;
2. Classe para família de colunas: todos os atributos da classe são agrupados em uma mesma família de colunas. O OID, utilizado para identificar um atributo em UML, é convertido para o atributo chave da coluna;
3. Classe associativa em família de colunas: como qualquer outra classe em UML convertida para o modelo proposto por Abdelhedi et al. (2016), uma classe associativa deve ser convertida para uma família de colunas, onde haverá um atributo chave para identificar as colunas que possuem associação;
4. Relacionamento associativo em família de colunas: cada associação n-ária deve ser convertida para uma nova coluna composta de n colunas, onde essa coluna deverá ter um atributo chave das demais colunas. Assim, isto permite o uso da referência em outras famílias de colunas;
5. Relacionamento de composição/ agregação: é realizada por meio de referências, onde cada composição/ agregação é denotada por uma nova coluna, onde a chave da coluna fica responsável por identificar o relacionamento entre as colunas.

Os autores quiseram demonstrar que, aplicando as regras de conversão propostas, essas poderiam ser aplicadas em qualquer sistema NoSQL orientado à colunas. Para exemplificar isso, Abdelhedi et al. (2016) utilizaram suas regras de conversão em dois SGBD NoSQL orientados à colunas: Cassandra e Hbase.

Em BENMAKHLOUF (2018) é proposto o mapeamento de UML para modelos lógicos DOM (Modelo Orientado a Documentos), que consiste em cinco passos lógicos:

1. Classe UML deve ser mapeada para uma coleção e os atributos da classe são convertidos em atributos do documento dessa coleção;
2. Associação 1: 1
 - DOM: das duas coleções associadas, a chave primária de uma delas deve se tornar chave estrangeira da outra coleção.
 - DOM aninhado: cria-se uma coleção aninhada representando uma das classes com seus atributos, a outra classe se torna uma sub-coleção.

3. Associação 1: N

- DOM: a chave primária da coleção que possui a cardinalidade 1, obrigatoriamente se torna chave estrangeira na coleção que tem a cardinalidade N.
- DOM aninhado: uma nova coleção principal precisa ser criada representando a classe do lado N, enquanto a classe do lado 1 se torna uma sub-coleção da coleção principal.

4. Associação N: N

- DOM: uma nova coleção deve ser criada, sendo composta pelas duas chaves primárias e atributos das classes participantes da associação.
- DOM aninhado: cria-se uma nova coleção, contendo as chaves primárias das classes participantes. Essas se tornam sub-coleções da nova coleção.

5. Associação por herança, onde o modelo DOM passa a ser composto por coleções do tipo pai e filho:

- DOM: a coleção filha deve ter os seus atributos não chave contidos nela e uma chave primária para identificá-la e diferenciá-la de suas irmãs. Já na coleção pai, adiciona-se um atributo *tipo*, que especifica os filhos que a coleção pai tem.
- DOM aninhado: uma nova coleção é criada representando a classe pai, que tem todos os seus atributos e sua chave primária, além das chaves primárias de todos os seus filhos.

O principal objetivo de BENMAKHLOUF (2018), ao propor o mapeamento de UML para o DOM, foi possibilitar a construção de esquemas de dados para modelos de dados NoSQL orientado a documentos, esse mapeamento baseou-se em coleções aninhadas. Onde o uso dessas coleções teve o intuito de manter a consistência de dados no processo de mapeamento, e conseqüentemente, reduzir o tempo de análises multidimensionais em grandes quantidades de dados pela remoção das junções. A validação do mapeamento de UML para o DOM foi feita usando dois SGBD: MongoDB (orientado a documento) e SQL (relacional).

Shin, Hwang e Jung (2017), propuseram um processo de conversão de modelos conceituais para modelos lógicos NoSQL. Contudo, apenas uma conversão foi feita: de UML para um modelo lógico orientado a documentos. O processo de conversão ocorreu da seguinte forma:

1. Conversão de classe: uma classe em UML deve ser convertida para uma coleção;
2. Conversão de atributo: um atributo deve ser mapeado para uma coluna;

3. Relacionamento associativo: é convertido para um relacionamento de referências (contendo referências dos objetos que possuem um relacionamento) ou relacionamento embutido (todos os objetos participantes do relacionamento devem estar em uma mesma classe).

Shin, Hwang e Jung (2017) apresentaram conceitos de chaves na conversão do UML, onde: a chave primária identifica um item da coleção; a chave estrangeira é utilizada nos relacionamentos associativos, sendo esses relacionamentos por referência ou embutidos. Para exemplificar o modelo proposto, os autores utilizaram uma aplicação de comércio eletrônico. Contudo, Shin, Hwang e Jung (2017) não apresentaram benefícios oferecidos do processo de conversão de modelos de dados, ou qual seria o seu impacto quando aplicado em projetos orientados a documentos.

De modo diferente dos estudos já citados (ABDELHEDI et al., 2016), (BENMAKHLLOUF, 2018), (SHIN; HWANG; JUNG, 2017) que utilizam o diagrama UML e realizaram conversões para modelos lógicos de aplicações NoSQL, Bugiotti et al. (2014) propuseram uma nova metodologia (intitulada NoAM) para o desenvolvimento de esquemas de dados voltados para BD NoSQL, baseando-se em UML. A metodologia NoAM se fundamenta na visão de aplicações que possuem dados agregados, levando em consideração duas formas de representar agregados:

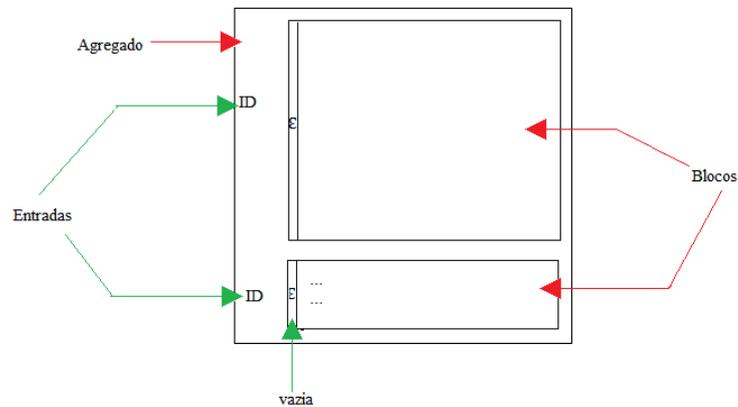
1. Entrada por objeto agregado: cada agregado individual possui uma única entrada. A entrada da chave é vazia e o valor de entrada é representado por todo agregado;
2. Entrada por campo de nível superior: cada agregado possui múltiplas entradas, onde cada entrada identifica um valor desse agregado.

A partir das considerações de como representar agregados, os autores definiram como um BD deve ser representado no modelo proposto, a notação gráfica deste modelo é exposta na Figura 8, assim, cada uma das propriedades do NoAM são descritas a seguir:

1. Coleção – um banco de dados é um conjunto de coleções, assim, para cada coleção, deve ser atribuído um nome;
2. Blocos – uma coleção é um conjunto de blocos, dessa forma, cada bloco na coleção, é identificado pela sua chave;
3. Entradas – as entradas no NoAM podem seguir duas formas: entrada por objeto agregado e entrada por campo de nível superior, ambas apresentadas anteriormente.

Bugiotti et al. (2014) implementaram o modelo proposto no domínio de jogos, os resultados obtidos no estudo mostraram que operações envolvendo a recuperação de dados foi superior quando a entrada utilizada foi a entrada por objeto agregado. Contudo, o foco do trabalho foi apresentar uma nova metodologia de modelagem de dados que pudesse ser

Figura 8 – Notação NoAM



Fonte: O autor (2019)

aplicada a qualquer sistema NoSQL. Porém, apenas três categorias NoSQL (chave-valor, orientada à coluna, orientada a documento) foram contempladas com a abordagem, por possuírem algumas características em comum, por exemplo, o conceito de dados agregados.

3.1.4 Modelagem lógica para representação de dados agregados

Em Jovanovic e Benson (2013), é proposto um modelo lógico para dados agregados destinado a banco de dados NoSQL. O modelo lógico foi construído através da inclusão de três novos operadores na linguagem de modelagem de dados IDEF1X (Modelagem de Definição e Integração da Informação). São eles:

- *ROOT*: operador que designa quais entidades serão raízes e quais entidades serão referenciadas à entidade raiz. Esse operador possui um identificador para auxiliar na identificação de relacionamentos com outros agregados;
- *REFI*: usado ao lado da entidade que possui a referência de outra entidade;
- *EMBED*: tem o propósito de representar uma entidade dependente de outra, ou seja, uma entidade que necessita de outra para existir. Esse operador também pode ser usado para designar subtipos.

Os autores aplicaram o modelo lógico proposto em um domínio na área de saúde. Por meio desse domínio, foi explicado como utilizar os novos operadores e como construir esquemas de dados lógicos a partir da extensão proposta por Jovanovic e Benson (2013), usando linguagem de modelagem de dados IDEF1X.

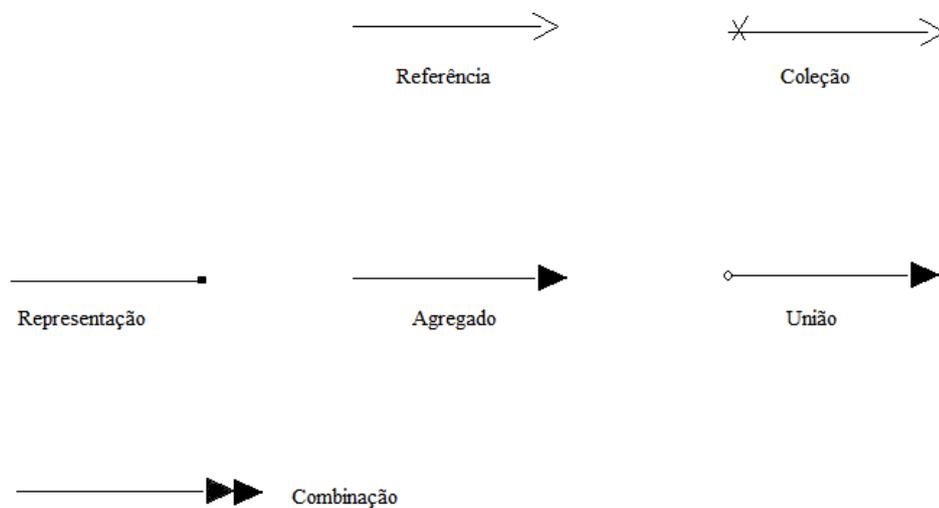
3.1.5 COMN (Modelagem de Notação Objeto Conceitual)

Hills (2016) propôs a notação gráfica COMN, que torna possível a representação de dados convencionais e dados NoSQL em um único modelo de dados. Além disso, a mesma consegue representar *arrays*, tipos aninhados e propriedades que outras notações não abordam. A notação proposta por Hills (2016) possui as três fases de modelagem de um banco de dados: conceitual, lógica e física.

Na fase conceitual, o autor definiu três formas de representar os dados:

1. Objeto é uma classe que possui uma coleção de dados. Um objeto é representado por um losango, como visto na Figura 9;
2. O tipo indica a extensão de uma classe. Graficamente, um tipo é representado no COMN na forma de um retângulo, sua representação gráfica é vista na Figura 9;
3. O Conceito indica um estado do dado. É representado no modelo por um retângulo com as pontas arredondadas (ver Figura 9).

Figura 9 – Notação COMN



Fonte: O autor (2019)

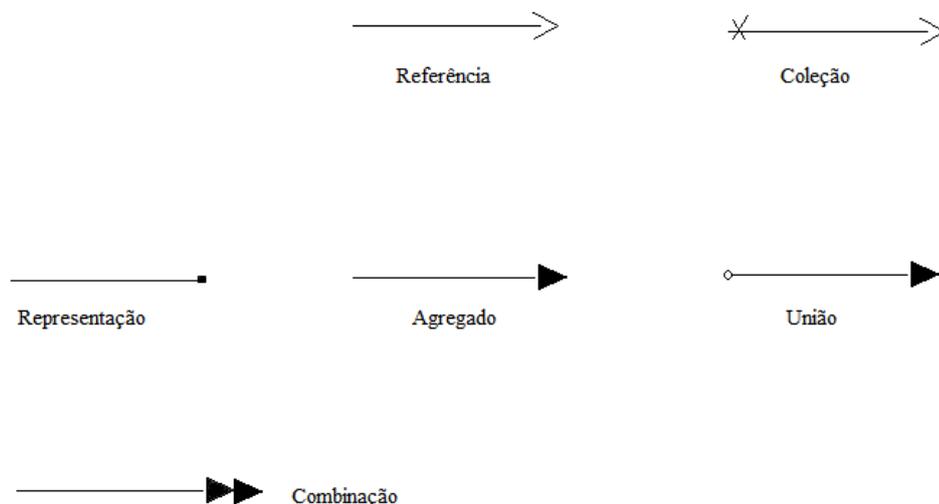
Os relacionamentos definidos no modelo COMN dividem-se em:

- Relacionamentos de composição, que possuem os seguintes subtipos:
 1. Agregação: indica que determinado objeto possui dados de outro objeto, que devem estar agregados um ao outro, sua representação gráfica é vista na Figura 10;

2. União: esse relacionamento indica que uma quantidade de dados deve estar unida para tornar mais simples a recuperação de uma informação, notação descrita na Figura 10;
 3. Combinação: indica a dependência de um dado para outro, ou seja, esse dado necessita de outro para a informação ter algum significado, ver Figura 10.
- Relacionamentos de não composição:
 1. Referência: relacionamento que indica que determinado objeto possui referência em outro, sua representação gráfica é mostrada na Figura 10;
 2. Coleção: o relacionamento coleção deve ser utilizado quando um objeto possui um conjunto de suas variáveis em outro objeto, sua representação gráfica é descrita na Figura 10;
 3. Representação: é utilizado para indicar que um objeto é dependente de outro (ver Figura 10).

Para Hills (2016), os relacionamentos entre os dados e sua organização podem ser vistos de duas formas: físicos e conceituais. Cada objeto ou relacionamento físico é aquele que existe fisicamente, e esse é representado da forma descrita acima. Enquanto relacionamentos ou objetos conceituais são aqueles que não existem fisicamente, e sua forma de representação muda da descrita acima, usando formas tracejadas para representá-los.

Figura 10 – Notação COMN Relacionamentos



Fonte: O autor (2019)

A cardinalidade dos relacionamentos no COMN consiste no uso dos símbolos + e *, sendo (+) caracterizado por multiplicidade, ou seja, um ou mais; enquanto (*) representa

zero ou mais. Onde há ausência explícita de um dos sinais na extremidade da linha indicativa de relacionamento, esse, por padrão, é representando por 1.

A conversão do modelo conceitual para o lógico é feita através da definição de detalhes de implementação, como indicando os tipos de dados que serão armazenados. Já a conversão do modelo lógico para o físico é feita através da adição de aspas (« »), onde dentro delas define-se a forma de armazenamento escolhida e detalhes de implementação, ou seja, o modelo construído é dependente do SGBD. A conversão do modelo conceitual para o lógico, e em seguida para o físico, não é aprofundada nesta dissertação, pois o objetivo aqui é apresentar a parte conceitual. Contudo, uma explicação mais completa pode ser vista em Hills (2016).

O autor ilustrou através do estudo de caso (domínio de uma cafeteria) como a modelagem de dados deve ser feita em cada nível do COMN. Na fase conceitual, Hills (2016) mostrou que, graças aos relacionamentos, o modelo conceitual pode se adequar a qualquer banco de dados, sendo o principal objetivo do autor, solucionar algumas limitações que outros diagramas possuem, como a não capacidade de expressar dados semiestruturados ou não estruturados, arrays, tipos aninhados, e ao mesmo tempo prover de um modelo que possa ser utilizado por sistemas relacionais e NoSQL.

3.1.6 GOOSSDM (Modelo de dados semiestruturado orientado a grafos)

Em Banerjee et al. (2015), o modelo GOOSSDM, proposto em Sarkar, Ganguly e Chatterjee (2013), foi revisitado e mapeado para um modelo lógico, com o objetivo de dar suporte à representação de dados semiestruturados do tipo *big data*. O mapeamento realizado por Banerjee et al. (2015) consiste em doze passos:

1. Cada GSE (Grupo Semântico Elementar) passa a ser expresso como propriedades e itens de um objeto ou *array* em JSON;
2. Cada GSC (Grupo Semântico Conceitual) deve ser expresso como um objeto ou *array*;
3. Notação passa a ser um campo de descrição, que pode conter detalhes do esquema e do objeto do esquema;
4. Referências de GSE e GSC são expressas pela *tag* cifrão ref;
5. O GSC de nível mais alto deve ser tratado como objeto raiz ou *array* raiz;
6. Qualquer GSC de nível mais baixo que possua algum GSE passa a ser expresso como um *array*, ou objeto aninhado ao objeto, ou *array* raiz;
7. GSC que possui alguma referência ao GSE com caminho adjacente mais baixo se torna objeto ou *array* raiz desse GSE;

8. Os relacionamentos entre *links* do GSC com outros GSC de níveis mais baixos devem ser expressos usando *extends*;
9. Relacionamento entre um conjunto do GSC espalhados em vários níveis deve ser expresso como objeto ou *array* aninhado;
10. Qualquer nível GSC que tenha relacionamento de associação com um nível adjacente passa a ser expresso como um objeto ou *array* aninhado;
11. Relacionamento de associação entre dois GSC de mesmo nível deve ser expresso como um objeto ou *array* aninhado em um esquema JSON;
12. Um conjunto de GSC espalhados através de vários níveis deve ser expresso como um objeto ou *array*.

Após a definição dos doze passos de conversão citados acima, Banerjee et al. (2015) ilustraram através de um sistema de gerenciamento de projetos, como se daria a construção desse sistema utilizando o modelo lógico proposto, sendo depois o esquema de dados modelado e implementado no MongoDB. A partir da construção do esquema de dados, os autores analisaram se o modelo lógico atendia as especificações *big data*, como a representação de dados aninhados. Os resultados obtidos no estudo apontaram que o modelo atendia parcialmente algumas características para a modelagem de dados *big data*. Contudo, Banerjee et al. (2015) indicaram que o modelo proposto atendia todas as especificações necessárias para a construção de esquemas NoSQL orientados a documentos, assim, podendo ser utilizado com esses sistemas.

3.2 AVALIAÇÃO DE MODELOS DE DADOS POR ESTUDOS EMPÍRICOS

Nesta seção, são abordados estudos experimentais que analisaram modelos conceituais de dados, já que uma parcela relativamente grande dos estudos realizaram testes com o intuito de obter informações a respeito da qualidade dos modelos conceituais. As métricas e os objetivos de cada um desses estudos serão discutidos nessa seção.

3.2.1 Análise do diagrama ER (Entidade Relacionamento)

Em Genero et al. (2001) foi investigada a complexidade de diagramas produzidos a partir do modelo ER e se essa complexidade afetava a manutenção desses diagramas. Para isso, os autores realizaram um estudo experimental, onde o perfil dos participantes que executaram o experimento foi de professores e alunos. Os professores tinham no mínimo cinco anos de experiência em projetos utilizando o ER, enquanto os alunos estavam no último ano da graduação em Ciência da Computação. Antes da execução do experimento, os estudantes participaram de um minicurso, onde foi ministrado conceitos importantes a respeito da modelagem de dados utilizando o ER.

Genero et al. (2001) elaboraram um conjunto de métricas para auxiliar na investigação efetuada em seu estudo. As métricas foram:

1. O número total de entidades dentro de um Diagrama Entidade Relacionamento (DER);
2. O número total de relacionamentos contidos em um DER;
3. O número total de relacionamentos N: M dentro de um DER;
4. O número total de relacionamentos 1: N e 1:1 contidos em um DER;
5. O número total de relacionamentos binários de um DER;
6. O número total de atributos de um DER;
7. O número total de atributos derivados de um DER;
8. O número total de atributos compostos de um DER;
9. O número total de atributos multivalorados de um DER;
10. O número total de especializações e generalizações contidas no DER;
11. O número total de relacionamentos redundantes existentes no DER;
12. O número total de relacionamentos reflexivos do DER.

A partir das métricas definidas, a execução do experimento se deu da seguinte maneira: cada participante recebeu problemas de um domínio com complexidades similares, com o prazo de uma semana para a sua resolução. Após a conclusão dessa atividade, foi dado a cada participante, um questionário que incluiu detalhes pessoais, como experiências, além de uma segunda parte com perguntas a respeito do DER.

Os autores definiram as seguintes subcategorias para analisar a manutenção dos diagramas produzidos: compreensão, simplicidade, analisabilidade, modificabilidade, testabilidade e estabilidade. Os participantes puderam classificar as subcategorias em sete níveis, do extremamente difícil ao extremamente simples. Os achados encontrados em Genero et al. (2001) apontaram para uma tendência de, quanto melhor as subcategorias fossem avaliadas, melhores resultados seriam obtidos na manutenção dos diagramas desenvolvidos. Contudo, uma parcela dos participantes selecionados já possuía um período de experiência no uso do ER, o que pode ter favorecido os resultados coletados. Ademais, como o estudo foi executado com uma parcela pequena de participantes, não há como saber se as métricas utilizadas foram ideais para obtenção dos resultados esperados, devendo o estudo ser executado mais vezes, a fim de obter mais dados que ajudem na investigação.

Em Genero, Poels e Piattini (2002), foi investigada a relação entre a complexidade estrutural de modelos conceituais ER e a manutenção de subcaracterísticas desses modelos

(como entendimento e facilidade de modificação). A investigação foi feita a partir de um experimento, onde as métricas analisadas foram as mesmas definidas no estudo de Genero et al. (2001). Contudo, o material utilizado em Genero, Poels e Piattini (2002) consistiu de quatro diagramas modelados no ER (com níveis de complexidade distintos). Esse nível de complexidade diferente foi utilizado propositalmente, a fim de obter dados de como a complexidade estrutural poderia impactar no entendimento do DER (Diagrama Entidade Relacionamento).

Os participantes do estudo (estudantes de Computação) receberam os quatro diagramas, sendo que a análise de cada diagrama se dividiu em duas etapas:

1. Consistiu em identificar se os participantes entenderam o domínio modelado no diagrama. Foi entregue a cada participante, um questionário contendo cinco questões, observando-se o tempo de conclusão para cada questionário.
2. A inclusão de novos requisitos no diagrama foi pedida, sendo observado também o tempo de conclusão dessa atividade.

A partir do tempo levantado nas duas etapas, os autores montaram uma correlação entre as métricas selecionadas e analisaram como a complexidade do diagrama influenciou os tempos medidos (entendimento e modificação). De acordo com Genero, Poels e Piattini (2002), as métricas que mais influenciaram nos tempos medidos foram:

1. O número total de entidades de um DER;
2. O número total de relacionamentos contidos em um DER;
3. O número total de relacionamentos N: M de um DER;
4. O número total de relacionamentos 1: N e 1:1 contidos em um DER;
5. O número total de relacionamentos binários de um DER;
6. O número total de atributos de um DER;

Os resultados obtidos em Genero, Poels e Piattini (2002) corroboraram com a hipótese feita pelos autores que, quanto maior a complexidade estrutural de um diagrama ER, mais tempo o usuário precisa para entendê-lo e realizar modificações nesse diagrama.

Em Hussain et al. (2006), foi executado um experimento controlado, com o objetivo de expor como a complexidade estrutural de diagramas ER (no nível conceitual) impactava na compreensão (entendimento), análise e tempo de modificação do esquema ER. O estudo foi conduzido utilizando as cinco primeiras métricas listadas acima. Além dessas métricas, Hussain et al. (2006) definiram duas novas métricas:

1. Número total de atributos do DER, incluindo os atributos de relacionamentos;

2. Número total de relacionamentos n-ários presentes no DER.

O experimento foi executado por 20 estudantes de Ciência da Computação (bacharelado e pós-graduação) com experiência mínima de um ano e máxima de doze anos. Cada estudante recebeu um questionário com 24 questões a respeito do domínio representado pelo diagrama. As informações obtidas nos questionários aplicados aos estudantes foram analisadas usando o Processo Hierárquico Analítico (PHA), isto é, uma técnica que divide um problema complexo em estruturas hierárquicas, mostrando os relacionamentos entre o objetivo, os critérios e as alternativas.

Os resultados obtidos em Hussain et al. (2006) confirmaram a hipótese levantada de que, quanto maior a complexidade do domínio representado, maior o tempo necessário para modificação e entendimento, corroborando com os achados de Genero et al. (2001) e Genero, Poels e Piattini (2002). Além disso, o estudo de Hussain et al. (2006) mostrou que o número de relacionamentos é mais importante que o número de atributos para o impacto da complexidade estrutural.

Em Moody (2003), foi avaliada a qualidade de modelos de dados conceituais sob a óptica da complexidade estrutural desses modelos. Para definir as métricas utilizadas no estudo, o autor se baseou nas especificações de qualidade de modelos conceituais definidas em Moody e Shanks (1994), que se dividem em:

1. *Stakeholders*: pessoas envolvidas no desenvolvimento ou uso do modelo de dados;
2. Fatores de qualidade: são características do modelo de dados que contribuem para a sua qualidade;
3. Qualidade das métricas: definem modos de medir um determinado fator de qualidade, e uma mesma métrica pode medir mais de um fator;
4. Comparação: define a importância de diferentes fatores de qualidade;
5. Aperfeiçoamento de estratégias: são técnicas utilizadas para o melhoramento de um ou mais fatores de qualidade.

A partir das especificações propostas em Moody e Shanks (1994), Moody (2003) definiu oito fatores de qualidade para modelos de dados conceituais:

1. Corretude: identifica se a representação está de acordo com as especificações do problema dado e de acordo com as regras definidas pelo modelo de dados;
2. Completude: indica se a representação contém toda a informação necessária para o funcionamento do sistema;
3. Integridade: verifica se o modelo de dados representa todas as regras necessárias para a sua aplicação;

4. Flexibilidade: é a facilidade que o modelo oferece para que mudanças sejam feitas;
5. Compreensão: é a facilidade com a qual os conceitos e as estruturas do modelo são entendidos;
6. Integração: é a consistência do modelo de dados com a organização dos dados;
7. Implementação: definida pela facilidade com que o modelo pode ser traduzido para o sistema.

Após a definição dos fatores de qualidade, Moody (2003) definiu as seguintes métricas para cada fator de qualidade apresentados acima. Essas métricas são expostas a seguir:

- Corretude:
 - (a) Número de violações do padrão de modelagem de dados;
 - (b) Número de instâncias com entidades redundantes;
 - (c) Número de instâncias com relacionamentos redundantes;
 - (d) Número de atributos redundantes.
- Completude:
 - (a) Número de requisitos perdidos;
 - (b) Número de requisitos supérfluos;
 - (c) Número de requisitos definidos incorretamente;
 - (d) Número de inconsistências com o processo do modelo.
- Integridade:
 - (a) Número de regras de negócio perdidas;
 - (b) Número de regras de negócio definidas incorretamente;
 - (c) Número de regras de negócio inconsistentes com o modelo de processo;
 - (d) Número de regras de negócio redundantes.
- Flexibilidade:
 - (a) Número de elementos do modelo de dados que foram alterados;
 - (b) Probabilidade ajustada com o custo de mudança;
 - (c) Impacto de mudança.
- Compreensão:
 - (a) Classificação de compreensão do usuário;
 - (b) Interpretação de erros do usuário;
 - (c) Classificação de compreensão do desenvolvedor da aplicação;
 - (d) Proporção área-entidade;
 - (e) Proporção entidade-atributo, obtida através do número de entidades e atributos presentes no diagrama.

- Simplicidade:
 - (a) Número de entidades [E];
 - (b) Complexidade do sistema, sendo essa obtida da soma do número de entidades e relacionamentos [E + R];
 - (c) Complexidade total, obtida pela soma de todas as entidades, relacionamentos e atributos [aE + bR + cA].
- Integração:
 - (a) Número de conflitos de dados com o modelo corporativo;
 - (b) Número de conflitos de dados com o sistema existente;
 - (c) Número de itens de dados duplicados no projeto;
 - (d) Classificação do modelo, isto é, habilidade de encontrar todas as necessidades do projeto;
- Por fim, a Implementabilidade foi medida através de duas métricas:
 - (a) Desenvolvimento de estimativa de custo;
 - (b) Classificação de risco técnico.

Após a definição das métricas, o estudo foi conduzido em um laboratório com estudantes universitários, onde foi passado a cada estudante, o mesmo problema de domínio a ser modelado no ER. Os diagramas obtidos dessa atividade foram avaliados por Moody (2003) e membros do grupo de administração de dados de duas companhias, responsáveis por assegurar a qualidade dos modelos de dados construídos em suas organizações.

O autor fez um levantamento para verificar se as métricas definidas por ele ajudariam ou não a medir a qualidade de um modelo de dados. Dentre as 29 métricas definidas, apenas 3 se mostraram eficientes para medir a qualidade, sendo essas: Desenvolvimento de estimativa de custo, que promoveu a participação do desenvolvedor mais cedo no projeto; Número de itens de dados duplicados no projeto, que auxiliou na identificação de possíveis reduções do esquema de dados e aperfeiçoamento no processo de modelagem; Complexidade do sistema [E + R], que se mostrou útil para analisar o nível de complexidade de esquemas de dados. Com os resultados obtidos em Moody (2003), foi exposto que as métricas definidas pelo autor devem ser repensadas, visto que, grande parte das métricas definidas não obtiveram dados que auxiliassem na identificação da qualidade de modelos de dados.

3.2.2 Análise de modelos de dados conceituais

Em Cherfi, Akoka e Comyn-Wattiau (2002) foram apresentadas métricas para a avaliação de esquemas conceituais construídos no EER e UML. Os autores realizaram um levantamento da literatura, que mostrou que há três formas de se analisar a qualidade em modelos conceituais, são elas:

1. Uso: indica se a forma de representação retrata de maneira completa a realidade observada. Esse aspecto normalmente é medido através de análises de completude e compreensibilidade.
2. Especificação: é voltada para a perspectiva do usuário. Nela é possível obter informações se, a partir da visão do usuário, os requisitos de modelagem foram cumpridos para determinado domínio. Métricas utilizadas nesse tipo de análise são - expressividade, legibilidade, simplicidade e corretude.
3. Implementação: relaciona-se com a implementação do sistema através do modelo, mede-se a implementação através da manutenção.

Dentre as três perspectivas para a avaliação da qualidade de modelos conceituais, Cherfi, Akoka e Comyn-Wattiau (2002) levaram em consideração apenas uma: a especificação, medida através de três métricas:

1. Legibilidade: indica se um esquema representado a partir do modelo pode ser lido facilmente;
2. Expressividade: indica a expressividade que avalia se o modelo conceitual captura os aspectos da realidade;
3. Simplicidade: quanto menos construtores o modelo possuir, mais simples ele é. Ou seja, com o acréscimo de construtores, a estrutura se torna mais complexa.

O estudo de Cherfi, Akoka e Comyn-Wattiau (2002) se deu a partir da construção de oito esquemas retratando um mesmo domínio hospitalar, e a qualidade de cada um dos modelos foi medida através das métricas selecionadas. Os próprios autores avaliaram os modelos desenvolvidos, analisando os aspectos escolhidos, onde os resultados mais favoráveis estão associados aos esquemas com uma menor complexidade. Dentre as contribuições do estudo, está a visualização de três perspectivas na análise de qualidade de modelos conceituais, e além disso, o levantamento feito pelos autores pode ser utilizado em outras pesquisas.

Em Poels (2003), foi apresentado um estudo empírico avaliando dois modelos: REA (Recurso-Evento-Agente) e ER. O objetivo do estudo foi comparar a qualidade de ambos os modelos quando utilizados na modelagem conceitual de processos de negócios.

Os participantes envolvidos no experimento foram alunos de pós-graduação, matriculados na disciplina de Sistemas de Informação Contábeis. Para introduzi-los ao tema da pesquisa, foi ministrado um curso introdutório de ER e REA de quatro horas e meia, onde foram apresentados conceitos básicos do ER e REA. Após o término do curso, cada participante realizou tarefas similares às planejadas para a execução do estudo, a fim de familiarizá-lo com as atividades. Contudo, os diagramas construídos nessas atividades representavam processos diferentes para não influenciar nos resultados do estudo.

Ao final do curso, o estudo foi executado a partir da percepção do usuário e de como esse compreendia os diagramas desenvolvidos a partir dos modelos ER e REA. A compreensão dos usuários foi observada de duas formas:

- Precisão das respostas: medida através de perguntas específicas de compreensão do domínio representado nos diagramas;
- Tempo de compreensão: medido por meio do tempo utilizado para responder o questionário.

Cada processo representado no diagrama possuiu um questionário com 13 perguntas relacionadas com o domínio representado no diagrama. Foram analisadas as respostas dadas e o tempo de conclusão do questionário. A partir da análise feita nos questionários de Poels (2003) e no tempo usado pelos participantes para respondê-lo, o autor expôs que a compreensão dos usuários foi mais alta quando observados os diagramas REA ao invés dos diagramas ER nos domínios estudados, mostrando a tendência de que o uso do REA em domínios financeiros é mais recomendável, por ter obtido um melhor nível de compreensão dos usuários. Apesar disso, novos estudos devem ser feitos para analisar se a tendência observada em Poels (2003) ocorre em outros domínios financeiros.

Poels et al. (2011) investigaram se o uso do modelo REA (Recurso-Evento-Agente) ajudava na compreensão do usuário de processos de negócios. Para obter informações se o uso do REA auxiliava ou não na compreensão dos usuários, foi realizado um estudo experimental com dois tratamentos: REA e ER. Os participantes do estudo foram 30 alunos de graduação matriculados na disciplina de Sistemas de Informação Contábeis. Foi ministrado para os participantes um curso, onde foram apresentados os modelos alvos do estudo. Logo após a apresentação dos modelos, foram entregues alguns exercícios de compreensão, para que os estudantes se familiarizassem com as abordagens. As questões utilizadas nos exercícios foram similares às questões tratadas depois no experimento. Ao final do curso, foi explicado que os estudantes executariam um experimento semelhante ao que foi visto durante o curso.

A execução do experimento consistiu de uma atividade de compreensão, onde foram dados aos participantes, esquemas conceituais de processos de negócios no REA (Recurso-Evento-Agente) e ER para cada grupo (houve a divisão dos participantes em dois grupos). Acompanhado dos diagramas, foi entregue a cada participante um questionário, que continha questões a respeito do processo de negócio representado no diagrama. O objetivo dessa atividade foi obter dados a respeito do entendimento do usuário, que foi medido através das respostas obtidas. As respostas desse questionário foram analisadas por meio de duas métricas definidas por Poels et al. (2011): eficiência, obtida pelo tempo usado pelo o participante para responder o questionário; eficácia, relacionada com o número de acertos obtidos no questionário.

Os resultados obtidos no estudo mostraram que o uso do modelo REA foi superior ao modelo ER, indicando que, para processos de negócios, o REA pode ser mais recomendado. Contudo, pelo experimento ter sido aplicado em problemas pontuais (de conversão e troca) e sob determinada métrica (compreensão) os resultados obtidos, podem não ser aplicáveis para outros processos de negócios.

3.2.3 Investigação em modelos de dados conceituais

Em Patig (2004), a autora apresentou um conjunto de métricas para a avaliação da expressividade de modelos conceituais. Para isso, a mesma definiu seis formas de checar a expressividade em modelos conceituais, aplicando-as no poder de descrição do modelo e/ou da linguagem. As métricas definidas foram:

1. Conjunto de requisições descritas de forma diferente: um modelo será expressivo, caso a modelagem descrita no modelo seja exatamente conforme pedido;
2. Elementos do modelo: avalia se conseguem expressar diferentes domínios de forma descritiva;
3. O esquema: verifica se há a necessidade de adicionar novos construtores para aumentar a sua expressividade.

As próximas três métricas utilizadas por Patig (2004) apenas são aplicadas quando não é possível medir a expressividade com as métricas apresentadas anteriormente.

4. Duas descrições: d1 e d2 devem ter seus resultados comparados;
5. Modelos disjuntos: se todos os modelos de duas descrições, d1 e d2 forem disjuntos, a sua expressividade não poderá ser comparada, desde que as requisições das descrições se refiram a diferentes conjuntos de elementos;
6. A expressividade do modelo: essa é maior do que o propósito de certo domínio.

A principal contribuição de Patig (2004) é que a forma de avaliação proposta não se aplica apenas a modelos de dados conceituais, mas também pode ser usada em estudos sobre linguagens de descrição.

Hussain (2014) fez uma revisão nas métricas propostas por Moody, Shanks e Darke (1998) e Hussain e Awais (2008), para medir a completude e corretude de modelos conceituais. As métricas foram analisadas a partir da definição de parâmetros de qualidade, isto é, a qualidade dos resultados produzidos em cada fase do desenvolvimento de sistemas (AS-SENOVA; JOHANNESSON, 1996). Esse aspecto foi analisado a partir de três perspectivas: sintaxe, semântico e pragmático. Todavia, Hussain (2014) apenas utilizou o semântico na sua revisão, pois, esse possui um grau de correspondência entre o esquema e o problema do domínio, o que se adequa a modelos conceituais.

A revisão feita por Hussain (2014) em Moody, Shanks e Darke (1998) analisou seis métricas correspondentes a completude e corretude. Segundo Hussain (2014), das quatro métricas definidas para medir a completude de modelos conceituais, apenas uma foi definida adequadamente, que seria ‘Número de requisitos do usuário que não estão inclusos no modelo’, uma vez que o significado de completude seria que o modelo contém todas as afirmações sobre o domínio abordado (HUSSAIN, 2014). De acordo com Hussain (2014), as demais métricas definidas em Moody, Shanks e Darke (1998) para observar a completude de modelos conceituais foram estipuladas de modo errôneo, pois, duas dessas métricas deveriam ser aplicadas para medir a corretude ao invés da completude. Uma vez que estão diretamente relacionadas com a representação correta do domínio, sendo essas ‘Número de itens incluídos no modelo não correspondem às requisições do usuário’ e ‘Número de inconsistência do modelo’.

Por fim, para Hussain (2014), a última métrica definida em Moody, Shanks e Darke (1998) para avaliar a completude deveria ser descartada. As duas métricas restantes foram utilizadas por Moody, Shanks e Darke (1998) para analisar a corretude dos modelos conceituais e, segundo Hussain (2014), apenas a métrica ‘Número de violações na conversão de dados’ seria apropriada para medir a corretude de um modelo conceitual, pois, violações no modelo afetam a sua corretude. Já, ‘Número de violações das normas formais’, não se aplicaria na avaliação de modelos conceituais, pois as normas formais são discutidas apenas em modelos relacionais, ou seja, em um nível lógico.

O segundo estudo analisado na revisão foi o estudo de Hussain e Awais (2008), onde foram definidas quatro métricas para analisar a completude e corretude de modelos conceituais.

1. Índice de Completude (IC): média de dependências funcionais representadas pelo número total de dependências funcionais do domínio.
2. Índice de Completude Difusa (ICD): checa se as dependências funcionais são parcialmente representadas no modelo.
3. Completude baseada no Esforço (CbE): mede o esforço necessário para modelar um esquema ER completo.
4. Corretude Semântica (CS): calcula a média dos requisitos representados pelo total de requisitos solicitados pela definição do problema modelado.

Das métricas apresentadas em Hussain e Awais (2008), Hussain (2014) indica que apenas uma delas estaria incorreta, que é a CbE, pois essa não é clara o suficiente, podendo enviar os resultados.

A revisão realizada por Hussain (2014) levantou incoerências nas métricas definidas por Hussain e Awais (2008) e Moody, Shanks e Darke (1998), sugerindo modificações.

Além disso, Hussain (2014) proveu sugestões de combinações de métricas para definição do nível de qualidade de modelos conceituais.

Em Hammar (2016), foi apresentado o modelo de qualidade CODP (Padrões de Desenvolvimento de Ontologias), que consiste em um meta-modelo que provê de formas para analisar a qualidade conceitual de ontologias. O CODP é constituído por três conceitos principais:

- Contexto de uso: representa o contexto no qual o CODP é utilizado para construir a ontologia.
- Qualidade das características: denota aspectos particulares de qualidade do CODP, que podem ter maior ou menor importância, conforme o contexto que está sendo observado.
- Indicadores: propriedades individuais de medida do CODP, que podem ser afetadas pelo contexto usado, podendo aumentar ou diminuir algumas características de qualidade.

Os três conceitos apresentados acima são considerados na análise de qualidade feita a partir do CODP, que é dividida em quatro categorias:

1. Adequação funcional, divide-se em:

- Completude funcional: avalia se todo o conhecimento requerido se encontra no modelo conceitual;
- Adequação funcional: característica relacionada à como a ontologia pode facilitar o entendimento de como armazenar e recuperar as informações definidas para o sistema;
- Consistência: verifica se o modelo de dados possui construtores lógicos;
- Precisão: observa se o modelo de dados representado é realmente condizente com o domínio do mundo real.

2. Usabilidade, subdivide-se em seis categorias, e pode ser utilizada quando se quer obter informações a respeito de eficiência, satisfação e simplicidade.

- Adequabilidade: se refere a capacidade do usuário de reconhecer quando suas necessidades são atendidas;
- Entendimento: verifica se a estrutura do modelo é a certa para o aprendizado de novos usuários;
- Operabilidade: avalia se o modelo tratado possui os atributos necessários para tornar simples o seu uso e aplicação;

-
- Proteção a erros: verifica se o modelo previne erros no processo de modelagem;
 - Interface amigável: analisa se a documentação (textos, construtores, gráficos) é de fácil compreensão;
 - Acessibilidade: analisa se qualquer pessoa, independentemente do nível de conhecimento, é capaz de usar o modelo.
3. Manutenção: está relacionada com as ontologias criadas a partir do uso do CODP e como essas podem ser modificadas para se adaptarem a diferentes casos de uso. A manutenção tem cinco categorias:
- Modularidade: avalia se um componente tem impacto em outros;
 - Analisabilidade: avalia o impacto de mudanças na ontologia;
 - Modificabilidade: pode ser eficazmente utilizado para identificar problemas de degradação e defeitos na qualidade do CODP;
 - Testabilidade: estabelece se testes de efetividade e eficiência podem ser executados pelo CODP;
 - Estabilidade: analisa os fluxos de mudança do CODP. Uma alta estabilidade significa um baixo nível de mudança esperado, e vice-versa.
4. Compatibilidade divide-se em três categorias:
- Coexistência: CODP pode coexistir com outros CODP como módulos em uma ontologia, isto é, não há impacto negativo em outros módulos CODP;
 - Reusabilidade: um CODP pode ser usado em mais de um sistema ou ser aplicado na modelagem de novos sistemas;
 - Interoperabilidade: indica se dois ou mais CODP compartilham representações de conceitos.

Hammar (2016) apresentou um conjunto de características que deveriam ser consideradas no desenvolvimento de ontologias. O autor apontou que não há muitos estudos que avaliem a qualidade de ontologias no nível conceitual, sendo essa de extrema importância em outras fases do projeto de banco de dados. Assim, Hammar (2016) propôs um conjunto de métricas, para medir algumas características de qualidade de ontologias. Todavia, o estudo só abordou o levantamento das métricas e discutiu maneiras de utilizá-las, não sendo feito qualquer experimento ou estudo de caso que indiquem que se os resultados alcançados auxiliam a medir a qualidade de ontologias.

3.3 CONSIDERAÇÕES DO CAPÍTULO

Neste capítulo foram apresentados estudos na modelagem de dados NoSQL e na avaliação de modelos de dados conceituais. Na Seção 3.1, foram apresentadas propostas de modelos

de dados para sistemas de BD NoSQL. A partir, do levantamento feito na Seção, pode-se selecionar os modelos de dados conceituais que seriam aplicados no estudo proposto nesta dissertação. A escolha dos modelos observados no estudo se deve ao fato de ambos serem os únicos modelos de dados conceituais encontrados no levantamento realizado que oferecem a possibilidade de construção de esquemas de dados conceituais para mais de um sistema de BD NoSQL, são eles: o COMN e NoAM. Onde o NoAM e COMN podem ser utilizados para construir esquemas de dados para três dos quatro modelos de dados de sistemas de BD NoSQL. Os domínios abordados no experimento descrito nesta dissertação, também foram obtidos a partir do levantamento realizado na Seção 3.1, onde foi observado quais domínios foram mais abordados nos estudos correlatos, assim, obtendo-se o grau de relevância desses domínios para a área abordada, os domínios escolhidos foram aplicações de comércio eletrônico e na área da saúde.

Na Seção 3.2, foram expostos estudos relacionados com a investigação de modelos de dados conceituais, a partir dos estudos levantados nessa Seção, pode-se ter um melhor entendimento da metodologia aplicada para a investigação de modelos de dados conceituais. Dessa forma, auxiliando na construção da metodologia aplicada no estudo da dissertação. Além disso, foram observadas as métricas aplicadas nos estudos e a partir dessas foram selecionadas as três métricas aplicadas ao estudo descrito na dissertação, sendo essas: simplicidade, representatividade e qualidade conforme apresentadas na Seção 1.3. A escolha das métricas deve-se ao fato dessas serem as mais adequadas aos objetivos do trabalho e as questões de pesquisa levantadas. Ademais, muitas das propriedades levantadas nos trabalhos correlatos não se adequam para a avaliação de modelos de dados conceituais para sistemas de BD NoSQL, sendo esse o foco da pesquisa desenvolvida.

No próximo capítulo é descrito o plano experimental desenvolvido para a execução das atividades do experimento realizado, neste capítulo são abordadas as hipóteses levantadas, a construção dos questionários aplicados no experimento, a forma como cada atividade foi aplicada, as métricas observadas e como obter dados a respeito das mesmas. Além disso, neste capítulo o perfil dos participantes é descrito, as ameaças a validade do estudo, variáveis independentes e dependentes.

4 DESENHO EXPERIMENTAL PROPOSTO PARA A AVALIAÇÃO DE MODELOS DE DADOS DE SISTEMAS DE BD NOSQL

Para facilitar o entendimento dos procedimentos metodológicos usados no estudo, neste capítulo são discutidos os principais tópicos do planejamento da execução do experimento realizado, como: objetivos, metas, hipóteses, instrumentação, universo populacional e amostra. Na Seção 4.1, é descrito o desenho do plano utilizado. A Seção 4.2 apresenta a definição e o que se espera obter a partir da execução do experimento, enquanto que na Seção 4.3, são expostas as hipóteses levantadas para o estudo. Já na Seção 4.4 os tratamentos utilizados no estudo são expostos. A Seção 4.5 aborda os objetos experimentais utilizados no estudo e a Seção 4.6 descreve os sujeitos experimentais que participaram do estudo. A Seção 4.7 mostra as variáveis do estudo, enquanto a Seção 4.8 explica como os questionários utilizados na execução dos estudos descritos nesta dissertação, foram construídos. Por fim, na Seção 4.9 são descritas as ameaças à validade do estudo.

4.1 DESENHO DO PLANO

Esse estudo possui um fator (o quanto a escolha do modelo interfere na modelagem de um domínio e no seu entendimento) e dois tratamentos: o modelo COMN e NoAM.

O uso do quadrado latino é o mais recomendado para o estudo, pois esse envolve pessoas e atividades que se repetem. Essa repetição pode favorecer mais uma abordagem que a outra sem o uso do quadrado latino. Pois, uma vez em que o participante tenha trabalhado com Tratamento X e, se esse participante utilizar o mesmo domínio abordado do tratamento anterior para o Tratamento Y, esse participante já estará familiarizado com o domínio, podendo gerar melhores resultados para o Tratamento Y. A fim de tentar evitar isso e obter resultados mais confiáveis, o desenho do experimento por meio do quadrado latino foi escolhido e esse é explanado a seguir.

Tabela 1 – Desenho Experimental

Unidade Experimental 1	Unidade Experimental 2
Grupo 1 Tratamento X	Grupo 1 Tratamento Y
Grupo 2 Tratamento Y	Grupo 2 Tratamento X

Fonte: O autor (2019)

Os participantes selecionados para o estudo foram divididos em dois grupos como visto na Tabela 1. Cada grupo utilizou os dois tratamentos (COMN e NoAM) durante o decorrer do experimento. Contudo, o experimento foi dividido em duas unidades experimentais, essas foram: a representação de um domínio na área da saúde e uma aplicação de um co-

mércio eletrônico. Assim, evitando que os participantes se familiarizem com os domínios abordados no estudo. Cada unidade experimental divide-se da seguinte maneira: O grupo sorteado para utilizar o COMN primeiro, realizou as atividades da unidade experimental no modelo selecionado. Após o término dessa unidade experimental, esse grupo realizou as mesmas atividades realizadas na primeira unidade experimental, variando apenas o tratamento que foi o NoAM e a unidade experimental, que também foi nova (ver Seção 4.5). Os participantes foram separados em dois grupos: um remoto e um presencial. A alocação dos participantes para cada grupo se deu ao fato de que se este participante teria ou não disponibilidade para realizar as atividades nos horários definidos para a execução presencial. Dessa forma, os participantes que não puderam participar presencialmente foram alocados para o grupo remoto, onde foi agendado uma data para a execução das atividades. Dessa forma, mesmo remotamente, um responsável pela execução das atividades estaria disponível para esclarecer dúvidas destes participantes e aplicar as atividades na ordem planejada para o estudo.

Antes de executar as atividades com os participantes selecionados no estudo, foi executado um piloto. Esse piloto serviu para observar se atividades e os questionários construídos estavam claros. Assim, possíveis problemas apresentados durante a execução do piloto foram corrigidos, a fim de, durante a execução do experimento descrito nesta dissertação não ocorressem erros durante a execução das atividades ou problemas com os materiais utilizados no experimento que pudessem impactar nos resultados do estudo. O piloto foi executado com dois participantes, no qual esses participantes possuem um perfil semelhante aos participantes do estudo (Seção 4.6). Mais detalhes sobre a execução do experimento estão contidos no Apêndice A desta dissertação.

4.2 DEFINIÇÃO DOS OBJETIVOS DO EXPERIMENTO

Por meio do levantamento do estado da arte realizado na Seção 3.2, pode-se observar estudos empíricos que realizaram análises em modelos conceituais, assim, auxiliando a entender quais propriedades foram investigadas e o porquê de sua investigação. Dessa forma, ajudando no desenvolvimento de quais propriedades seriam analisadas no estudo descrito nesta dissertação.

Vale ressaltar que boa parte dos estudos encontrados analisaram modelos relacionais, dessa maneira, algumas das propriedades investigadas são relevantes apenas para esses modelos, por essa razão não se aplicando para modelos de dados conceituais voltados para BD NoSQL. Contudo, algumas das definições abordadas nos trabalhos correlatos foram utilizadas neste estudo, sendo essas: a simplicidade, representatividade e qualidade dos modelos.

1. A simplicidade pode ser observada a partir de duas métricas:

-
- usabilidade, foi definida a partir do quão fácil para os participantes do estudo foi utilizar o modelo;
 - compreensão, foi observado se o uso do esquema de dados modelado a partir do uso dos modelos estudados torna mais fácil a compreensão do domínio.
2. Representatividade, esta propriedade também foi observada a partir de duas propriedades, sendo essas:
- expressividade gráfica, observou se a notação disposta nos modelos é intuitiva;
 - poder de expressividade, se observa se os modelos conseguem expressar os domínios.
3. Qualidade dos modelos, foi observada por meio de três propriedades, estas são:
- corretude, nesta investiga-se o quão corretos os esquemas construídos nos modelos estudados estão com o que foi solicitado no domínio;
 - completude, se observa se os esquemas construídos possuem propriedades não necessárias para o entedimento do domínio abordado;
 - independência de *software*, foi observado se os modelos estudados são independentes de paradigma.

Dentre as possíveis características as descritas acima foram selecionadas, pois por meio da investigação dessas é possível ter uma avaliação completa de modelos de dados conceituais voltados para BD NoSQL. Dessa forma, o objetivo geral da execução do experimento é caracterizar como a escolha de um modelo conceitual influencia na modelagem e no entendimento de um domínio. Considerando esse objetivo, foram definidos os seguintes objetivos de medição e como se obteria dados para cada um desses objetivos no experimento:

- Simplicidade do modelo de dados: as observações a respeito desta característica foram obtidas por meio da aplicação de dois questionários empregados aos participantes do estudo (Apêndice C e D). O primeiro questionário observado foi totalmente voltado para a compreensão dos participantes, dessa forma sendo construídas questões relacionadas ao nível de compreensão desses participantes quanto aos modelos de dados investigados (ver Seção 4.8). Já no segundo questionário foi observado a usabilidade, e uma das propriedades de outra métrica. Em respeito a usabilidade foi observado o nível de satisfação dos participantes em relação aos modelos de dados conceituais. Ambos os questionários elaborados (Apêndice C e D) foram construídos utilizando a escala de Likert, essa escala foi escolhida por ser uma das melhores formas de observar o nível de concordância para determinada questão. O nível utilizado nas respostas de ambos os questionários foi o nível 5 (1- o nível mais baixo, 5 o mais alto) ;

- Representatividade do modelo de dados: os dados para esta métrica foram obtidos através do segundo questionário aplicado aos participantes (Apêndice D), neste foi indicado o quanto os construtores dispostos em cada modelo auxiliaram no entendimento do domínio segundo os participantes, sendo assim, observado as duas propriedades da métrica representatividade do modelo de dados;
- Qualidade dos modelos de dados: esta propriedade foi obtida à partir da validação dos esquemas construídos: essa validação foi feita por um par de especialistas, que avaliaram se os esquemas construídos condizem com o domínio modelado, sendo avaliadas três características corretude, completude e independência do paradigma. Essa validação contou com um questionário, este auxiliou os especialistas a observar se as propriedades levantadas para essa métrica foram ou não atendidas nos esquemas de dados construídos. O questionário contou com 10 questões objetivas, nestas foram usadas a escala Likert em 3 níveis (3- Máximo, 2-Parcial, 1- Mínimo).

Assim, no estudo, espera-se:

Analisar modelos de dados conceituais voltados para BD NoSQL, com o propósito de caracterizar o impacto que a adoção desses tem ao que diz respeito a compreensão, usabilidade, expressividade e qualidade dos esquemas construídos.

A partir da realização do estudo, espera-se obter dados para responder ao seguinte questionamento: quais as consequências do uso dos modelos COMN ou NoAM no que se refere à representatividade dos dados e na compreensão e usabilidade destes modelos?

4.3 DEFINIÇÃO DAS HIPÓTESES

Cada uma das métricas consideradas possui duas variações (COMN e NoAM) que correspondem aos modelos conceituais utilizados no experimento. A nomenclatura definida a seguir será adotada neste capítulo para se referir às métricas.

Compreensão do Modelo de Dados - CMD

Usabilidade do Modelo de Dados - UMD

Representatividade do Modelo de Dados - RMD

Qualidade dos Modelos de Dados - QMD

A hipótese principal deste estudo é a nula, onde não há diferença significativa ou essa diferença é pequena nos resultados coletados para os modelos analisados. Dessa forma, o teste de Wilcoxon foi escolhido para investigar se a hipótese nula seria ou não aceita nas métricas levantadas no estudo. O teste de Wilcoxon é um teste de hipótese não paramétrico (CUZICK, 1985).

Hipótese nula: na hipótese nula cada métrica foi analisada individualmente e essa possui duas variações como dito anteriormente, sendo essas variações: o NoAM e o COMN.

Dessa forma, a compreensão do modelo de dados foi analisada através dos dados do primeiro questionário aplicado aos participantes durante a execução do experimento (formulação dos questionários na Seção 4.4). Os dados obtidos por meio deste questionário foram aplicados à hipótese nula da seguinte maneira:

$$H_0: \text{CMDCOMN}_p = \text{CMDNoAM}_p$$

A análise foi realizada da seguinte forma: Os resultados obtidos do questionário entregue aos participantes para o modelo COMN foram comparados com os resultados obtidos do mesmo questionário sendo esse para o modelo NoAM, onde p representa os participantes (ver na Seção 4.6 e 4.8, para informações sobre os sujeitos experimentais e o desenho do plano).

A segunda métrica observada foi a UMD, sendo essa observada a partir dos dados obtidos do segundo questionário aplicado aos participantes (Apêndice D), onde esse questionário possui questionamentos sobre a usabilidade e expressividade dos modelos (sendo essa relacionada a métrica de representatividade). Assim, cada questionamento observado foi aplicado na sua devida métrica. Dessa forma, a usabilidade dos modelos de dados foi analisada da seguinte maneira:

$$H_0: \text{UMDCOMN}_p = \text{UMDNoAM}_p$$

Onde UMD é a métrica, p representa os participantes, e a comparação foi feita entre os modelos NoAM e COMN observando se a hipótese nula foi aceita na métrica.

Enquanto, a análise da expressividade dos modelos de dados escolhidos no estudo foi efetuada como segue:

$$H_0: \text{RMDCOMN}_p = \text{RMDNoAM}_p$$

Por fim, a última métrica analisada foi a QMD, essa foi obtida a partir dos resultados da avaliação dos esquemas construídos. Nessa métrica foram observadas três características - corretude, completude e independência de *software*. Essa validação foi efetuada por dois especialistas, onde cada especialista avaliou de forma individual os esquemas construídos pelos estudantes. Assim, foi dado aos especialistas, um questionário, contendo questões sobre as propriedades observadas. O intuito desse questionário foi investigar se os esquemas construídos pelos participantes possuíam as três características analisadas (mais informações sobre esse questionário podem ser encontradas na Seção 4.4 e no Apêndice E). E se não havia diferença na QMD quando observados os modelos NoAM e COMN. A métrica QMD foi observada da seguinte forma:

$$H_0: \text{QMDCOMNES}_p = \text{QMDNoAMES}_p$$

Nessa hipótese, ES representa os dois especialistas e p representa os estudantes. Nessa métrica foram avaliados se houve diferença ou não nos esquemas de dados que foram construídos (nos modelos de dados NoAM e COMN) pelos participantes nas propriedades investigadas: corretude, completude e independência de paradigma. Contudo, antes de investigar os resultados dessa métrica, o teste de *Kappa* foi aplicado nas respostas obtidas, tornando possível saber se a concordância entre os especialistas foi meramente aleatória ou não. O teste de *Kappa* é utilizado com a finalidade de identificar se um determinado número de pessoas concorda ou não a respeito de determinado assunto. Assim, quanto mais próximo de zero for o valor obtido, significa que a concordância entre os especialistas foi aleatória, ou seja, não houve concordância entre as duas avaliações necessitando de um terceiro especialista para avaliar os esquemas construídos. Após a realização do teste de *Kappa*, a mediana apresentada no questionário aplicado aos especialistas (Apêndice E) foi analisada. E por fim, executou-se o teste de Wilcoxon, para verificar se a hipótese foi significativa ao nível nominal de 5%.

A principal premissa da hipótese nula é que a escolha do modelo de dados não impacta nas características observadas, ou seja, qualquer modelo conceitual obterá os mesmos resultados. Contudo, este estudo tenta rejeitar essa hipótese e demonstrar que há sim diferença nos resultados das características analisadas, quando aplicados os modelos conceituais investigados neste trabalho. Assim, foi definida uma hipótese alternativa, caso a hipótese nula seja refutada.

Hipótese alternativa: caso algumas das métricas apresentem resultados diferentes dos esperados na hipótese nula, os dados devem ser analisados na hipótese alternativa, onde cada propriedade observada possui a sua métrica.

A CMD deve ser analisada da seguinte maneira:

$$H1: \text{CMDCOMN}_p \neq \text{CMDNoAM}_p$$

A análise da UMD deve ser observada da seguinte forma:

$$H1: \text{UMDCOMN}_p \neq \text{UMDNoAM}_p$$

A análise da RMD deve ser observada da seguinte forma:

$$H1: \text{RMDCOMN}_p \neq \text{RMDNoAM}_p$$

Por fim, caso a QMD seja refutada na hipótese nula, essa deve ser aplicada na seguinte hipótese:

$$H1: \text{QMDCOMNES}_p \neq \text{QMDNoAMES}_p$$

Assim, as métricas que forem refutadas na hipótese nula devem ser aplicadas à hipótese alternativa.

4.4 TRATAMENTOS

Os tratamentos utilizados neste estudo foram dois modelos de dados conceituais, a escolha dos mesmos se deve ao fato de que no levantamento do estado da arte esses modelos foram os únicos capazes de serem aplicados a mais de uma sub-categoria de BD NoSQL, esses modelos foram:

- O COMN, proposto por Hills (2016), pode ser usado na modelagem de sistemas NoSQL e relacionais;
- NoAM, proposto por Bugiotti et al. (2014) para a modelagem de aplicações NoSQL contendo dados agregados.

Assim, esse estudo experimental possui um fator que corresponde ao modelo conceitual a ser aplicado como tratamento, ou seja, o fator possui dois níveis (COMN e NoAM).

4.5 UNIDADES EXPERIMENTAIS

No estudo, foram utilizados dois domínios: Comércio eletrônico e área médica. Esses foram selecionados, pois foram os mais discutidos no estado da arte apresentado na Seção 3.1 deste trabalho. O uso de dois domínios, um para cada unidade experimental, se deve ao fato do desenho do estudo (Seção 4.1).

A ordem com que cada um dos domínios foi aplicado determinou-se a partir de sorteio, sendo sorteado a unidade experimental e o tratamento a ser aplicado primeiro por cada grupo.

4.6 SUJEITOS EXPERIMENTAIS

Os participantes desse estudo foram doze pessoas, sendo 10 destas alunos da área de TI (Ciência da Computação e Sistemas de Informação), cinco desses alunos do programa de pós graduação e os cinco restantes, alunos de graduação recém formados. A faixa etária dos estudantes ficou em 25,5 anos na mediana e 27,3 anos na média das idades, no qual a maior idade foi de 36 anos e a menor idade 22. Os estudantes selecionados foram seis mulheres e quatro homens. Para participar das atividades planejadas no desenho do estudo, esses participantes deviam cumprir três requisitos:

- Alunos de Banco de Dados;
- Ter experiência em modelagem conceitual;
- Possuir experiência na construção de esquemas conceituais de dados.

Essas três premissas foram elaboradas para que os participantes se sentissem mais confortáveis na hora de realizar as tarefas pedidas na execução do experimento. Estas

premissas também evitam que os achados obtidos fossem afetados pela inexperiência do participante na área em que estudo foi desenvolvido. A experiência mínima apontada pelos participantes em modelagem conceitual foi de um ano e a máxima de cinco anos. Dentre os participantes, seis costumam utilizar algum modelo de dados na modelagem conceitual de dados com regularidade.

O perfil dos especialistas que participaram do estudo foram: um homem e uma mulher, sendo a idade desses 26 e 30 anos. Um desses especialistas é professor na área de banco de dados há cinco anos e tem uma experiência anterior na área de dois anos, totalizando sete anos de experiência. Já o outro especialista, possui cinco anos de experiência na área, no qual este participou de projetos relacionados com modelagem de dados conceituais e atualmente desenvolve pesquisas nessa área.

4.7 VARIÁVEIS

Independentes

- Os domínios escolhidos;
- Modelos conceituais utilizados;
- Os grupos selecionados.

Dependentes As características observadas no estudo são: a compreensão do usuário a respeito dos esquemas modelados no NoAM e COMN, a usabilidade e expressividade dos modelos conceituais COMN e NoAM. Além disso, foi observada a validação dos esquemas construídos no estudo.

4.8 FORMULAÇÃO DOS QUESTIONÁRIOS

Como forma de obter dados para serem avaliados no estudo, foi elaborada uma série de formulários, que possui questões específicas diretamente conectadas com as variáveis selecionadas do estudo. Segundo Saris e Gallhofer (2014), essa conexão entre uma variável que se quer observar e a elaboração de uma questão que colete informações sobre essa variável é uma das melhores formas de se realizar estudos descritivos.

Cada um dos questionários foi aplicado em um momento distinto. O primeiro questionário dado aos participantes diz respeito à compreensão dos mesmos sobre o domínio abordado, este questionário está presente no Apêndice C desta dissertação. A partir das respostas desse questionário, a CMD foi observada. Com o levantamento feito nos estudos de Genero, Poels e Piattini (2002), Poels (2003) e Poels et al. (2011), que analisaram a compreensão de usuários sobre modelos conceituais, foi usada a forma como esses estudos definiram a compreensão e a observaram.

Assim, considera-se que a compreensão de modelos conceituais está relacionada com a facilidade com a qual os conceitos e as estruturas do modelo são entendidos pelo usuário. O questionário aplicado possui sete questões elaboradas a partir desta definição de compreensão, no qual o propósito deste questionário foi obter dados que auxiliem na análise da compreensão dos modelos COMN e NoAM. As questões presentes nesse questionário são de acordo com o nível de compreensão do usuário dos modelos estudados, e se esses ajudam na compreensão do domínio representado. A forma como esse questionário foi aplicado se baseia no experimento reportado em Genero, Poels e Piattini (2002), onde um questionário foi dado junto com um esquema de dados construído a partir do modelo conceitual investigado. A execução da atividade realizada para esse questionário está descrita no Apêndice A da dissertação.

O uso do segundo questionário foi pensado da seguinte maneira: dado um problema de um domínio, o participante modelaria esse problema e, após a conclusão da atividade, seria dado um questionário a respeito da experiência de uso que o participante tenha tido com o modelo conceitual investigado. As questões abordadas nesse questionário dizem respeito a usabilidade do modelo de dados, a métrica UMD e a expressividade de cada um dos modelos estudados, característica relacionada com a métrica RMD.

O questionário divide-se em perguntas voltadas para a usabilidade do modelo e sua expressividade. São tratadas questões sobre a usabilidade do modelo conceitual, como: o modelo possui uma estrutura que ajude novatos no seu uso? Os construtores definidos são claros? Houve dificuldade para saber qual construtor usar? Entre outros questionamentos. A parte do questionário que aborda a expressividade dos modelos conceituais foi medida sob a ótica do desenvolvedor, baseando-se no estudo de Cherfi, Akoka e Comyn-Wattiau (2002), que analisaram oito diagramas construídos a partir de dois modelos conceituais, onde as questões foram voltadas para a expressividade dos modelos observados.

A respeito das questões do questionário descrito nesta dissertação, voltadas para a representatividade dos modelos foram abordadas questões como a capacidade do modelo conceitual representar mais de um domínio e a não necessidade de adição de novos construtores para melhorar a sua representatividade, assim como no trabalho de Patig (2004), onde a autora utilizou as duas características citadas para medir a expressividade em modelos conceituais (Seção 3.2.3).

Os dois questionários, até então explicados, tiveram suas questões elaboradas a partir das definições de cada propriedade selecionada dos estudos aqui descritos. A construção dos questionamentos aplicados, foi feita com o intuito de obter informações à respeito de três métricas definidas no estudo a CMD, UMD e a RMD. Dessa forma, foi investigado se os modelos de dados COMN e NoAM apresentaram resultados diferentes quando observadas as três métricas. Esses dois questionários podem ser vistos na íntegra nos Apêndices C e D desta dissertação.

Por fim, o terceiro e último questionário elaborado foi aplicado com dois especialistas

da área de banco de dados, onde esse questionário foi entregue junto com os esquemas construídos pelos participantes. Para cada esquema construído, os especialistas preencheram o questionário, avaliando os esquemas modelados pelos participantes. Nesse questionário foi analisada a QMD, a partir da qual são observadas três características: a completude, corretude e a independência de *software*.

A elaboração das perguntas feitas nesse questionário foi obtida através do estado da arte apresentado nesta dissertação. Onde, a partir da definição de Hussain (2014) sobre como medir a completude (sendo que essa investigou se toda a informação necessária para o entendimento do domínio está contida na representação desenvolvida em um esquema), foram elaboradas questões que abordassem a completude dos esquemas construídos, como: se todas as propriedades necessárias relevantes para o domínio estão representadas, se alguma propriedade relevante foi esquecida, se há alguma propriedade definida de forma redundante, e se o diagrama desenvolvido está de acordo com o que foi pedido para ser feito. Essas foram as perguntas voltadas para a completude dos esquemas construídos.

Foi extraída dos estudos de Moody (2003) e Hussain (2014) a forma como eles definem a corretude de um modelo de dados, sendo a corretude definida como a representação correta do domínio. Assim, as perguntas elaboradas a partir dessa premissa foram levantadas com o intuito de obter informação sobre a corretude dos esquemas construídos, como: se os construtores estavam sendo corretamente utilizados, se os relacionamentos foram definidos corretamente e se alguma regra foi definida de forma errônea. A última característica observada nesse questionário entregue aos especialistas foi a independência do *software*. Essa foi levantada a partir da definição de Elmasri e Navathe (2010), sobre o que é um modelo independente de *software*. Um modelo independente de *software* é aquele que não especifica detalhes de armazenamento de uma determinada plataforma ou sistema. Assim, duas perguntas foram construídas a fim de se saber se os esquemas construídos são independentes de software: se esses são independentes de ferramenta de implementação e se há alguma característica presente no diagrama que indica algum paradigma de armazenamento.

O questionário aplicado aos especialistas pode ser visto na íntegra no Apêndice E desta dissertação, onde também há uma explicação detalhada de como foi executado o experimento no Apêndice A.

4.9 AMEAÇAS A VALIDADE

Nesta seção, são apresentadas as ameaças à validade deste estudo. Essas ameaças se dividem em quatro: **validade de conclusão**, relaciona-se com problemas para se obter a conclusão correta sobre os relacionamentos entre o tratamento e os resultados do experimento (WOHLIN et al., 2012); **validade interna**, de acordo com Wohlin et al. (2012), uma ameaça à validade interna de um experimento ocorre quando os parâmetros independentes afetam a casualidade do experimento sem o conhecimento do pesquisador, portanto,

podendo ser uma ameaça à conclusão do estudo; **validade de construção**, diz respeito a generalização dos resultados obtidos em um experimento para uma teoria ou conceito (WOHLIN et al., 2012); **validade externa** são condições que limitam a generalização dos resultados de um experimento para a prática. Os três tipos de interações com o tratamento que podem afetar a generalização são: pessoas, lugar e tempo (WOHLIN et al., 2012). Sabendo-se disso, cada uma das possíveis ameaças à validade deste estudo é discutida a seguir.

4.9.1 Validade de conclusão

A fim de evitar uma ameaça na validade de conclusão, um responsável pelo estudo esteve sempre presente durante a execução do experimento, com o intuito de garantir que os tratamentos fossem aplicados como planejado no desenho experimental (Apêndice A). A presença do responsável também é necessária para sanar possíveis dúvidas que os participantes possam ter durante a execução de alguma das atividades, evitando que os participantes cometam erros que possam prejudicar os dados obtidos.

Com o propósito de evitar conclusões equivocadas a respeito dos dados obtidos foram executados testes nas hipóteses levantadas no estudo (Seção 4.3). O teste escolhido no estudo foi um teste não paramétrico, pois, esse tipo de teste é mais adequado para dados ordinais e nominais. Dessa forma, como os dados obtidos no estudo foram ordinais sendo esses obtidos a partir de uma escala de maior a menor intensidade, um teste não paramétrico é o mais recomendado para o estudo. Além disso, testes não paramétricos também são recomendados quando estão sendo observadas amostras muito pequenas, sendo o caso do estudo.

O teste não paramétrico selecionado no estudo foi o Teste de Wilcoxon, a escolha do mesmo foi devido a esse teste poder observar se duas amostras diferem ou não, sendo esse o objetivo das hipóteses levantadas no estudo.

4.9.2 Validade interna

Uma ameaça a validade interna seria se os resultados do grupo remoto fossem afetados por não haver uma pessoa responsável fisicamente presente. Assim, podendo enfraquecer a causa e o efeito que os modelos investigados tem sobre as métricas investigadas. Para evitar que isso viesse a ocorrer, um responsável pela execução das atividades do estudo marcou com cada participante do grupo remoto um horário em que este acompanhou o passo a passo da execução das atividades apenas entregando os materiais com a finalização das atividades.

4.9.3 Validade de construção

Para evitar problemas nos materiais produzidos no estudo foi executado um piloto, neste piloto foi observado se o material estava de acordo com o planejado. A partir deste piloto pode-se melhorar a execução das atividades que seriam realizadas durante o estudo. Além do piloto, as atividades foram realizadas primeiro com o grupo que executou-as presencialmente. A escolha da execução das atividades com esse grupo se deu ao fato de que o responsável pela execução das atividades presente, poderia observar quais dúvidas que os participantes desse grupo apresentaram. Dessa forma, podendo melhorar o processo das atividades para o grupo remoto, a fim de sanar os problemas apresentados ao decorrer da execução das atividades do grupo presencial não se repetissem no grupo remoto. Uma vez que o responsável não estaria presente fisicamente para esclarecer dúvidas, apenas virtualmente. No grupo remoto como o presencial as atividades foram passadas aos participantes a medida em que estes finalizavam cada atividade, evitou-se entregar todo o material aos participantes para que estes não realizassem as atividades de maneira inadequada do qual foi planejado.

4.9.4 Validade externa

A ameaça a validade externa foram os domínios utilizados no estudo. Pois, mesmo esses domínios sendo relevantes na área. A modelagem desses não surgiu de um ambiente real, onde os participantes mesmo possuindo experiência na área são estudantes. O que vem a fugir de um ambiente mais realístico (que seria um ambiente real com profissionais para realizar as atividades).

5 RESULTADOS

Neste capítulo, são apresentados os resultados obtidos da análise dos materiais aplicados no estudo descrito nesta dissertação.

5.1 ANÁLISE DOS DADOS OBTIDOS DO QUESTIONÁRIO QUE OBSERVOU A COMPREENSÃO DOS PARTICIPANTES

O primeiro questionário analisado foi o aplicado na primeira atividade executada no experimento. O questionário foi dado junto com o esquema de dados construído usando os modelos COMN e NoAM. Nesse questionário, foi usada a Escala de *Likert*, onde os participantes puderam expressar sua opinião em cinco níveis. Além das cinco questões presentes nesse questionário, há mais duas questões abertas que deveriam ser respondidas, caso o participante expressasse um nível de opinião específico, sendo pedido a esse que explicasse seu ponto de vista.

Os dados obtidos desse questionário foram analisados de três formas: a partir do grupo que realizou as atividades remotamente, do grupo que participou do estudo presencialmente e por fim, foi realizado o teste de hipótese para saber se foi apresentando diferença entre utilizar o NoAM ou COMN na CMD. Dessa forma, nas Seções 5.1.1 e 5.1.2 são apresentadas as medianas encontradas nos grupos para as questões relacionadas a compreensão, nestas seções são feitas análises de qual dos modelos investigados teve melhores resultados em cada grupo. Finalmente na Seção 5.1.3 foi descrito o teste de hipótese da CMD, assim, nesta seção se observa se a hipótese nula foi aceita ou não.

5.1.1 Resultados do grupo presencial ao questionário de compreensão dos modelos de dados

As respostas dos participantes foram agrupadas por intensidade e as medianas encontradas para cada questão é exposta nesta seção. Os resultados encontrados na primeira questão, sendo esta *O modelo de dados utilizado ajudou a compreender o domínio?*, mostraram que de acordo com a mediana obtida pelo os resultados do grupo presencial o modelo COMN obteve uma mediana maior no que diz respeito à compreensão do domínio apresentado nos modelos estudados, como pode ser visto na segunda coluna da Tabela 2.

A próxima mediana exposta na Tabela 2, esta relacionada com a seguinte questão: *O diagrama entregue é claro quanto à organização de dados por ele modelado?*. O modelo COMN teve um mediana maior que a apresentada no modelo NoAM, apresentando uma diferença de 0,5 entre as medianas o que indica que as propriedades do modelo auxiliaram os participantes deste grupo a entender melhor a organização dos dados representados no diagrama COMN.

Para a questão *O esquema de dados produzido torna mais simples a visualização do domínio estudado?* foi feita para investigar se os modelos estudados produzem esquemas de dados que auxiliam na compreensão dos usuários sobre o domínio por eles representados. De acordo com a mediana obtida nesse grupo, pode-se observar que o NoAM é inferior ao COMN em tornar a visualização do domínio mais simples, apresentando uma mediana inferior aos outros questionamentos já explanados, como pode ser visto na Tabela 2 quarta coluna, onde a diferença entre as medianas é 1,5. Essa diferença indica que, para esse grupo, o uso do COMN é mais recomendando quando se quer tornar mais fácil a visualização dos dados de um domínio.

Quando indagados se *Os construtores do modelo de dados são intuitivos?*, o COMN mais uma vez foi melhor avaliado por esse grupo, apresentando uma diferença de 0,5 ponto a mais para o modelo, como pode ser observado na Tabela 2.

A última questão observada foi: *A aplicação dos construtores de modelagem do modelo auxilia na visualização de como os objetos do domínio se relacionam?*. O COMN mostrou um melhor resultado para essa questão, sendo essa diferença vista na última coluna da Tabela 2, onde a diferença apresentada é 0,5 ponto para a mediana obtida em comparação com o NoAM. Esses resultados mostram que, de acordo com os participantes deste grupo, os construtores de modelagem de dados do COMN auxiliam mais que os construtores do NoAM, quando se quer ter um maior entendimento de como estão organizados os objetos do domínio representado e como esses objetos relacionam-se.

Tabela 2 – Medianas do COMN e NoAM para o Grupo Presencial em Relação ao Questionário de Compreensão.

Modelos	Compreensão	Claro	Simples	Intuitivo	Relacionamentos entre objetos
NoAM	4	4	3,5	4	4
COMN	4,5	4,5	5	4,5	4,5

Fonte: O autor (2019)

De acordo com a análise efetuada para este grupo, pode-se dizer que, em relação ao nível de compreensão dos usuários, o COMN obteve melhores avaliações quando observadas as características discutidas.

5.1.2 Resultados do grupo remoto ao questionário de compreensão dos modelos de dados

O questionário aplicado ao grupo que executou as atividades remotamente teve duas de suas questões com resultados iguais na mediana, sendo estas: *O esquema de dados produzido torna mais simples a visualização do domínio estudado?*; *Os construtores do modelo de dados são intuitivos?*. Em ambos questionamentos os modelos (NoAM e COMN) obtiveram o valor de mediana igual a 4 na escala de *Likert*, como visto nas primeiras colunas

da Tabela 3. Contudo, mesmo os dois modelos tendo apresentado o mesmo valor, dois participantes questionaram o modelo NoAM dizendo que o esquema produzido por ele ajudou pouco na visualização do domínio. Por essa razão, os dois participantes responderam à questão opcional, e explicaram o motivo do modelo não ter auxiliado na visualização do problema.

Tabela 3 – Medianas do COMN e NoAM para o Grupo Remoto em Relação ao Questionário de Compreensão.

Modelos	Simple	Intuitivo	Compreensão	Claro	Relacionamentos entre objetos
NoAM	4	4	4	4	4
COMN	4	4	4,5	4,5	4,5

Fonte: O autor (2019)

A pergunta que estes responderam foi a seguinte: *Se a resposta anterior for pouco ou de forma alguma, o que você acha que deveria ser alterado no modelo, para tornar os esquemas de dados modelados por ele mais claros?*. Um dos participantes indagou sobre a forma de representar os relacionamentos no NoAM, expressando o seguinte pensamento: "Modelos que exibem os relacionamentos de forma mais gráfica são mais claros". O segundo participante também comentou sobre a forma escolhida pelo NoAM para representar relacionamentos, afirmando: "Linhas de conexões de relacionamentos poderiam tornar mais clara a leitura, não havendo necessidade de ler tudo para entender o todo. Imagens valem mais do que palavras". Como pôde ser observado, ambos participantes acharam que a forma escolhida pelo modelo para representar relacionamentos entre objetos não é clara o suficiente, impactando na visualização do domínio, tornando-a mais difícil.

Nos demais questionamentos observados do grupo remoto, o modelo COMN foi melhor avaliado em comparação com o NoAM, sendo a diferença entre eles de 0,5 ponto na mediana para cada questão como pode ser visto nas colunas restantes da Tabela 3.

Vale ressaltar, mesmo dois dos questionamentos investigados terem resultados iguais na mediana, o COMN obteve melhores resultados nos demais questionamentos, o que sugere que para o grupo analisado, o COMN auxiliou mais que o NoAM na compreensão dos domínios representados no estudo. Contudo, esses resultados apenas se referem a esse grupo, para uma investigação mais profunda seria necessário aumentar o tamanho dos participantes.

5.1.3 Teste de hipótese CMD

Após a avaliação individual dos grupos, as respostas dadas pelos participantes foram agrupadas, e sob esses dados foi executado o teste de Wilcoxon, analisando se houve diferença ou não entre os modelos investigados em relação a CMD. O resultado apresentado em Wilcoxon mostrou que para os dados observados a hipótese nula levantada foi rejeitada

ao nível de 5% de significância. Assim, houve diferença entre utilizar o NoAM e o COMN quando se quer observar a CMD.

E a hipótese alternativa foi aceita **H01: $CMD_{COMN} \neq CMD_{NoAM}$** , mostrando que há diferença entre usar o COMN ou NoAM quando está sendo observada a compreensão.

5.2 ANÁLISE DOS DADOS OBTIDOS A PARTIR DO QUESTIONÁRIO APLICADO À USABILIDADE E EXPRESSIVIDADE DOS MODELOS

Nesta seção, os dados observados foram obtidos do questionário de usabilidade e expressividade dos modelos de dados, esse foi aplicado aos participantes após a conclusão da modelagem de novas propriedades para o esquema de dados visto. Posteriormente, esses esquemas foram analisados por dois especialistas e os resultados da análise feita por esses especialistas podem ser vistos na Seção 5.3 desta dissertação. O questionário aqui investigado possui dez questões nas quais seis delas são voltadas para a RMD, enquanto as quatro restantes são voltadas para a UMD.

Esse questionário também foi elaborado com base na escala de *Likert* e o nível de escala utilizado nos questionamentos feitos foi o nível cinco. O questionário também possui questões abertas opcionais a serem respondidas, caso o participante expresse-se certo nível em sua resposta nas questões objetivas.

A análise dos resultados obtidos foi feita da mesma forma como o questionário a respeito da compreensão dos participantes, ou seja, primeiro, os resultados dos grupos foram investigados separadamente e, após isso, houve o teste de hipótese para as métricas investigadas, sendo essa a UMD e a RMD.

5.2.1 Resultados do grupo presencial ao questionário de usabilidade e expressividade dos modelos de dados

Cada questionamento foi analisado separadamente e as medianas obtidas nos modelos foram comparadas. Todos os questionamentos investigados apresentaram resultados diferentes na mediana.

5.2.1.1 Investigação da Usabilidade

A usabilidade foi investigada a partir de quatro questões, os resultados das medianas encontradas para o grupo presencial são apresentados na Tabela 4. Como pode ser visto nesta tabela os participantes do grupo presencial relataram uma experiência de uso mais favorável ao utilizar o modelo COMN. Em relação a facilidade de uso a mediana encontrada para esse grupo foi a máxima definida na escala quando observado o COMN, isso indica que os participantes deste grupo acreditam que dentre os modelos investigados o COMN foi mais fácil de ser utilizado.

Para o questionamento a respeito da dificuldade no uso, visto também na Tabela 4, quanto menor o valor da mediana obtida nos dados investigados, melhor o modelo foi avaliado. Dessa forma, como pode ser visto na Tabela 4 o COMN apresentou um melhor resultado, pois, esse obteve uma mediana inferior ao NoAM. Assim, expressando que os participantes desse grupo tiveram menos dificuldade ao utilizá-lo, a diferença entre os resultados do COMN e NoAM foi de 1 ponto na mediana.

Três dos participantes desse grupo expressaram alguma dificuldade ao utilizar o NoAM. Essas dificuldades são expressas nas seguintes respostas:

- *"Algumas dificuldades em representar os relacionamentos";*
- *"Em realizar o relacionamento entre os objetos" e*
- *"Fiquei na dúvida como relacionar os objetos para determinado relacionamento."*

Isso indica que, talvez a escolha utilizada para expressar os relacionamentos no NoAM não seja a ideal, uma vez que a forma escolhida tornou o uso do modelo difícil pelo reporte feito pelos três participantes. Em relação a última questão a respeito da usabilidade, sendo esta relacionada a um possível uso dos modelos investigados, foi evidenciado um maior desejo dos participantes deste grupo de utilizarem o COMN novamente como pode ser visto na Tabela 4. Assim, indicando que para o grupo a experiência de uso foi melhor utilizando o COMN, uma vez que esses mostram uma inclinação maior em utilizar novamente o modelo.

Tabela 4 – Resultados de usabilidade do grupo presencial.

Modelos	Experiência de uso	Facilidade no uso	Dificuldade no uso	Possibilidade de uso futuro
NoAM	3,5	3	3,5	3
COMN	4	5	2,5	4

Fonte: O autor (2019)

5.2.1.2 Investigação da Representatividade

A primeira questão investigada foi *houve alguma dificuldade para saber qual construtor do modelo seria mais adequado para uma situação?*. Para essa questão quanto maior a mediana encontrada maior foi a dificuldade em escolher algum dos construtores do modelo para representar algo. O NoAM apresentou uma mediana maior como pode ser visto na Tabela 5 na coluna *escolha*, mostrando que para os participantes, houve uma maior dificuldade de compreensão dos construtores do modelo, dificuldade essa que pode ser vista na Seção 5.1.1, onde o NoAM obteve medianas inferiores ao COMN quando analisada a compreensão dos participantes. Consequentemente, essa falta de compreensão

prejudicou o entendimento do modelo NoAM de acordo com os participantes. Enquanto o COMN apresentou um melhor resultado, o que indica que, segundo o grupo investigado foi mais fácil identificar qual construtor selecionar quando o modelo foi COMN.

Para a pergunta feita anteriormente, havia uma questão opcional, caso o participante expressasse algum nível de dificuldade. Dentre os participantes, dois apresentaram muita dificuldade ao usar o NoAM. Assim, perguntou-se qual foi o problema encontrado. As respostas foram as seguintes:

- *"As figuras são pouco representativas e a forma que os objetos estão relacionados não facilita em saber qual construtor utilizar"* e
- *"Ao representar os relacionamentos surge uma dúvida ao escolher"*.

De acordo com essas respostas foi demonstrado que os construtores do modelo não são claros o suficiente, o que dificulta a escolha de qual melhor construtor utilizar em cada situação, o que está ligado ao poder de expressividade do modelo.

O próximo questionamento feito foi: *Houve dificuldade para a escolha de um ou outro construtor do modelo para representar algum aspecto do domínio escolhido?* O COMN apresentou um resultado inferior ao NoAM, como pode ser visto na Tabela 5, esses resultados mostram que, do ponto de vista desse grupo houve uma menor dificuldade para identificar qual construtor escolher para representar determinado relacionamento ou objeto. Segundo os resultados da Tabela 5, os participantes expressaram menos dificuldade ao utilizar o COMN. Isto indica que seus construtores são mais intuitivos, o que impacta na sua representatividade. Segundo um dos participantes, a principal dificuldade ao utilizar o NoAM foi ser capaz de diferenciar algumas propriedades utilizadas em seus construtores, como pode ser visto em sua resposta

- *"Em identificar o tipo de estrutura de dados que está sendo utilizada, por exemplo, se representa uma Array ou Objeto. No modelo dá para identificar o que é cada um, mas a forma de representar aparentemente é a mesma, deixando uma sensação de dúvida."*

Para os próximos questionamentos quanto maior a mediana obtida melhor avaliado os modelos investigados foram, o COMN apresentou melhores resultados nas demais questões analisadas como pode ser observado na Tabela 5. Na coluna da Tabela 5, a mediana obtida foi expressiva, sendo a diferença de 1,5 na mediana de um modelo para o outro. A questão da qual essa mediana foi apresentada refere-se: *os construtores do modelo selecionado conseguem expressar todas as características necessárias do domínio abordado?* O NoAM apresentou um resultado menor ao COMN. Assim, para esse grupo, o COMN consegue expressar de forma mais completa o domínio que foi abordado no estudo.

Quando indagados, sobre o porquê desse modelo (NoAM) não representar o domínio estudado ou outro modelo representar melhor, os participantes responderam da seguinte forma:

- *"Eu acho que a maneira de representar o domínio poderia ser melhorado, a fim de acrescentar o tipo e explicar como estão organizados os dados."*; e
- *"Um modelo que pudesse expressar melhor o que cada construtor representa, ou seja, com definições de que se trata. Por exemplo, definir o que é um array, objeto, etc."*

Ambas as informações obtidas, a de intensidade e a escrita, expressam que os construtores definidos no NoAM deveriam ser repensados de acordo com os dados analisados desse grupo.

Na coluna da Tabela 5, a diferença entre as medianas foi uma das mais expressivas, sendo essa diferença de 3 pontos na escala de um modelo para o outro. A questão analisada foi: *A respeito dos construtores de modelagem de relacionamentos do modelo, estes são capazes de expressar qualquer relacionamento entre objetos de um domínio NoSQL?*

Foi pedido aos participantes que explicassem, por meio de um exemplo, porque o NoAM não conseguia expressar qualquer relacionamento entre objetos NoSQL. Dois participantes desse grupo responderam essa questão, o primeiro afirmou:

- *"Eu acho que modelo não é intuitivo. Existe uma certa dificuldade ao escolher qual construtor a ser utilizado, para representar uma coleção, por exemplo"*.

. Enquanto o segundo deu um exemplo de um *array*, como pode ser visto no seguinte trecho:

- *"O relacionamento entre os objetos, deve ser representado por um Array, Objeto ou outra estrutura de dados"*.

Assim, concluí-se que a dificuldade encontrada impactou na forma com que esse grupo viu o poder de expressividade do NoAM e na capacidade desse modelo conseguir representar domínios NoSQL.

Tabela 5 – Resultados de representatividade do grupo presencial.

Modelos	Escolha	Identificação	Expressivo	Elementos	Poder de expressividade	NoSQL
NoAM	3,5	3,5	3,5	3	3,5	1,5
COMN	2	3	4	4	5	4,5

Fonte: O autor (2019)

Como foi visto nas questões relacionadas a representatividade, o modelo COMN obteve melhores resultados, dessa forma, para o grupo esse modelo de dados é mais expressivo que o NoAM, quando observado os domínios que foram investigados.

5.2.2 Resultados do grupo remoto ao questionário de usabilidade e expressividade

A análise realizada no grupo remoto foi feita da mesma forma que a investigação no grupo presencial. Primeiro as questões a respeito da usabilidade foram investigadas e, em seguida as questões a respeito da representatividade dos modelos de dados.

5.2.2.1 Análise da Usabilidade

A respeito da experiência de uso do grupo remoto, grande parte das questões onde essa foi observada obtiveram as mesmas medianas, assim, mostrando que para esse grupo não houve diferença entre os modelos investigados ao analisar a usabilidade como pode ser visto na Tabela 6. A única questão que apresentou um resultado diferente foi *se os participantes gostariam de utilizar um dos modelos investigados futuramente*. O modelo COMN teve uma mediana maior que a mediana obtida pelo NoAM, sendo a diferença de 0,5 ponto de uma mediana para a outra. Essa diferença como os demais resultados achados são apresentados na Tabela 6.

Tabela 6 – Resultados de usabilidade do grupo remoto.

Modelos	Experiência de uso	Facilidade no uso	Dificuldade no uso	Possibilidade de uso futuro
NoAM	4,5	5	3	3
COMN	4,5	5	3	3,5

Fonte: O autor (2019)

Os achados desse grupo em relação a usabilidade mostram que não houve diferença entre utilizar o COMN ou NoAM.

5.2.2.2 Análise da Representatividade

A investigação da representatividade pode ser observada a partir dos dados obtidos do questionário de usabilidade e expressividade dos modelos de dados (presente no Apêndice D), das dez questões objetivas, seis dessas foram voltadas para a representatividade. Apenas uma das questões obteve a mesma mediana para ambos os modelos, sendo essa questão *Houve dificuldade para a escolha de um ou outro construtor do modelo para representar algum aspecto do domínio escolhido?*, a mediana obtida ficou 2,5 para cada modelo (Tabela 7). Assim, mostrando que para esse grupo não houve diferença na identificação de um construtor ou outro para representar algum aspecto do domínio.

As demais questões apresentaram medianas diferentes como pode ser visto na Tabela 7, em relação a *se os participantes tiveram alguma dúvida em escolher um ou outro construtor do modelo*, o COMN obteve uma mediana menor, sendo a diferença apresentada de 0,5 ponto na mediana dos modelos, pois para essa questão quanto menor a mediana fosse

melhor avaliado o modelo foi. Para essa questão foi criada uma questão aberta, para que os participantes que reportaram alguma dúvida falasse qual foi a dúvida que esses encontram em escolher um outro construtor para representar algo. No grupo remoto os participantes reportaram dúvidas para ambos os modelos.

A respeito do NoAM um dos participantes disse: *"Fiquei em dúvida sobre em qual agregado representar os relacionamentos."* Outro participante também expressou dificuldade em representar a cardinalidade dos relacionamentos no NoAM, essa dificuldade foi expressa na seguinte frase: *"Não sabia se deveria representar o relacionamento de 1 para n, de N para N da maneira como representei"*.

Já os participantes que citaram alguma dúvida no uso dos construtores do COMN, disseram as seguintes frases: *"Saber o escopo de um 'valor' ou 'significado' é confuso. Saber quando usar tracejado ou não também é confuso."*; *"Não tive muita segurança ao utilizar as representações de relacionamento."* e *"Entender como expressar a cardinalidade nessa representação e entender a diferença entre conceitual ou não dentro desse modelo."* Dois dos participantes reportaram a dificuldade de entender quando usar ou não relacionamentos ou objetos conceituais disponíveis no modelo. Apenas um dos três que reportaram alguma dificuldade não falou sobre a forma de representar objetos conceituais, esse só mencionou a dificuldade de representar relacionamentos no COMN.

Tabela 7 – Resultados de representatividade do grupo remoto.

Modelos	Escolha	Identificação	Expressivo	Elementos	Poder de expressividade	NoSQL
NoAM	3,5	2,5	3	3,5	4	2,5
COMN	2	2,5	3,5	4	5	4,5

Fonte: O autor (2019)

As demais questões o modelo COMN teve medianas maiores que as medianas obtidas no NoAM, indicando que para esse grupo o modelo COMN apresenta uma maior variabilidade. Uma das questões que mais ficou expressa essa opinião foi *A respeito dos construtores de modelagem de relacionamentos do modelo, estes são capazes de expressar qualquer relacionamento entre objetos de um domínio NoSQL?*. Houve uma diferença de 2 pontos do NoAM para o COMN como descrito na Tabela 7. Essa diferença aponta que, segundo o grupo, o COMN pode ser usado para expressar qualquer objeto e relacionamento em um domínio NoSQL. Enquanto que, o modelo NoAM esses não souberam dizer, uma vez que a mediana da métrica do grupo ficou entre 3- talvez e 2- não sei.

5.2.3 Testes de hipótese para UMD e RMD

A hipótese nula levantada para a UMD e RMD (Seção 4.3) foi que não haveria diferença em usar o COMN e o NoAM quando observada a usabilidade e representatividade desses modelos de dados. Dessa forma, foram levantadas duas hipóteses nulas, uma para a UMD

e a outra para RMD. Cada uma dessas hipóteses possui uma hipótese alternativa, onde, essas se apresentam se uma das hipóteses nulas forem rejeitadas. Assim, a hipótese alternativa se apresenta quando há sim diferença entre o utilizar o COMN ou NoAM quando observados os dados das métricas UMD e RMD. O teste de Wilcoxon foi utilizado para observar se as métricas foram ou não aceitas pelas hipóteses nulas definidas no estudo.

Os dados observados para as duas métricas foram obtidos a partir das respostas dos participantes do questionário de usabilidade e expressividade dos modelos de dados (Apêndice D). Assim, os dados das respectivas métricas foram agrupados e o teste de Wilcoxon foi utilizado.

Os primeiros dados observados foram os relacionados com a usabilidade, o resultado obtido foi: $W = 978.5$, $p\text{-valor} = 0.07711$. O p -valor foi maior que o nível de significância, dessa forma, não foi dada nenhuma razão para concluir que os dados a respeito da UMD dos modelos investigados diferem. Assim, a hipótese nula para a UMD foi aceita, mostrando que em relação a usabilidade dos modelos de dados não há diferença entre utilizar o modelo NoAM ou COMN. A partir, da análise feita nos grupos (Seções 5.2.1.1 e 5.2.2.1) pode-se observar que em um dos grupos, o remoto, não houve diferença em usar o COMN ou NoAM quando a usabilidade foi analisada (onde as mesmas medianas foram encontradas para ambos os modelos de dados, com excessão de uma mediana como foi mostrado na Seção 5.2.2.1).

A respeito da métrica RMD está foi observada a partir de duas propriedades a expressividade gráfica e o poder de expressividade dos modelos de dados (Seção 4.2). Os dados obtidos das questões relacionadas as duas propriedades observadas na RMD foram agrupados e o teste de Wilcoxon foi realizado, assim, observando se a hipótese nula para a RMD foi aceita ou não.

De acordo com o resultado obtido no teste para a métrica RMD, a hipótese nula não foi rejeitada ao nível de significância de 5% para esse caso, o resultado obtido foi $W = 2112$, $p\text{-valor} = 0.09312$, onde o resultado do p -valor foi maior que o nível de significância de 5%, por essa razão, conclui-se que para o estudo realizado não houve diferença entre utilizar o modelo COMN e NoAM quando a RMD foi analisada.

5.3 ANÁLISE DOS DIAGRAMAS MODELADOS

Os diagramas produzidos pelos participantes foram avaliados por dois especialistas, como explicado na Seção 4.8. Cada diagrama modelado foi avaliado por ambos, onde essa avaliação contou com um questionário (Apêndice E), sendo esse preenchido conforme o diagrama obtido atendia ou não as especificações feitas para a construção do esquema. Assim, foi avaliado pelos especialistas a qualidade dos modelos de dados, por meio da corretude e completude dos esquemas construídos, isto é, se os esquemas modelados expressaram de forma completa e correta o domínio que foi abordado. Além disso, foi observada se os modelos de dados são independentes de paradigma de implementação.

Para que a avaliação efetuada pelos especialistas pudesse ocorrer foi dado um documento explicando a função dos construtores dos modelos estudados. Além disso, foi passado aos especialistas o minimundo (ou problema do domínio), dessa maneira, esses teriam a visão completa do que foi solicitado, podendo avaliar se os esquemas construídos expressavam o pedido feito de forma correta e completa.

Figura 11 – Resultado Kappa para o NoAM

```

3 library(irr)
4
5 #Base de dados
6 dadoskappa<- read.csv("C:\\users\\idacio.a\\Desktop\\dacio.csv", header = TRUE, sep = ";")
7 head(dadoskappa)
8
9 dadoskappa<-cbind(dadoskappa$AV1,dadoskappa$AV2)
10 head(dadoskappa)
11
12 kappa2(dadoskappa)
13
14 #Conclusão do teste
15 pvalor=kappa2(dadoskappa)$p.value
16 alfa=0.05
17
18
19
20
21
22:1 (Top Level)
R Script

```

```

~/R/R-4.0.2
> resposta<-function(pvalor) {
+   if(pvalor<alfa) { cat("pvalor é ",pvalor, " logo é pequeno >>> rejeita HO ", "\n") }
+   else if(pvalor>=alfa) { cat("pvalor é ",pvalor, " logo é grande >>> NÃO rejeita HO ", "\n")
+ }
+ }
> rTOTAL <- resposta(pvalor)
pvalor é 7.349676e-14 logo é pequeno >>> rejeita HO
> kappa2(dadoskappa)
Cohen's Kappa for 2 Raters (weights: unweighted)

Subjects = 100
Raters = 2
Kappa = 0.548

z = 7.48
p-value = 7.35e-14
> rTOTAL <- resposta(pvalor)
pvalor é 7.349676e-14 logo é pequeno >>> rejeita HO
> |

```

Fonte: O autor (2019)

Antes de analisar as avaliações dos especialistas no que diz respeito à qualidade dos modelos de dados, estas foram aplicadas ao teste de *Kappa*. O teste de *Kappa* é utilizado para descrever o nível de concordância entre dois ou mais juízes (UEBERSAX, 1982). Por essa razão, esse teste foi aplicado ao estudo para verificar o nível de concordância entre os dois especialistas.

Os resultados desse teste mostram que houve concordância entre esses dois especialistas, assim, os dados podiam ser aplicados ao teste de Wilcoxon, observando se a QMD foi aceita pela hipótese nula ou não (A definição da métrica e da hipótese é encontrada nas Seções 4.2 e 4.3).

Primeiro, o teste de *Kappa* foi utilizado na validação do NoAM, esse teste foi executado

Figura 12 – Resultado Kappa para o COMN

```

File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Run Source
32 dadoskappa<-cbind(dadoskappa$AV1,dadoskappa$AV2)
33 head(dadoskappa)
34
35 kappa2(dadoskappa)
36
37 #Conclusão do teste
38 pvalor=kappa2(dadoskappa)$p.value
39 alfa=0.05
40 resposta<-function(pvalor) {
41   if(pvalor<alfa) { cat("pvalor é ",pvalor, " logo é pequeno >>> rejeita Ho ", "\n") }
42   else if(pvalor>=alfa) { cat("pvalor é ",pvalor, " logo é grande >>> NÃO rejeita Ho ", "\n") }
43 }
44 rTOTAL <- resposta(pvalor)
45
45:1 (Untitled) R Script
Console Terminal
~/
> kappa2(dadoskappa)
Cohen's Kappa for 2 Raters (weights: unweighted)

Subjects = 100
Raters = 2
Kappa = 0.47

z = 6.92
p-value = 4.67e-12
> #Conclusão do teste
> pvalor=kappa2(dadoskappa)$p.value
> alfa=0.05
> resposta<-function(pvalor) {
+   if(pvalor<alfa) { cat("pvalor é ",pvalor, " logo é pequeno >>> rejeita Ho ", "\n") }
+   else if(pvalor>=alfa) { cat("pvalor é ",pvalor, " logo é grande >>> NÃO rejeita Ho ", "\n") }
+ }
> rTOTAL <- resposta(pvalor)
pvalor é 4.667156e-12 logo é pequeno >>> rejeita Ho
>

```

Fonte: O autor (2019)

no RStudio, *software* popular para análise de dados estatísticos. Os resultados achados mostram que o nível de concordância entre os especialistas não foi aleatório. Pois, de acordo com a tabela definida por Landis (BYRT; BISHOP; CARLIN, 1993) para descrever o nível de concordância entre os juízes, o nível de concordância obtido na análise do NoAM é dito como moderado. Uma vez que o valor moderado fica entre 0,41 a 0,6 e obtido na análise do NoAM foi 0,548 como visto na Figura 11, evidencia-se que os especialistas tiveram um nível de concordância moderada quando avaliaram os diagramas desenvolvidos pelos participantes no NoAM.

O teste de *Kappa* também foi utilizado na validação feita pelos especialistas no COMN. O nível de concordância obtido foi 0,47, como mostrado na Figura 12, o que indica também um nível moderado de concordância.

A partir dos dois testes executados, pode-se dizer que o nível de concordância entre os especialistas não foi meramente aleatório, assim, não foi necessário contatar um terceiro

especialista.

O questionário que os especialistas preencheram abrangia questionamentos a respeito da correteza dos esquemas construídos pelos participantes, nesse questionário as respostas foram construídas a partir da escala de *Likert*, possuindo três níveis (3-máximo, 2-médio e 1-mínimo). Assim, esse questionário serviu como um *checklist*, onde os especialistas observavam se as propriedades abordadas no questionário foram ou não atendidas nos diagramas construídos pelos participantes.

A análise foi feita da seguinte forma para cada questionamento foi: obtida a mediana e a moda das respostas dos especialistas para cada questão presente no questionário (Apêndice E). Assim, tornando possível saber qual dos modelos de dados estudados foi melhor avaliado pelos especialistas e qual foi a tendência na escala em suas avaliações.

As questões desse questionário se dividiram da seguinte forma: três questões a respeito da completude dos esquemas (questões 1, 3 e 4); correteza (questões 2, 5, 6, 7 e 8) e as questões 9 e 10 voltadas para a independência de paradigma de implementação.

5.3.1 Análise da avaliação dos especialistas sobre a completude

A completude dos diagramas desenvolvidos foi observada através de três questões, como dito anteriormente. Nesta seção cada uma das questões será discutida.

Tabela 8 – Resultados das medianas da completude de cada modelo.

Questões	NoAM	COMN
1- Propriedades relevantes	3	2
3- Características redundantes	3	3
4- Não necessidade de algum elemento	3	3

Fonte: O autor (2019)

Tabela 9 – Resultados das modas da completude de cada modelo.

Questões	NoAM	COMN
1- Propriedades relevantes	3	2
3- Características redundantes	3	3
4- Não necessidade de algum elemento	3	3

Fonte: O autor (2019)

A primeira questão abordou se todas as propriedades relevantes do domínio estavam descritas no esquema construído. Assim, quanto mais próximo do valor 3 a mediana chegasse, mais o esquema construído possuía todas as propriedades necessárias. Na avaliação

dos especialistas, o NoAM obteve um melhor resultado, como visto nas Tabelas 8 e 9, indicando que os esquemas construídos no modelo foram mais completos. A terceira e quarta questões abordaram a redundância dos esquemas construídos, ou seja, propriedades não necessárias que foram adicionadas no esquema pelos participantes não sendo importantes para a compreensão do domínio ou sendo essas propriedades prolixas. Para essas questões, quanto mais próximo do valor 3, menos propriedades redundantes foram adicionadas no esquema. Como visto na segunda e terceira linha das Tabela 8 e 9, ambos os modelos obtiveram bons resultados.

O que mostra que em relação à completude dos esquemas construídos ambos os modelos podem ser ditos modelos completos, por serem claros conterem toda a informação necessária para a compreensão do domínio e por não serem redundantes em relação às propriedades adicionadas para representar esse domínio. Contudo, na avaliação realizada pelos especialistas, o NoAM foi melhor avaliado em uma das três questões a respeito da completude, indicando que os esquemas modelados pelos participantes utilizando o NoAM foram mais completos.

5.3.2 Análise da avaliação dos especialistas sobre à corretude

Nesta seção é descrito como os especialistas avaliaram à corretude dos esquemas construídos. As questões que estes avaliaram também possuem três níveis, tais como as questões elaboradas para analisar à completude. Assim, quanto mais próximo do valor 3, mais correta a característica analisada estava. O primeiro questionamento feito foi se o diagrama desenvolvido estava de acordo com o requerido no minimundo aplicado no estudo, bem como é mostrado nas Tabelas 10 e 11, ambos os modelos NoAM e COMN tiveram resultados iguais, demonstrando que os diagramas desenvolvidos atenderam de forma igual os requisitos pedidos no estudo. A próxima questão feita foi se alguma requisição pedida no minimundo não foi incluída no diagrama final, o COMN mostrou um resultado menor em comparação ao NoAM, podendo ser visto nas Tabelas 10 e 11. Assim, expondo que os participantes, quando usaram o COMN esqueceram-se de adicionar propriedades necessárias no esquema construído.

As questões seis e sete são relacionadas com o uso correto dos construtores do modelo. Assim, essas questões analisaram se o participante estava usando o construtor certo para representar determinado relacionamento, por exemplo.

O NoAM apresentou um resultado maior se comparado ao COMN em um dos questionamentos (Tabelas 10 e 11), indicando que os participantes acertaram mais ao utilizar os construtores dispostos no NoAM. Como pode ser observado nas Tabelas 10 e 11, de modo geral o NoAM teve melhores avaliações, o que indica que o uso do NoAM teve um impacto maior na corretude dos esquemas modelados pelos participantes. Contudo, como apresentado nas Seções 5.1 e 5.2, o COMN apresentou melhores resultados quando analisadas as métricas CMD e RMD. O que contradiz com a avaliação feita pelos especialistas

Tabela 10 – Resultados das modas da corretude de cada modelo.

Questão	NoAM	COMN
2- Diagrama está de acordo com as especificações	2	2
5- Falta de inclusão de alguma propriedade	2,5	1
6- Uso correto dos construtores	3	2
7- Relacionamentos estão definidos corretamente	2	2
8- Algum erro no diagrama	3	1

Fonte: O autor (2019)

Tabela 11 – Resultados das medianas da corretude de cada modelo.

Questão	NoAM	COMN
2- Diagrama está de acordo com as especificações	2	2
5- Falta de inclusão de alguma propriedade	3	1
6- Uso correto dos construtores	3	2
7- Relacionamentos estão definidos corretamente	2	2
8- Algum erro no diagrama	3	1

Fonte: O autor (2019)

apresentando um resultado inferior, mas esse resultado pode ser justificado pelo seguinte pensamento: o COMN possui um número maior de construtores a ser usado o que pode ter confundido os participantes na hora de escolher qual construtor utilizar (como relatado por um dos participantes na Seção 5.2.2.2), prejudicando na corretude dos esquemas por eles modelados. Essa dificuldade pode ter impactado nos resultados inferiores que o COMN apresentou em relação ao NoAM. E como o NoAM possui um número menor de construtores, pode ter sido favorecido na observação desse questionamento. Contudo, essa teoria só pode ser confirmada com a realização de um novo estudo, onde isso seja observado de forma mais profunda.

O que leva a última questão a respeito da corretude, sendo essa: **Alguma regra do domínio foi expressa no diagrama erroneamente?**. Mais uma vez o COMN apresentou um resultado menor na mediana e na moda encontrada, como pode ser visto na última linha das Tabelas 10 e 11. Esse resultado pode estar conectado com o pensamento anterior, de que os participantes ficaram confusos em qual construtor utilizar, assim, estando mais fadado ao erro quando o modelo possui um número maior de construtores.

As duas perguntas restantes do questionário aplicado aos especialistas obteve resultados iguais. Assim, expondo que ambos os modelos, NoAM e COMN, são independentes

de paradigma de implementação, o que é o ideal, já que ambos são modelos de dados conceituais.

De acordo com a validação feita pelos especialistas, foi possível atribuir uma nota para cada modelo estudado. Essa nota foi obtida por meio do questionário, onde cada questão completamente tratada pelo diagrama desenvolvido pelos participantes vale 1 ponto; e se foi parcialmente vale 0,5 e se essa não foi atendida vale 0. As notas obtidas foram somadas e foi extraída a média da avaliação para cada modelo. Os esquemas construídos no NoAM obtiveram uma melhor nota, sendo essa 7,25 de média pelos diagramas modelados. Já o COMN teve sua média de 6,25, o que mostra que os esquemas construídos pelos participantes tiveram mais erros que os esquemas construídos pelo NoAM.

Na próxima seção a hipótese QMD é analisada. Assim, concluindo se há ou não diferença na qualidade dos modelos de dados quando observados o NoAM e o COMN.

5.3.3 Teste de hipótese para QMD

O teste de Wilcoxon foi feito para a QMD, onde foi observado se essa métrica foi aceita pela hipótese nula, assim, mostrando que não houve diferença entre os modelos de dados quando investigada a métrica QMD.

O resultado obtido foi $W = 16610$, $p\text{-valor} = 0.001012$, onde $p\text{-valor}$ refuta a hipótese nula. Ou seja, houve sim diferença entre utilizar os modelos investigados ao que diz respeito a métrica QMD.

6 CONCLUSÃO

Neste capítulo são elencados os resultados obtidos através do estudo realizado. Além disso, são expostos trabalhos futuros na área de pesquisa.

6.1 DISCUSSÃO

Os dados obtidos no estudo podem ser analisados por outras perspectivas que não faziam parte do foco da pesquisa discutida nesta dissertação, como: quais dos grupos deram o nível máximo da escala definida nos questionários em mais questões, se o sexo da pessoa afetou em suas respostas, faixa etária entre outras características.

O ambiente em que os participantes estavam pode também ser observado, investigando se o ambiente afeta ou não nos artefatos produzidos pelos participantes. Observou-se que os participantes que realizaram as atividades presencialmente tenderam a dar melhores avaliações para o modelo COMN, no qual isto pode ser visto nas Seções 5.1.1 e 5.2.1. Enquanto, os resultados do grupo remoto de forma geral ficaram bem próximos para ambos os modelos: o COMN e o NoAM. Assim, mostrando que não houve muita diferença nas características observadas, sendo essa diferença na avaliação notada em uma ou outra questão das métricas investigadas. Diferentemente do grupo presencial que teve quase todas as questões da UMD, CMD e RMD com medianas maiores para o COMN. A diferença nos resultados que foi mais evidente no grupo remoto foi em relação a métrica RMD (5.2.2.2), onde o COMN teve melhores avaliações se comparadas com as avaliações feitas para o NoAM.

Como pôde ser observado nos resultados obtidos dos dois grupos investigados, houve uma diferença em cada grupo, onde essa pode ter sido afetada pelo ambiente do grupo, impactando nas respostas desses participantes. Porém, se faz necessário realizar um novo estudo, onde possa ser observado se as condições físicas impactaram realmente na forma com qual esses participantes avaliaram os modelos investigados.

Além do que foi discutido sobre o ambiente ter ou não impactado nas respostas dos participantes, poderia também ter sido analisado se os diagramas modelados pelos participantes foram mais completos para aqueles que o modelaram de forma remota ou presencial. Assim, podendo traçar uma linha de pensamento se o ambiente afeta ou não no processo de modelagem de dados. No próxima seção são discutidas as contribuições do trabalho.

6.2 CONTRIBUIÇÕES

A análise de modelos de dados conceituais através de pesquisas com participantes não é muito usual na área de banco de dados. Normalmente, quando se propõe um modelo

de dados conceitual é feito um estudo de caso mostrando se o modelo proposto aborda o domínio para qual foi desenvolvido, sem observar características importantes como a facilidade com que novos usuários se adaptariam ao modelo caso esse fosse aplicado em seu ambiente de trabalho, se esse é de fácil entendimento e uso entre outras propriedades. Assim, em estudos que propõem modelos de dados conceituais é raro encontrar autores que se preocupem em observar essas características no levantamento realizado nesta dissertação foram encontrados alguns estudos que analisaram modelos de dados conceituais, sendo em sua maioria modelos de dados conceituais voltados para sistemas relacionais de dados.

A principal contribuição deste trabalho foi trazer a análise voltada para o ponto de vista do usuário e desenvolvedor no uso de modelos de dados conceituais. Essa foi possível a partir do levantamento realizado na Seção 3.2, onde foram levantadas as propriedades que seriam observadas, sendo essas a simplicidade, representatividade e qualidade do modelo de dados. Após a definição das características observadas, foi selecionado dois modelos de dados conceituais voltados para sistemas NoSQL, onde uma análise feita para tais modelos não foi encontrada no levantamento do estado da arte. Além da contribuição inédita que foi realizar esse tipo de análise em modelos de dados conceituais voltados para NoSQL, a elaboração do desenho experimental também foi uma contribuição, detalhando todo o processo utilizado para a obtenção dos dados a serem analisados. Assim, tornando possível a pesquisadores que queiram realizar pesquisas na mesma área utilizar como base o desenho desenvolvido para executar novas pesquisas ou até replicar o estudo descrito nesta dissertação, observando se os dados obtidos são diferentes.

Outra contribuição do trabalho foi a realização do estudo com dois grupos: um remoto e um presencial. Onde a partir da percepção de cada grupo pode-se observar se houve ou não concordância entre esses grupos quando observadas as propriedades investigadas no estudo.

6.3 TRABALHOS FUTUROS

Os dados obtidos na investigação podem servir de base para a construção de novos modelos conceituais para sistemas NoSQL, onde pode ser observado possíveis melhorias nos modelos analisados e propor um novo que atenda as características discutidas nesta dissertação.

Dentre os possíveis trabalhos futuros na área pode-se adicionar novas propriedades a serem observadas, assim, tornando mais completa a análise dos modelos de dados. Além disso, pode-se ofertar uma disciplina, onde os conceitos dos modelos serão explicados mais detalhadamente, possibilitando um maior aprendizado dos modelos de dados conceituais. Ademais, como um maior tempo para analisar os dados obtidos, pode-se adicionar novos domínios a serem estudados, proporcionando uma investigação mais ampla. Ademais, podem ser observados outros detalhes como os discutidos na Seção 6.1.

6.4 CONCLUSÕES DO TRABALHO

O trabalho apresentou uma forma de avaliar modelos de dados conceituais voltados para sistemas de BD NoSQL, nessa avaliação dois modelos de dados conceituais foram investigados: o NoAM e COMN.

O objetivo geral deste trabalho foi analisar o quanto a escolha do modelo de dados podia interferir na compreensão e no desenvolvimento de projetos conceituais de BD NoSQL. Para atingir esse objetivo foram levantadas quatro métricas a serem investigadas: a CMD, UMD, RMD e QMD. A investigação nas métricas CMD, UMD, RMD foram realizadas a partir dos resultados obtidos para cada grupo e por meio dos testes de hipóteses feitos para cada métrica. Além disso, os diagramas modelados pelos participantes utilizando os construtores do NoAM e COMN foram avaliados por dois especialistas, onde a QMD pôde ser observada.

A partir dos testes de hipóteses feitos pôde-se observar que em dois desses (para a métrica CMD e QMD) foram obtidos bons resultados, sendo refutados na hipótese nula e mostrando que há diferença entre utilizar o NoAM e COMN quando observadas a CMD e QMD.

REFERÊNCIAS

- ABDELHEDI, F.; BRAHIM, A. A.; ATIGUI, F.; ZURFLUH, G. Big data and knowledge management: How to implement conceptual models in nosql systems?. In: *KMIS*. [S.l.: s.n.], 2016. p. 235–240.
- AMIRIAN, P.; BASIRI, A.; WINSTANLEY, A. Implementing geospatial web services using service oriented architecture and nosql solutions. In: *The Third International Conference on Digital Information and Communication Technology and its Applications*. [S.l.: s.n.], 2013.
- ASSENOVA, P.; JOHANNESSON, P. Improving quality in conceptual modelling by the use of schema transformations. In: SPRINGER. *International Conference on Conceptual Modeling*. [S.l.], 1996. p. 277–291.
- BANERJEE, S.; SHAW, R.; SARKAR, A.; DEBNATH, N. C. Towards logical level design of big data. In: IEEE. *2015 IEEE 13th International Conference on Industrial Informatics (INDIN)*. [S.l.], 2015. p. 1665–1671.
- BENMAKHLOUF, A. Nosql implementation of a conceptual data model: Uml class diagram to a document-oriented model. In: . [S.l.]: IJDMS, 2018. v. 10, n. 2.
- BREWER, E. Cap twelve years later: how the. *Computer*, IEEE, n. 2, p. 23–29, 2012.
- BRODIE, M. L. On the development of data models. In: *On conceptual modelling*. [S.l.]: Springer, 1984. p. 19–47.
- BRODIE, M. L.; MYLOPOULOS, J.; SCHMIDT, J. W. On conceptual modeling: Perspectives from artificial intelligence, databases, and programming languages. In: . [S.l.]: Springer Science Business Media, 2012.
- BUGIOTTI, F.; CABIBBO, L.; ATZENI, P.; TORLONE, R. Database design for nosql systems. In: SPRINGER. *International Conference on Conceptual Modeling*. [S.l.], 2014. p. 223–231.
- BYRT, T.; BISHOP, J.; CARLIN, J. B. Bias, prevalence and kappa. *Journal of clinical epidemiology*, Elsevier, v. 46, n. 5, p. 423–429, 1993.
- CATTELL, R. Scalable sql and nosql data stores. *Acm Sigmod Record*, ACM, v. 39, n. 4, p. 12–27, 2011.
- CHEBOTKO, A.; KASHLEV, A.; LU, S. A big data modeling methodology for apache cassandra. In: IEEE. *2015 IEEE International Congress on Big Data*. [S.l.], 2015. p. 238–245.
- CHERFI, S. S.-S.; AKOKA, J.; COMYN-WATTIAU, I. Conceptual modeling quality-from eer to uml schemas evaluation. In: SPRINGER. *International Conference on Conceptual Modeling*. [S.l.], 2002. p. 414–428.
- CONNOLLY, T. M.; BOYLE, E. A.; MACARTHUR, E.; HAINEY, T.; BOYLE, J. M. A systematic literature review of empirical evidence on computer games and serious games. *Computers & education*, Elsevier, v. 59, n. 2, p. 661–686, 2012.

- CORONEL, C.; MORRIS, S. *Database systems: design, implementation, & management*. [S.l.]: Cengage Learning, 2016.
- CUZICK, J. A wilcoxon-type test for trend. *Statistics in medicine*, Wiley Online Library, v. 4, n. 4, p. 543–547, 1985.
- ELMASRI, R.; NAVATHE, S. *Fundamentals of database systems*. [S.l.]: Addison-Wesley Publishing Company, 2010.
- ENGELS, G.; GOGOLLA, M.; HOHENSTEIN, U.; HÜLSMANN, K.; LÖHR-RICHTER, P.; SAAKE, G.; EHRICH, H.-D. Conceptual modelling of database applications using an extended er model. *Data & Knowledge Engineering*, Elsevier, v. 9, n. 2, p. 157–204, 1992.
- FETTKE, P.; LOOS, P. Multiperspective evaluation of reference models—towards a framework. In: SPRINGER. *International Conference on Conceptual Modeling*. [S.l.], 2003. p. 80–91.
- FOSTER, E. C.; GODBOLE, S. *Database systems: a pragmatic approach*. [S.l.]: Apress, 2016.
- FOWLER, A. *NoSQL for dummies*. [S.l.]: John Wiley & Sons, 2015.
- GANIEE, N.; BHARGAVA, R. Nosql for interactive applications. *International Journal of Allied Practice, Research and Review*, Website: www.ijaprr.com (ISSN 2350-1294), retrieved on, 2015.
- GENERO, M.; OLIVAS, J. A.; PIATTINI, M.; ROMERO, F. P. Knowledge discovery for predicting entity relationship diagram maintainability. In: SEKE. [S.l.: s.n.], 2001. p. 203–211.
- GENERO, M.; POELS, G.; PIATTINI, M. Defining and validating measures for conceptual data model quality. In: SPRINGER. *International Conference on Advanced Information Systems Engineering*. [S.l.], 2002. p. 724–727.
- HAMMAR, K. *Quality of content ontology design patterns*. IOS Press, 2016.
- HAMOUDA, S.; ZAINOL, Z. Document-oriented data schema for relational database migration to nosql. In: IEEE. *2017 International Conference on Big Data Innovations and Applications (Innovate-Data)*. [S.l.], 2017. p. 43–50.
- HANNAY, J. E.; SJOBERG, D. I.; DYBA, T. A systematic review of theory use in software engineering experiments. *IEEE transactions on Software Engineering*, IEEE, v. 33, n. 2, p. 87–107, 2007.
- HARRISON, G. Sharding, amazon, and the birth of nosql. In: *Next Generation Databases*. [S.l.]: Springer, 2015. p. 39–51.
- HASEEB, A.; PATTUN, G. A review on nosql: Applications and challenges. *International Journal of Advanced Research in Computer Science*, v. 8, n. 1, 2017.
- HASHEM, H.; RANC, D. An integrative modeling of bigdata processing. *IJCSA*, v. 12, n. 1, p. 1–15, 2015.
- HILLS, T. *NoSQL and SQL data modeling: bringing together data, semantics, and software*. [S.l.]: Technics Publications, 2016.

- HUSSAIN, T. Revisiting quality metrics for conceptual models. In: IEEE. *TENCON 2014-2014 IEEE Region 10 Conference*. [S.l.], 2014. p. 1–6.
- HUSSAIN, T.; AWAIS, M. M. An effort-based approach to measure completeness of an entity-relationship model. In: IEEE. *Seventh IEEE/ACIS International Conference on Computer and Information Science (icis 2008)*. [S.l.], 2008. p. 463–468.
- HUSSAIN, T.; TAHIR, A. S.; AWAIS, M. M.; SHAMAIL, S. Analytical hierarchy process approach to rank measures for structural complexity of conceptual models. In: IEEE. *2006 IEEE International Multitopic Conference*. [S.l.], 2006. p. 255–258.
- JOVANOVIĆ, V.; BENSON, S. Aggregate data modeling style. In: *Proceedings of the Southern Association for Information Systems Conference*. [S.l.: s.n.], 2013. p. 70–75.
- JUDD, C. M.; SMITH, E. R.; KIDDER, L. H. *Research methods in social relations*. Harcourt Brace Jovanovich, 1991.
- JURISTO, N.; MORENO, A. M. *Basics of software engineering experimentation*. [S.l.]: Springer Science & Business Media, 2013.
- KASHYAP, S.; ZAMWAR, S.; BHAVSAR, T.; SINGH, S. Benchmarking and analysis of nosql technologies. *Int J Emerg Technol Adv Eng*, v. 3, n. 9, p. 422–426, 2013.
- KAUR, K.; RANI, R. Modeling and querying data in nosql databases. In: IEEE. *2013 IEEE International Conference on Big Data*. [S.l.], 2013. p. 1–7.
- KIM, S.-S.; YU, S.-H. Architecture of geospatial big-data batch processing model based on hadoop. *International Conference on Information and Communication Technology Convergence (ICTC)*, IEEE, n. D1, p. D964–D966, 2015.
- LAKE, P.; CROWTHER, P. Nosql databases. In: *Concise Guide to Databases*. [S.l.]: Springer, 2013. p. 97–134.
- LEAVITT, N. Will nosql databases live up to their promise? *Computer*, IEEE, v. 43, n. 2, p. 12–14, 2010.
- LEITE, L. A. Banco de dados cassandra: um estudo de caso para análise dos dados dos servidores públicos federais. 2014.
- LI, S.; DRAGICEVIC, S.; CASTRO, F. A.; SESTER, M.; WINTER, S.; COLTEKIN, A.; PETTIT, C.; JIANG, B.; HAWORTH, J.; STEIN, A. et al. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing*, Elsevier, v. 115, p. 119–133, 2016.
- LIMA, C.; MELLO, R. S. On proposing and evaluating a nosql document database logical approach. *International Journal of Web Information Systems*, Emerald Group Publishing Limited, v. 12, n. 4, p. 398–417, 2016.
- LONGLEY, P. A.; GOODCHILD, M. F.; MAGUIRE, D. J.; RHIND, D. W. *Geographic information systems and science*. [S.l.]: John Wiley & Sons, 2005.
- LUKYANENKO, R.; SAMUEL, B. M. Are all classes created equal? increasing precision of conceptual modeling grammars. *ACM Transactions on Management Information Systems (TMIS)*, ACM, v. 8, n. 4, p. 14, 2017.

- MCAFEE, A.; BRYNJOLFSSON, E.; DAVENPORT, T. H.; PATIL, D.; BARTON, D. Big data: the management revolution. *Harvard business review*, v. 90, n. 10, p. 60–68, 2012.
- MCCREARY, D.; KELLY, A. *Making sense of NoSQL*. [S.l.]: Greenwich, Conn.: Manning Publications, 2014.
- MEHMOOD, K.; CHERFI, S. S.-S.; COMYN-WATTIAU, I. Data quality through model quality: a quality model for measuring and improving the understandability of conceptual models. In: ACM. *Proceedings of the first international workshop on Model driven service engineering and data quality and security*. [S.l.], 2009. p. 29–32.
- MONIRUZZAMAN, A.; HOSSAIN, S. A. Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *arXiv preprint arXiv:1307.0191*, 2013.
- MOODY, D. L. Measuring the quality of data models: an empirical evaluation of the use of quality metrics in practice. *ECIS 2003 Proceedings*, p. 78, 2003.
- MOODY, D. L.; SHANKS, G. G. What makes a good data model? evaluating the quality of entity relationship models. In: SPRINGER. *International Conference on Conceptual Modeling*. [S.l.], 1994. p. 94–111.
- MOODY, D. L.; SHANKS, G. G.; DARKE, P. Improving the quality of entity relationship models—experience in research and practice. In: SPRINGER. *International Conference on Conceptual Modeling*. [S.l.], 1998. p. 255–276.
- MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. [S.l.]: Editora Saraiva, 2017.
- PAPASTERGIOU, M. Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Computers & education*, Elsevier, v. 52, n. 1, p. 1–12, 2009.
- PATIG, S. Measuring expressiveness in conceptual modeling. In: SPRINGER. *International Conference on Advanced Information Systems Engineering*. [S.l.], 2004. p. 127–141.
- POELS, G. Conceptual modeling of accounting information systems: A comparative study of rea and er diagrams. In: SPRINGER. *International Conference on Conceptual Modeling*. [S.l.], 2003. p. 152–164.
- POELS, G.; MAES, A.; GAILLY, F.; PAEMELEIRE, R. The pragmatic quality of resources-events-agents diagrams: an experimental evaluation. *Information Systems Journal*, Wiley Online Library, v. 21, n. 1, p. 63–89, 2011.
- QUE, X.; CHEN, R.; LIU, J.; LU, C.; WU, C. Spatiotemporal data model for geographical process analysis with case study. *International Symposium on Parallel and Distributed Computing*, n. 15, p. 390–394, 2016.
- ROY-HUBARA, N.; ROKACH, L.; SHAPIRA, B.; SHOVAL, P. Modeling graph database schema. *IT Professional, IEEE*, v. 19, n. 6, p. 34–43, 2017.
- SARIS, W. E.; GALLHOFER, I. N. *Design, evaluation, and analysis of questionnaires for survey research*. [S.l.]: John Wiley & Sons, 2014.

SARKAR, A.; GANGULY, R.; CHATTERJEE, R. Graph semantic based approach for querying semi-structured database system. In: *22nd International Conference on Software Engineering and Data Engineering*. [S.l.: s.n.], 2013. p. 79–84.

SCHILDGEN, J.; LOTTERMANN, T.; DESSLOCH, S. Cross-system nosql data transformations with notaql. In: *ACM. Proceedings of the 3rd ACM SIGMOD Workshop on Algorithms and Systems for MapReduce and Beyond*. [S.l.], 2016. p. 5.

SHAW, M. The impact of modelling and abstraction concerns on modern programming languages. In: *On conceptual modelling*. [S.l.]: Springer, 1984. p. 49–83.

SHIN, K.; HWANG, C.; JUNG, H. Nosql database design using uml conceptual data model based on peter chen's framework. *International Journal of Applied Engineering Research*, v. 12, n. 5, p. 632–636, 2017.

SJØBERG, D. I.; DYBÅ, T.; ANDA, B. C.; HANNAY, J. E. Building theories in software engineering. In: *Guide to advanced empirical software engineering*. [S.l.]: Springer, 2008. p. 312–336.

TANG, E.; FAN, Y. Performance comparison between five nosql databases. In: *IEEE. 2016 7th International Conference on Cloud Computing and Big Data (CCBD)*. [S.l.], 2016. p. 105–109.

THALHEIM, B. *Entity-relationship modeling: foundations of database technology*. [S.l.]: Springer Science & Business Media, 2013.

THURM, B. Einsatz von nosql-datenbanksystemen für geodaten. *Bachelorarbeit Hochschule für Technik und Wirtschaft Dresden*, 2012.

TOTH, R. M. Abordagem nosql-uma real alternativa. *Sorocaba, São Paulo, Brasil: Abril*, v. 13, 2011.

UEBERSAX, J. S. A generalized kappa coefficient. *Educational and Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 42, n. 1, p. 181–183, 1982.

VAISH, G. *Getting started with NoSQL*. [S.l.]: Packt Publishing Ltd, 2013.

VERA, H.; BOAVENTURA, M. H. W.; GUIMARAES, V.; HONDO, F. Data modeling for nosql document-oriented databases. In: *CEUR Workshop Proceedings*. [S.l.: s.n.], 2015. v. 1478, p. 129–135.

VILLARI, M.; CELESTI, A.; GIACOBBE, M.; FAZIO, M. Enriched er model to design hybrid database for big data solutions. In: *IEEE. 2016 IEEE Symposium on Computers and Communication (ISCC)*. [S.l.], 2016. p. 163–166.

WOHLIN, C.; RUNESON, P.; HÖST, M.; OHLSSON, M. C.; REGNELL, B.; WESSLÉN, A. *Experimentation in software engineering*. [S.l.]: Springer Science & Business Media, 2012.

ZEČEVIĆ, I.; BJELJAC, P.; PERIŠIĆ, B.; STANKOVSKI, S.; VENUS, D.; OSTOJIĆ, G. Model driven development of hybrid databases using lightweight metamodel extensions. *Enterprise Information Systems*, Taylor & Francis, v. 12, n. 8-9, p. 1221–1238, 2018.

ZHENG, Z.; DU, Z.; LI, L.; GUO, Y. Bigdata oriented open scalable relational data model. In: IEEE. *2014 IEEE International Congress on Big Data*. [S.l.], 2014. p. 398–405.

APÊNDICE A – SCRIPT DO EXPERIMENTO

Aqui é detalhado como foi executado o experimento descrito nesta dissertação. Primeiramente, foi apresentado o questionário de perfil aos participantes, sendo explicando que o principal objetivo desse questionário seria obter informações a respeito das experiências dos participantes com a área do estudo, a modelagem conceitual de dados. Os participantes foram divididos em dois grupos: os que participaram presencialmente do estudo e aqueles que participaram remotamente. Primeiro, o experimento foi executado com os participantes presenciais, a escolha de se realizar primeiro com esses participantes se deve ao fato de que qualquer problema durante a execução das atividades poderia ser relatada ao experimentador, e esse poderia auxiliar de forma mais direta o participante. Assim, repensando os problemas ocorridos na fase presencial e elaborando um novo script para evitar que as dúvidas ou problemas ocorridos na fase presencial, acontecessem na fase remota, pois seria mais difícil do pesquisador auxiliar os participantes em caso de dúvidas. Por essa razão foi executado presencialmente primeiro, para relatar possíveis dúvidas e saber-se como lidar com elas da melhor forma possível.

A execução presencial ocorreu em dois dias, cada dia com participantes diferentes. Em cada dia foi sorteado qual seria a primeira unidade experimental a ser utilizada pelos participantes: *e-Commerce* ou médico. Além disso, foi sorteado de forma individual qual seria o primeiro tratamento dado a esses participantes: COMN ou NoAM. Após, a definição da unidade experimental e tratamento. Foi explanado aos participantes a primeira atividade do estudo, esta foi uma atividade de compreensão dos modelos estudados. A realização dessa atividade consiste no preenchimento de um questionário com perguntas voltadas para a compreensão do participante do modelo, junto com esse questionário foi dado um esquema de dados construído no modelo selecionado, sendo explicado ao participante que esse deveria observar o esquema antes de responder o questionário.

Após a conclusão da primeira atividade foi explicado aos participantes, que esses iriam estender o esquema dado a eles na primeira atividade realizada no estudo, mas que essa atividade seria simples, onde seria adicionado apenas algumas novas propriedades no esquema já dado na primeira atividade. Antes de sua execução, foi apresentado os construtores do modelo utilizado, além disso, foi dado a cada participante um documento, onde nesse explicava todos os construtores do modelo e, esse documento permaneceu com os participantes durante a execução da atividade. Ao fim da explicação do modelo, foi passado aos participantes um problema de domínio, para que os participantes realizassem a adição desses novos elementos no esquema. Feita a adição dos novos elementos solicitados, foi pedido que o participante respondesse um questionário, o intuito deste questionário é obter dados a respeito da usabilidade e representatividade dos modelos estudados.

A conclusão da primeira unidade experimental se deu a partir da finalização do questi-

onário dado na segunda atividade. Foi explanado aos participantes que, esses realizariam os mesmos passos novamente, apenas mudando a unidade experimental, e o tratamento, se esse foi o COMN, seria usado o NoAM e vice e versa.

A execução com os participantes que executaram as atividades remotamente ocorreu da mesma forma que a presencial, sendo sorteado a ordem da unidade experimental usada e o tratamento. Foi marcado um horário com cada participante para que esses realizassem as atividades e para que o experimentador estivesse disponível em caso de dúvidas. As atividades foram enviadas por email. O corpo do email continha todas as informações necessárias para a execução da primeira unidade experimental, sendo ressaltado no email que em caso de qualquer dúvida como proceder. O participante deveria contactar o experimentador, que esse tentaria resolver por mensagens o questionamento do participante, caso essa não fosse sanada. Seria realizada uma videochamada para explicar a dúvida que esse participante possuísse. Após a conclusão da primeira unidade experimental, o mesmo processo foi realizado com a segunda unidade experimental.

Os esquemas construídos pelos participantes serão avaliados por dois especialistas, essa avaliação será feita de forma individual. Cada especialista avaliará todos os esquemas construídos, essa avaliação possuirá um questionário, que serve como um *checklist* para o especialista, onde são feitos questionamentos se o esquema construído atende a algumas premissas, todos os questionamentos feitos nesse questionário possuem um grau de intensidade 3 (atendeu Totalmente à de forma alguma), exceto um questionamento possui intensidade 2, sendo esse de resposta: sim ou não. Os questionários possuem um campo ID e tratamento, sendo esse apenas utilizado para identificar se o participante fez parte do grupo que executou as atividades presencialmente ou remotamente e identificar o tratamento avaliado em cada em questionário.

Após a conclusão de todos os questionários, esse foram analisados. Os questionários respondidos pelos participantes foram aplicados na escala de likert, pois esse tipo de escala é habitualmente usada em questionários onde se quer saber a opinião de certo grupo de pessoas sobre determinado assunto. Ao analisar as respostas dadas nos questionários avaliou-se o nível de concordância dos participantes com os questionamentos. Assim, obtendo-se informações a respeito da usabilidade, expressividade e nível de compreensão dos participantes em relação aos modelos estudados.

Enquanto os questionários aplicados aos participantes tiveram o intuito de obter informações sobre a opinião dos mesmos sob tópicos específicos. O questionário ou *checklist* aplicado aos especialistas teve o objetivo de saber se os esquemas construídos nos modelos estudados eram completos, corretos, de acordo com o que foi solicitado na proposta de extensão dos esquemas. Além de analisar se os esquema construídos são independentes de paradigma de programação. Nesse questionário foi aplicado o teste Kappa, onde foi analisado o coeficiente de concordância dos dois especialistas. Os resultados obtidos do estudo foram apresentados no Capítulo 5 da presente dissertação.

APÊNDICE B – QUESTIONÁRIO DE PERFIL

1. Endereço de e-mail
2. Data de nascimento
3. Possui experiência com modelagem de dados? sim não
4. Se a resposta para a questão anterior for positiva, quantos anos de experiência possui.
5. Costuma utilizar algum modelo de dados para modelar algum problema de domínio em banco de dados? Se sim, qual (is)?
6. No projeto de banco de dados, costuma utilizar sistemas NoSQL para o armazenamento de dados? sim não
7. Já construiu algum esquema lógico de dados para sistemas NoSQL, que agrupasse por exemplo dados relacionados em um único objeto?
8. Você acredita que os construtores do (s) modelo (s) citado (s) por você na questão 5 conseguiria (m) expressar um esquema de dados para um sistema de banco de dados NoSQL? sim talvez não
9. Se a resposta anterior for positiva. Justifique o porquê dela?
10. Você acredita que a fase de concepção do projeto de banco de dados, isto é, a construção de esquemas de dados lógicos para determinado domínio auxiliam na visualização de como os dados são organizados? sim não
11. Se a resposta anterior for negativa, justifique o porquê da mesma.

APÊNDICE C – QUESTIONÁRIO DE COMPREENSÃO DOS MODELOS DE DADOS

1. O modelo de dados utilizado ajudou a compreender o domínio? Ajudou bastante Ajudou Ajudou em quase nada Não ajudou de maneira alguma
2. O diagrama entregue é claro quanto à organização de dados por ele modelado? Claríssimo Claro Indiferente Pouco claro Confuso
3. O esquema de dados produzido torna mais simples a visualização do domínio estudado? Bastante Muito Neutro Pouco De forma alguma
4. Se a resposta anterior for (pouco ou de forma alguma), o que você acha que deveria ser alterado no modelo, para tornar os esquemas de dados modelados por ele mais claros?
5. Os construtores do modelo de dados são intuitivos? Muito intuitivos Intuitivos Neutro Pouco Intuitivos Vago
6. A aplicação dos construtores de modelagem do modelo auxiliam na visualização de como os objetos do domínio se relacionam? Ajudam Bastante Ajudam Nem dificultam e nem ajudam Ajudam pouco Não ajuda de forma alguma
7. Se sua resposta anterior for a opção "Não ajuda de forma alguma", o que poderia ser acrescentado ao modelo para ajudar na visualização dos relacionamentos entre os objetos do domínio.

APÊNDICE D – QUESTIONÁRIO DE USABILIDADE E EXPRESSIVIDADE DOS MODELOS DE DADOS

1. Como foi a sua experiência ao utilizar o modelo de dados? Ótimo Boa Razóavel Ruim Péssima
2. O modelo escolhido é de fácil uso? sim não
3. Teve alguma dificuldade ao utilizar o modelo? Bastante Muita Pouca Quase nenhuma Nenhuma
4. Se na resposta anterior foi apresentada alguma dificuldade, explique qual foi esta?
5. Teve alguma dificuldade para saber qual construtor do modelo seria mais adequado para uma situação? Bastante Muita Pouca Quase nenhuma Nenhuma
6. Caso tenha apresentado alguma dificuldade explique qual foi esta e porque.
7. Houve dificuldade para a escolha de um ou outro construtor do modelo para representar algum aspecto do domínio escolhido? Bastante Muita Pouca Quase nenhuma Nenhuma
8. Se a resposta anterior foi Bastante ou muita, explique qual foi a dificuldade encontrada.
9. Os construtores dispostos no modelo são claros quanto ao que cada construtor representa? Claríssimo Claro Normal Pouco claro Nada claro
10. Você utilizaria este modelo novamente? Com certeza Provavelmente Talvez Provavelmente Não
11. Os elementos gráficos definidos no modelo são sugestivos? Bastante sugestivos Sugestivos Medianos Pouco sugestivos nada sugestivos
12. Na sua opinião, os construtores do modelo selecionado conseguem expressar todas as características necessárias do domínio abordado? Sim,completamete Sim, parcilmente Atende, mas outro se adequaria melhor Não atende
13. Se a sua resposta a questão anterior foi "atende, mas outro se adequaria melhor", justifique o porquê. Que modelo seria e quais características os construtores deste modelo possuem que o estudado não têm para atender esse domínio?
14. Se a resposta da questão 12 foi não atende. Explique porque o modelo não atende e o que esse não atende.

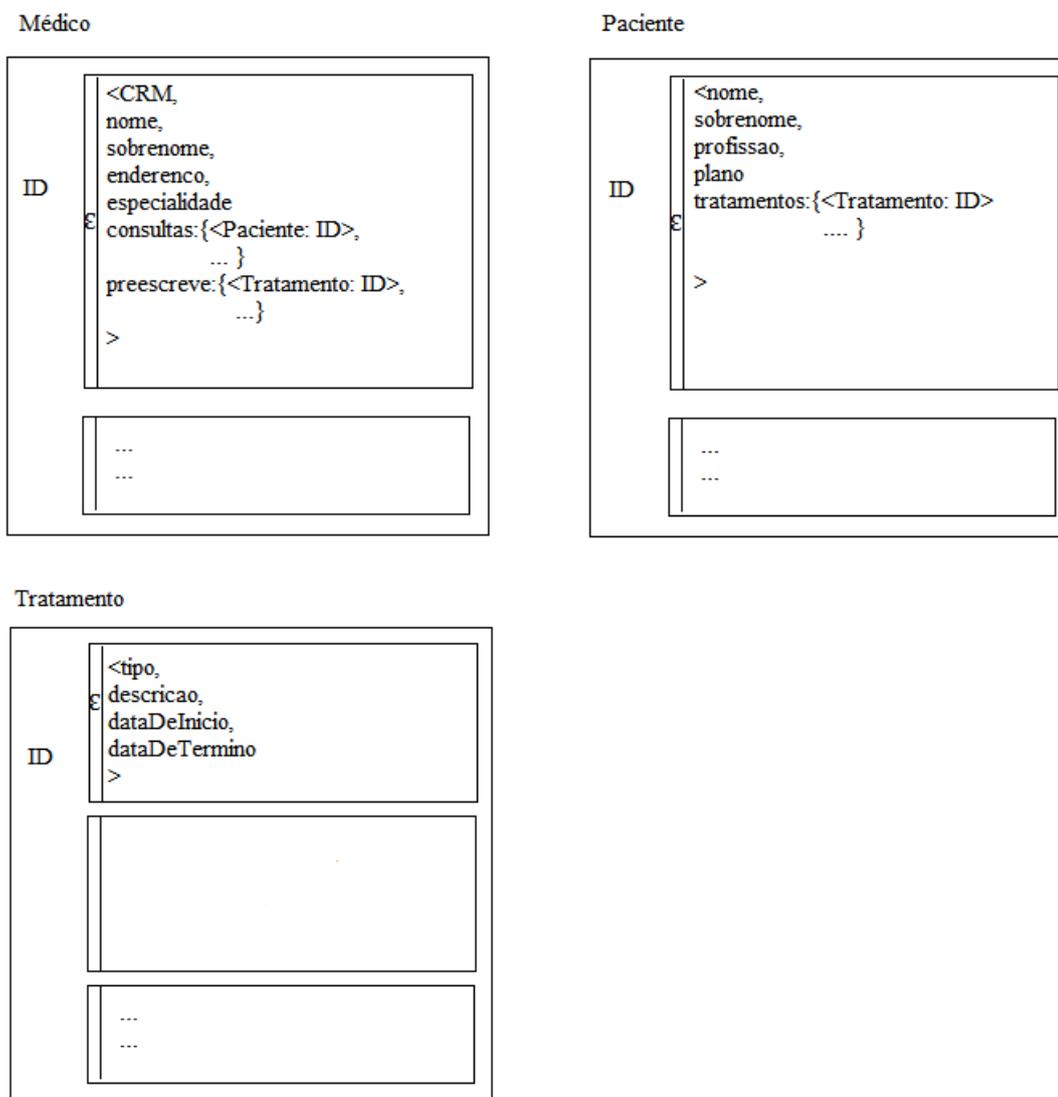
15. A respeito dos construtores de modelagem de relacionamentos do modelo, estes são capazes de expressar qualquer relacionamento entre objetos de um domínio NoSQL?
() Com Certeza () Sim, parcialmente () Talvez () Não ()
16. Se a resposta anterior foi não. Cite um exemplo de relacionamento NoSQL em que os construtores do modelo estudado não conseguem modelar e explique o porquê?

APÊNDICE E – QUESTIONÁRIO À RESPEITO DA QUALIDADE DOS MODELOS DE DADOS

1. Todas as propriedades relevantes do domínio da aplicação utilizado no experimento estão definidas no modelo? Totalmente Parcialmente De forma alguma
2. O diagrama desenvolvido está de acordo com o requerido no estudo? Totalmente Parcialmente De forma alguma
3. Há alguma característica implementada no diagrama é redudante? Sim, mais de uma Sim, apenas uma Não
4. Alguma propriedade foi definida no diagrama sem necessidade? Sim, mais de uma Sim, apenas uma Não
5. Alguma requisição do domínio não foi incluída no diagrama? Sim, mais de uma Sim, apenas uma Não
6. Os construtores do modelo conceitual estão sendo usados corretamente? Totalmente Parcialmente De forma alguma
7. Os relacionamentos dispostos no modelo estão sendo usados corretamente? Totalmente Parcialmente De forma alguma
8. Alguma regra do domínio foi expressa no diagrama erroneamente? Sim, mais de uma Sim, apenas uma Não
9. O diagrama apresenta alguma característica de algum paradigma de armazenamento? sim não
10. O modelo conceitual exposto é independente de ferramenta de implementação? Totalmente Parcialmente De forma alguma

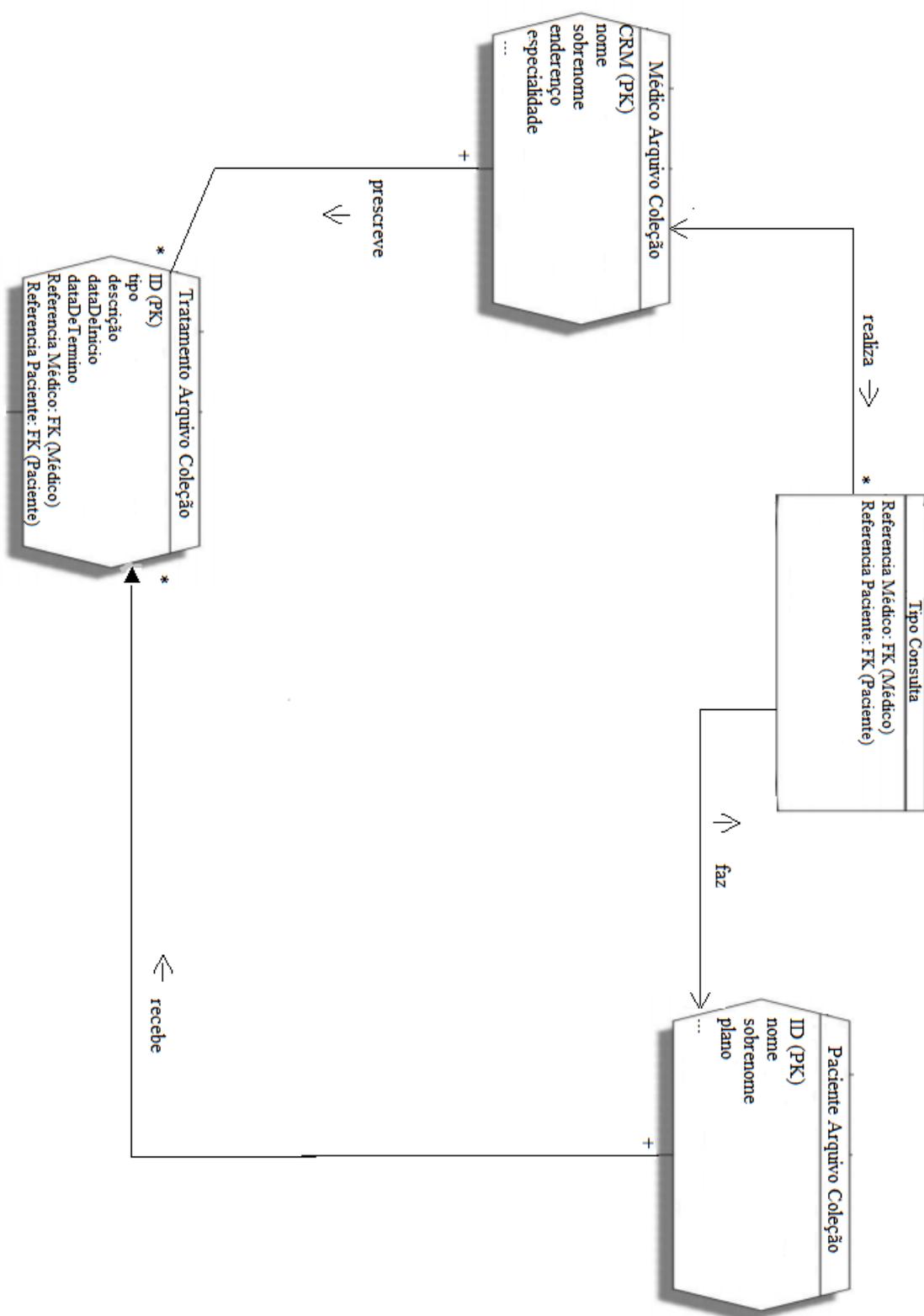
APÊNDICE F – ESQUEMAS ABORDADOS

Figura 13 – Domínio hospitalar representado no NoAM



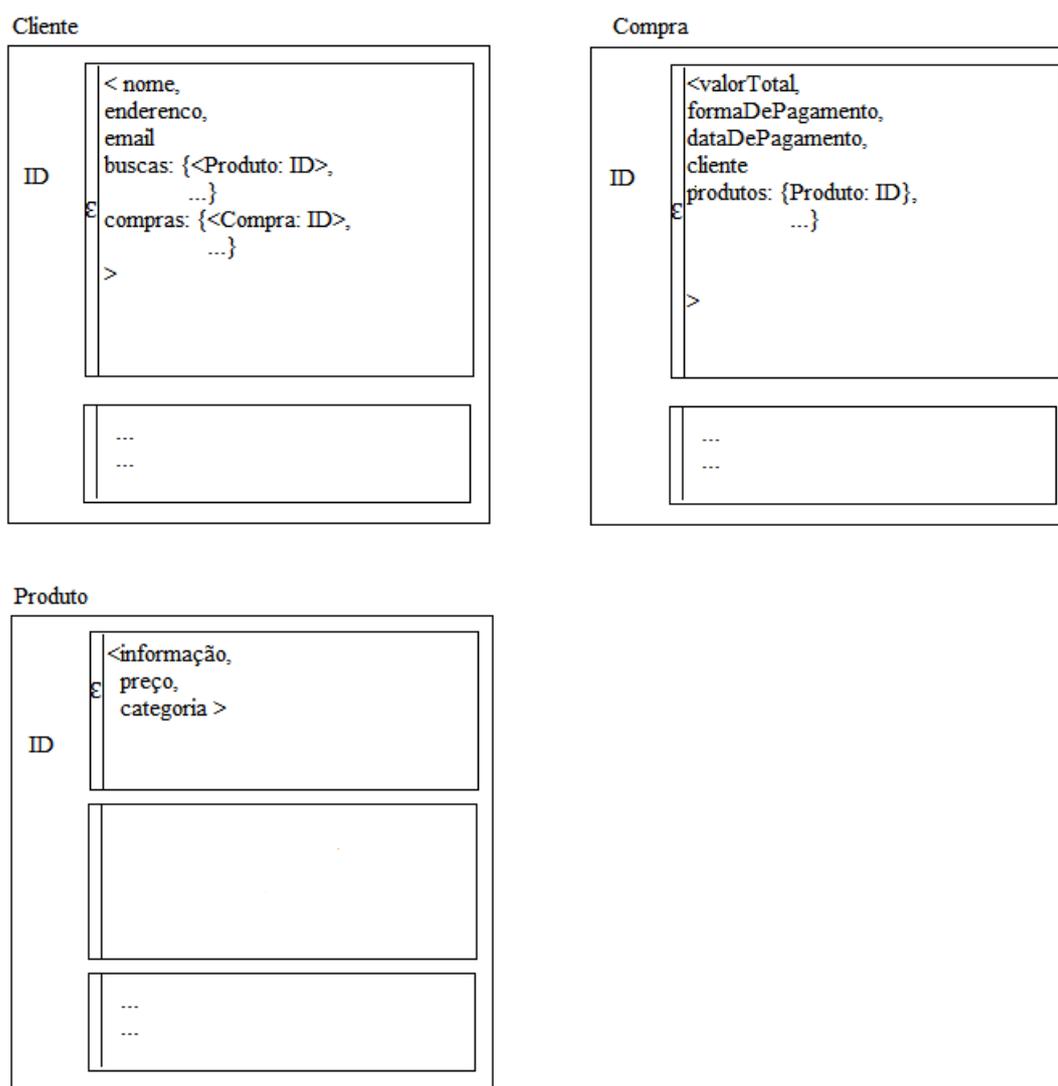
Fonte: O autor (2019)

Figura 14 – Domínio de hospitalar representado no COMN



Fonte: O autor (2019)

Figura 15 – Comércio eletrônico representado no NoAM



Fonte: O autor (2019)

Figura 16 – Comércio eletrônico representado no COMN

