UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Vitória Maria da Silva Maciel

Um Modelo de Suporte para Conformidade de Data Lake com a LGPD

Recife
2022

Vitória Maria da Silva Maciel

Um Modelo de Suporte para Conformidade de Data Lake com a LGPD

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
Orientador (a): Profa. Dra. Bernadette Farias Lóscio
Coorientador (a): Prof. Dr. Marcelo Iury de Sousa Oliveira

Recife
2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

M152m Maciel, Vitória Maria da Silva
Um modelo de suporte para conformidade de data lake com a LGPD / Vitória Maria da Silva Maciel. – 2022.
92 f.: fig.

Orientadora: Bernadette Farias Lóscio.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2022.
Inclui referências e apêndice.

1. Banco de dados. 2. Rastreabilidade. 3. Metadados. I. Lóscio, Bernadette Farias (orientadora). II. Título.

025.04 CDD (23. ed.) UFPE - CCEN 2023-35

Vitória Maria da Silva Maciel

“Um Modelo de Suporte para Conformidade de Data Lake com a LGPD”

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. Área de Concentração: Banco de Dados.

Aprovado em: 10/03/2022.

BANCA EXAMINADORA

_____

Profa. Dra. Jessyka Flavyanne Ferreira Vilela
Centro de Informática/UFPE

_____

Profa.Dra. Priscilla Kelly Machado Vieira Azevêdo
Universidade Federal do Agreste de Pernambuco

_____

Profa.Dra. Bernadette Farias Lóscio
Centro de Informática/UFPE
(**Orientadora**)

Dedico esta dissertação a todas as pessoas que, de alguma forma, me ajudaram a chegar até aqui.

AGRADECIMENTOS

Palavras não são suficientes para descrever o alívio de finalmente poder enxergar uma luz no fim do túnel. O caminho foi longo e árduo, pautado por dezenas de empecilhos e muito sacrifício. Mas Deus esteve comigo até aqui, então o primeiro agradecimento é para Ele, por ter cuidado de meus entes queridos quando não pude estar presente.

Agradeço ao meu coorientador Prof. Dr. Marcelo Iury, por ter segurado minha mão durante essa jornada e ter me ajudado a trilhar um caminho possível. Foram muitas trocas importantes e este trabalho é só parte do terreno que cultivamos. Nunca tivemos a chance de nos conhecermos pessoalmente, mas o contato próximo supera isso. Muito obrigada por todo encorajamento, ensinamentos, parceria e especialmente as imagens de "super-hero pose".

Agradeço também à Karina e Ruan, que me ajudaram na construção dessa dissertação. Além de apoio prático eles também foram meu ombro amigo nas dificuldades durante o processo. Espero ter saúde para retribuí-los, torcendo muito para vocês serem os próximos.

Outro agradecimento emocionado vai para Raissa e Ana Carolina, profissionais que eu tive muita sorte em encontrar. Até aqui elas me ajudaram a estar com o corpo e cabeça no lugar. Especialmente Raissa quem acompanha o processo há mais tempo, muito obrigada por tanto empenho e acolhimento. Ambas fazem parte desse resultado, muito mais do que imaginam. Gratidão pelo suporte emocional.

Agradeço à minha mãe Jeanne, por me encorajar a persistir sempre e por ter suportado tanta ausência. Agradeço também à minha outra mãe Helena, por todo carinho e cuidado mesmo a distância. Vocês refletem tudo o que tenho como inspiração na vida: Ser do bem, ter coragem, fazer o trabalho bem feito e com vontade, fazer a diferença todos os dias. Muito obrigada!

Agradeço à Rodrigo por todo incentivo, à Arthur por ter feito a jornada em alguns momentos mais leve, e também à meus companheiros de graduação, Pedro, Valdi e Neto.

Agradeço ao casal JJ, Vivi e Wagner, que sempre estiveram presentes, aguentando por muitas vezes o deságue de problemas em todas as ordens. Juntos, vocês dividiram comigo o peso, que só quem vive sabe, e por mais que eu agradeça, nada traduz minha imensa gratidão.

Agradeço aos meus amigos do 006 + Liz. Foram 3 anos me ouvindo falar disso aqui repetidamente, e em todas as vezes oferecendo como resposta palavras amigas e de motivação. Somente 10 anos de amizade suportam isso.

Por fim, agradeço à minha orientadora Profa. Dra. Bernadette, por toda compreensão, apoio e confiança. Nossos momentos de aprendizado ficarão guardados com carinho.

RESUMO

Na era do Big Data, um grande volume de dados estruturados, semi-estruturados, e principalmente não estruturados é gerado muito mais rápido por tecnologias digitais e sistemas de informação. Neste contexto, Data Lakes surgiram como uma alternativa aos tradicionais Data Warehouses, tornando-se uma das soluções de Big Data mais utilizadas para análise e gerenciamento distribuído de grande volumes de dados. A ideia principal do Data Lake é ingerir dados brutos e processá-los durante seu uso, caracterizando a abordagem *schema on-read*. Durante seu ciclo de vida em um Data Lake, um dado pode passar por inúmeras transformações, levando a questões de rastreabilidade. Com a Lei Geral de Proteção de Dados Pessoais - LGPD em vigor, as organizações precisam ter ao seu dispor, além das mudanças ocorridas nos dados, informações sobre quem modificou, onde modificou e as dependências geradas. Visando atender esse problema, alguns modelos de metadados foram propostos na literatura. No entanto, nenhum deles foca em apresentar metadados que descrevam o ciclo de vida dos dados. Sendo assim, essa dissertação propõe um Modelo de Suporte para Conformidade de Data Lake com a LGPD (Data Lake Compliance Model - DLCM), que tem como objetivo descrever os conjuntos de dados no Data Lake e os tratamentos aplicados sobre eles. Para isso, o DLCM subdivide-se em duas partes: A primeira reúne todos os elementos de metadados necessários para atendimento de uma solicitação de acesso aos dados, enquanto que a segunda parte, é composta pelo agrupamento desses metadados por categorias, onde cada categoria possui um modelo associado. Os resultados obtidos a partir da avaliação do DLCM mostraram a relevância da solução proposta no contexto de Data Lakes.

Palavras-chaves: data lakes; LGPD; rastreabilidade; metadado; conformidade.

ABSTRACT

In the age of Big Data, a large volume of structured, semi-structured, and mostly unstructured data is generated much faster by digital technologies and information systems. In this context, Data Lakes emerged as an alternative to traditional Data Warehouses, becoming one of the most used Big Data solutions for distributed analysis and management of large volumes of data. The main idea of Data Lake is to ingest raw data and process it during its use, characterizing the schema on-read approach. During its life cycle in a Data Lake, data can undergo numerous transformations, leading to traceability issues. With the General Personal Data Protection Law - LGPD in place, organizations need to have at their disposal, in addition to the changes that have occurred, information about who modified the data, where they modified it and the dependencies generated. In order to address this problem, some metadata models have been proposed in the literature. However, none of them focus on presenting metadata that describes the data life cycle. Therefore, this dissertation proposes a Support Model for Data Lake Compliance with the LGPD (Data Lake Compliance Model - DLCM), which aims to describe the datasets in the Data Lake and the treatments applied to them. For this, the DLCM is subdivided into two parts: The first part gathers all the metadata elements necessary to fulfill a data access request, while the second part is composed by the grouping of these metadata by categories, where each category has an associated model. The results obtained from the DLCM evaluation showed the relevance of the proposed solution in the context of Data Lakes.

Keywords: data lakes; LGPD; traceability; metadata; compliance.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ADLA	Azure Data Lake Analytics
ADLS	Azure Data Lake Store
AIMQ	Assessment Information Methodology Quality
AWS	Amazon Web Services
CSV	Comma Separated Value
DW	Data Warehouse
GDPR	General Data Protection Regulation
HDFS	Hadoop Distributed File System
HTML	HyperText Markup Language
IoT	Internet of Things
JSON	JavaScript Object Notation
LGPD	Lei Geral de Proteção de Dados Pessoais
PDF	Portable Document Format
SQL	Standard Query Language
XLS	Microsoft Excel file format
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

Este Capítulo fornece uma visão geral desta pesquisa e apresenta o contexto no qual este trabalho está inserido. A Seção 1.1 apresenta uma breve motivação para o desenvolvimento deste trabalho. A caracterização do problema é descrita na Seção 1.2 e os objetivos e contribuições são apresentados na Seção 1.3. A Seção 1.4 discorre sobre a método de pesquisa usado e, por fim, a Seção 1.5 apresenta a estrutura desta dissertação.

1.1 MOTIVAÇÃO

Diariamente, uma massiva quantidade de dados é gerada por tecnologias digitais e sistemas de informação. Com a ascensão dos dados na geração de valor, muitas organizações tem buscado adaptar e melhorar sua infraestrutura de TI para processar esses dados e gerar informações úteis para a tomada de decisão (MILOSLAVSKAYA; TOLSTOY, 2016). Aos poucos, os tradicionais sistemas de banco de dados se mostraram não tão eficientes para o processamento de grandes volumes de dados, abrindo espaço para uma nova abordagem denominada Data Lake.

Segundo (NARGESIAN et al., 2019), um Data Lake pode ser definido como uma coleção massiva de conjuntos de dados que: 1) podem ser hospedados em sistemas de armazenamento distintos; 2) podem variar em seus formatos; 3) podem não ter nenhum metadado ou catálogo de dados associado; e 4) podem mudar de forma autônoma ao longo do tempo.

Em contraste com os tradicionais sistemas de suporte à decisão (a exemplo, Data Warehouses), Data Lakes geralmente são construídos para lidar com grandes volumes de dados não estruturados que chegam rapidamente. Outra característica importante que o difere do Data Warehouse é a flexibilidade na ingestão e disponibilização dos dados. Já que a estrutura é definida mediante o uso através da abordagem *schema on read*, logo após a ingestão, os dados já estão disponíveis para consumo, evitando a etapa complexa de modelagem e o esforço de integração de dados (FANG, 2015).

No entanto, essa flexibilidade promovida pela ausência de um esquema pode acabar transformando o Data Lake em um pântano de dados (*data swamp*) inacessível e incompreensível pelos usuários. Sem modelos ou descrições associadas, não é possível descobrir facilmente quais dados foram adicionados no Lake, ou como foram adicionados. Os processos de ingestão e manutenção ficam comprometidos, pois não há como evitar a repetição de dados sem saber quais deles já foram inseridos, nem se estão corretos ou completos. Logo, para garantir um funcionamento adequado e cumprimento de seu propósito em gerar valor através dos dados para a organização, Data Lakes precisam incorporar sistemas para gerenciamento de metadados que forneçam descrições apropriadas sobre os dados armazenados.

Tipicamente, Data Lakes reúnem dados operacionais (vendas, finanças, estoque), dados gerados automaticamente (dispositivos IoT, logs) e/ou dados gerados por humanos (postagens

de mídia social, e-mails, conteúdo da web) vindos de dentro ou de fora da organização. Esses dados serão eventualmente objeto de análise, mas os riscos de privacidade não são desprezíveis, pois Data Lakes também estão sujeitos a violações da privacidade de dados, tais como vazamento de dados, uso indevido e abuso de informações privadas. Em resposta, diversos países e organizações multi-nacionais editaram normas que estabelecem controles de privacidade e segurança para rastrear, bloquear ou restringir o acesso a dados pessoais.

A norma mais conhecida internacionalmente de proteção a dados pessoais é a GDPR (*General Data Protection Regulation*). No Brasil, em 2018, foi promulgada a LGPD (Lei Geral de Proteção de Dados Pessoais). Estas legislações visam fornecer mais proteção e recursos aos indivíduos para controle de seus dados pessoais. Por esta razão, as organizações foram obrigadas a reformularem o modo como abordam a gestão de dados pessoais armazenados e tratados durante a execução de seus processos diários, para alcançar a conformidade com as leis (AGOSTINELLI et al., 2019).

Referindo-se a LGPD especificamente, seu propósito é servir como um mecanismo para regular o tratamento de dados pessoais em território nacional. Em seu regulamento, ela define como tratamento de dados, qualquer operação realizada com dados pessoais, tais como: coleta, produção, transmissão, dentre outros. Além disso, são caracterizados três atores: (i) o Titular dos Dados, pessoa natural a quem se referem os dados pessoais que são objetos de tratamento, (ii) o Controlador, agente responsável por tomar as decisões acerca dos tratamentos, e o (iii) Operador, que é a figura encarregada de realizar os tratamentos em nome do Controlador.

Além de regulamentar as diretrizes sobre como as organizações devem lidar com dados pessoais, a LGPD também assegura o direito dos Titulares dos Dados. O direito de acesso aos dados é um dos tipos de solicitação que o Controlador pode receber do Titular. Esta solicitação pode ser feita em formato simplificado, de forma imediata ou por meio de uma declaração mais completa. Quando feita por meio de declaração, o Controlador tem um prazo de até 15 (quinze) dias para atender ao solicitante, contados a partir da data do requerimento. Em ambos os formatos é exigido que o Controlador dos dados informe quais dados possui relacionados ao indivíduo e quais tratamentos foram aplicados sobre eles.

Em Data Lakes, a ingestão dos dados ocorre inicialmente sem a preocupação de rotulagem ou auditorias (COTTRELL, 2020). Não há informações sobre os fins para quais os dados foram extraídos, e além disso, estes dados podem passar por diversos *workflows* em ferramentas distintas e tratamentos para diferentes finalidades. Para alcançar a conformidade com a LGPD, é necessário que os Data Lakes possuam informações de toda linhagem dos dados armazenados e processados, comprovando que é possível identificar onde eles podem ser encontrados e como eles fluem e são usados dentro do ambiente.

Neste contexto, o gerenciamento de metadados executa um papel fundamental, não somente para suporte a conformidade com a Lei, mas também para funcionamento do Data Lake em geral. Um aspecto importante do gerenciamento de metadados são os modelos de metadados, que podem ser utilizados tanto para representação de domínios quanto para o

desenvolvimento de métodos automatizados que verificam a conformidade com a legislação.

Alguns modelos foram propostos na literatura, como por exemplo o de (UJCICH; BATES; SANDERS, 2018), que representa os principais conceitos do GDPR e sobre como a proveniência de dados pode ser aplicada para alcance da conformidade. No domínio de Data Lakes, Zhao, Megdiche e Ravat (2021) propõe um modelo de metadados que foca na ingestão de dados, provendo informações básicas acerca dos conjuntos de dados que auxiliam os usuários na compreensão de sua estrutura e conteúdo. Mesmo apresentando metadados importantes, os modelos não incorporam metadados sobre os tratamentos que podem ser aplicados nos conjuntos de dados. Além disso, também não foram encontrados modelos adaptados do GDPR para representação da LGPD. Sendo assim, é nítida a necessidade de um modelo que guie a coleta de metadados essencialmente sobre os conjuntos de dados armazenados em um Data Lake, os tratamentos aplicados sobre eles e adicionalmente, represente os conceitos básicos da lei brasileira.

Sendo assim, esta Dissertação propõe um Modelo de Suporte para Conformidade de Data Lake com a LGPD (*Data Lake Compliance Model* - DLCM), que tem como objetivo descrever quais metadados devem ser capturados acerca dos dados e dos tratamentos realizados sobre eles em um Data Lake. Estes metadados devem auxiliar o Controlador no retorno à solicitação de acesso aos dados efetuada pelo Titular dos dados. Além disso, o modelo também almeja dar suporte no gerenciamento de metadados no Data Lake, provendo informações sobre os conjuntos de dados para facilitar sua exploração e uso por usuários técnicos e de negócio.

1.2 CARACTERIZAÇÃO DO PROBLEMA

Com a crescente preocupação sobre proteção de dados e privacidade, é cada vez mais importante avaliar o cumprimento das legislações. Uma das motivações para o surgimento das regulamentações, GDPR e LGPD, é fornecer mais proteção e recursos aos indivíduos para controlarem seus dados pessoais em face de novos desenvolvimentos tecnológicos. Apesar de benéfica, a realidade é que, com estas novas leis de proteção, as empresas estão tendo dificuldades em compreender o que significa conformidade neste novo cenário e em como implementá-la (TORRE et al., 2019).

As organizações que descumprirem a LGPD poderão sofrer sanções administrativas aplicáveis pela Autoridade Nacional. Acarretando, assim, em multa de até cinquenta milhões de reais por infração, publicização da infração, suspensão parcial do funcionamento do banco de dados a que se refere a infração ou até mesmo suspensão do exercício da atividade de tratamento dos dados pelo período máximo de 6 meses (LEI..., 2018).

(UJCICH; BATES; SANDERS, 2018) mencionam uma pesquisa realizada em organizações afetadas pela GDPR, na qual mais de 50% delas acreditam que serão penalizadas por descumprimento e quase 70% pensam que o regulamento irá impactar nos custos de seus negócios. A mesma pesquisa ainda observou que as tecnologias analíticas foram consideradas extremamente necessárias para demonstrar que os dados pessoais foram armazenados e processados

de acordo com o consentimento fornecido pelos titulares dos dados.

Alcançar conformidade com a LGPD ou GDPR não é algo trivial. As organizações devem implantar mecanismos que rastreiem e gerenciem seus dados. No entanto, adaptar sistemas e processos já em uso há bastante tempo requer esforços e investimentos para associar noções jurídicas de alto nível com noções técnicas de baixo nível, incorporando as leis na prática à realidade operacional dos negócios (UJCICH; BATES; SANDERS, 2018). No mais, muitas organizações não têm ideia de quais informações pessoais elas possuem e como elas estão sendo combinadas (UJCICH; BATES; SANDERS, 2018).

Uma estratégia de gerenciamento de metadados eficiente é fundamental para a governança de dados e é também um elemento chave para conformidade, pois pode prover informações que respondem a questões sobre quais dados existem, em quais formatos, sua localização e uso. Além do mais, metadados são um requisito importante em soluções de Data Lakes para garantir integração, identificação e organização dos conjuntos de dados (ZGOLLI; COLLET; MADERA, 2020).

Sendo assim, um modelo que guie a coleta de metadados pode contribuir para garantir que o Controlador dos dados, proprietário do Data Lake, tenha a seu dispor os recursos necessários, no que diz respeito a metadados, para atendimento as solicitações efetuadas pelos Titulares de dados. Além disso, ele também servirá como fonte de consulta para usuários técnicos e de negócios, provendo informações sobre o quais conjuntos de dados estão disponíveis no Data Lake.

Ademais, esse trabalho teve seu desenvolvimento guiado pela seguinte questão de pesquisa: **Quais metadados devem ser coletados sobre os tratamentos realizados nos dados em um Data Lake que auxiliem o Controlador dos dados no atendimento às requisições da LGPD?**

1.3 OBJETIVOS

O principal objetivo desta Dissertação é propor um Modelo de Suporte para Conformidade de Data Lake com a LGPD. Este modelo busca descrever os dados armazenados no Data Lake e os tratamentos aplicados sobre eles. Esse modelo leva em consideração conceitos relacionados à: 1) Data Lakes, 2) leis de proteção de dados e 3) boas práticas de catalogação de conjuntos de dados existentes na literatura (*e.g.* (PANDIT; LEWIS, 2017), (UJCICH; BATES; SANDERS, 2018), (ZGOLLI; COLLET; MADERA, 2020), (CONSORTIUM et al., 2014), (ZHAO; MEGDICHE; RAVAT, 2021), (GAO; CHEN; DU, 2020), (LÓSCIO; BURLE; CALEGARI, 2016), (RAVAT; ZHAO, 2019b), (ZHAO et al., 2021), (MEGDICHE; RAVAT; ZHAO, 2021), (EICHLER et al., 2021), (SCHOLLY et al., 2021), (TORRE et al., 2020)), a fim de reunir todos os metadados necessários para dar suporte ao Controlador dos dados, no atendimento as requisições legais efetuadas pelos Titulares de dados.

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

- Levantamento dos modelos existentes na literatura que abordam as leis de proteção de dados (GDPR ou LPGD) ou Data Lakes;
- Identificação dos conceitos/constructos relacionados a conformidade com leis de proteção a dados pessoais;
- Avaliação do modelo proposto através de um questionário com especialistas.

1.4 MÉTODO DE PESQUISA

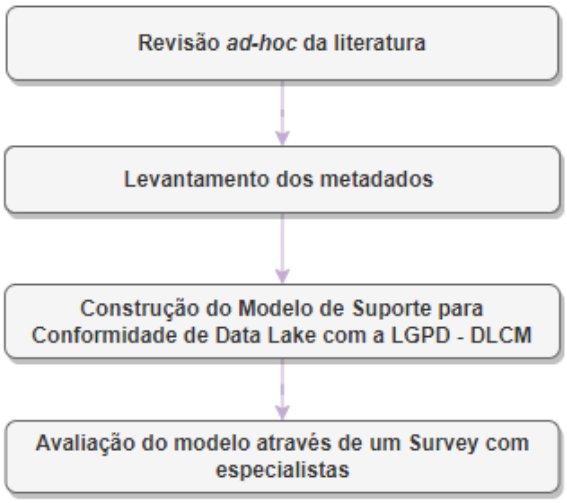
Uma vez que diferentes metodologias de pesquisa servem a diferentes propósitos, um dos primeiros passos é escolher um paradigma filosófico de pesquisa apropriado (EASTERBROOK et al., 2008). O paradigma filosófico diz respeito à fonte, natureza e desenvolvimento do conhecimento (CRESWELL, 2010). Em geral, afirmar a filosofia de pesquisa envolve estar ciente e formular as crenças e suposições da pesquisa. Nesse sentido, esta dissertação se baseia em um paradigma filosófico pragmático.

De acordo com (EASTERBROOK et al., 2008) e (CRESWELL, 2010), essa postura filosófica se caracteriza por aceitar diferentes conceitos para embasar a pesquisa. Em vez de focar nos métodos, o problema é mais importante, e os pesquisadores usam todos os meios para entender o problema.

Esta pesquisa também é baseada no paradigma *Design Science Research*, que visa construir e avaliar artefatos para resolver classes concretas de problemas relevantes usando métodos científicos rigorosos (HEVNER et al., 2008). O paradigma da *Design Science Research* está alinhado com a postura filosófica pragmática.

Em resumo, esta pesquisa foi realizada em quatro etapas, como mostra a Figura 1. Inicialmente foi realizada uma revisão *ad-hoc* da literatura para encontrar trabalhos relacionados ao nosso que pudessem nos fornecer um embasamento teórico. Uma revisão *ad-hoc* é uma busca informal para encontrar insumos de um determinado assunto ou área temática. Alguns dos trabalhos selecionados serão apresentados e discutidos no Capítulo 3.

Figura 1 – Etapas da Metodologia de Pesquisa.



Fonte: a autora, 2022

A segunda etapa é caracterizada pelo levantamento dos metadados. Com base nos trabalhos selecionados na revisão *ad-hoc*, utilizamos a técnica de mapa mental para elencar e evoluir todos os metadados necessários, classificando-os em categorias. Após a especificação dos metadados, a terceira fase teve como objetivo construir e melhorar continuamente o modelo proposto nesta dissertação. Essa etapa passou por vários ciclos de verificação e refinamento até ter como resultado uma versão inicial do Modelo de Suporte para Conformidade de Data Lake com a LGPD.

Por último, uma avaliação utilizando um questionário fundamentado no método *Assessment Information Methodology Quality* (AIMQ) foi realizada com especialistas. Baseado nos resultados obtidos nesta avaliação, o modelo foi refinado a fim de atender as melhorias propostas pelos participantes e sua versão final é apresentada no Capítulo 4.

1.5 ESTRUTURA DA DISSERTAÇÃO

Os próximos capítulos estão organizados da seguinte forma. No Capítulo 2, é apresentada a fundamentação teórica, enquanto que no Capítulo 3 descrevemos os modelos de metadados no domínio do regulamento europeu GDPR e Data Lakes. Já no Capítulo 4, é apresentado o modelo de Suporte para Conformidade de Data Lake com a LGPD proposto nesta dissertação. No Capítulo 5 é apresentado o método AIMQ utilizado para avaliação do modelo e os resultados obtidos. Por fim, no Capítulo 6 é feita uma pequena discussão sobre o trabalho realizado e sugestões para os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo, iremos apresentar os conceitos essenciais para o entendimento desta dissertação. Inicialmente, na Seção 2.1, apresentamos uma visão geral sobre o impacto das leis de proteção de dados em Data Lakes e sobre como metadados podem ser úteis nesse contexto. Na Seção 2.2 descrevemos as particularidades de Data Lakes, arquiteturas e tecnologias, enquanto que na Seção 2.3 apresentamos a Lei Geral de Proteção de Dados (LGPD). Em seguida, na Seção 2.4 conceituamos metadados e suas classificações. Na Seção 2.5 incluímos as definições de rastreabilidade e sua aplicabilidade no contexto de Data Lakes. Por fim, na 2.6 apresentamos as Considerações Finais do Capítulo.

2.1 VISÃO GERAL

Na era do Big Data, um grande volume de dados estruturados, semi-estruturados, e principalmente não estruturados, é gerado muito mais rápido do que antes por dispositivos conectados, mídias sociais e também por outros tipos de produtores de dados. Esses dados agregam valor para as organizações e seus sistemas de suporte à decisão, geralmente caracterizados como Data Warehouses (DW). No entanto, tais sistemas lidam constantemente com o desafio de processar dados heterogêneos e volumosos, e disponibilizá-los para análises em tempo hábil com performance.

Neste contexto, Data Lakes surgiram como uma alternativa aos tradicionais Data Warehouses, tornando-se uma das soluções de Big Data para análise e gerenciamento distribuído de grande volumes de dados (GIEBLER et al., 2021). Data Lakes são utilizados em vários domínios de negócios, como os setores de saúde, financeiro e dentre outros. Eles fornecem as organizações a capacidade de explorar seus dados por meio de análises avançadas, como *Machine Learning*. Para atender a este propósito, dados heterogêneos são armazenados em seu formato mais bruto, sem a pré definição de esquemas.

Segundo Zhao, Megdiche e Ravat (2021), a ideia principal do Data Lake é ingerir dados brutos e processá-los durante seu uso. Entretanto, Data Lakes que possuem muitos conjuntos de dados sem modelos ou descrições, podem facilmente se tornar incompreensíveis e inacessíveis. Logo, é de fundamental importância que Data Lakes incorporem sistemas de gerenciamento de metadados que forneçam informações acerca dos dados armazenados, facilitando assim, sua exploração.

Durante seu ciclo de vida em um Data Lake, um dado pode passar por inúmeras transformações através de diferentes sistemas. Isto é, em um estágio, um dado encontra-se em seu formato original, em outro estágio, este mesmo dado é tratado e refinado por ferramentas distintas. Estas modificações durante seu ciclo de vida reforçam a necessidade de mecanismos de rastreabilidade eficientes.

Rastreabilidade é a capacidade de rastrear elementos de um sistema ao longo de seu ciclo de vida (AZEVEDO; JINO, 2019). Uma das principais vantagens da rastreabilidade é que ela permite uma avaliação completa sobre as mudanças ocorridas e seus impactos, pois identifica não somente o que mudou, mas também quem modificou, onde realizou e as dependências geradas.

Com as novas leis de proteção de dados, as organizações passaram a reformular o modo como abordam a gestão de dados pessoais armazenados e tratados durante a execução de seus processos diários (AGOSTINELLI et al., 2019). Além disso, também são compelidas a responder quando solicitado, quais dados possuem sobre uma pessoa natural e quais foram os tratamentos aplicados sobre eles. Deste modo, metadados de rastreabilidade podem ser úteis para descrever as transformações realizadas no dados em um Data Lake. Nas próximas seções, discutiremos as particularidades de Data Lakes assim como os conceitos de metadados e rastreabilidade, relacionando-os com a Lei Geral de Proteção de Dados - LGPD e seus impactos.

2.2 DATA LAKES

O termo Data Lake foi introduzido pela primeira vez por James Dixon em 2010 (CHIHOUB et al., 2020). Em seu artigo inicial, Dixon descreve Data Lakes como sistemas que armazenam dados de uma única fonte em seu formato original. A medida que estes sistemas tornaram-se mais populares e utilizados, as pessoas passaram a compreender que na verdade, Data Lakes armazenam dados de inúmeras fontes e em diversos formatos.

O conceito de Data Lake possui diversas definições tanto no universo acadêmico quanto industrial. Apesar da diversidade, todas as definições existentes respeitam a ideia de que Data Lake é um repositório que armazena dados brutos em seu formato nativo (RAVAT; ZHAO, 2019a). Estes conceitos evoluíram ao longo do tempo a partir das experiências e feedback, incorporando novas características e funcionalidades.

Chihoub et al. (2020) afirma que um Data Lake deve ser capaz de obter dados de qualquer fonte, mas para evitar se transformar em um pântano de dados, requisitos mínimos de governança devem ser cumpridos, como por exemplo, o fornecimento de metadados que descrevam os conjuntos dados e o formato dos dados brutos. Ravat e Zhao (2019a) descrevem Data Lake como uma solução de Big Data que ingere dados heterogêneos de várias fontes, armazena estes dados brutos em seu formato nativo, permite processar dados com diferentes requisitos e disponibiliza-os para diferentes usuários (cientistas de dados, analistas, profissionais de *Business Inteligence*, etc) realizarem análises estatísticas ou de Aprendizado de Máquina, garantindo o controle, segurança e ciclo de vida dos dados.

Sawadogo e Darmont (2021) acrescenta que as características de um Data Lake também incluem:

- Um catálogo de metadados que reforça a qualidade dos dados;
- Ferramentas e políticas para governança de dados;

- Acessibilidade para vários tipos de usuários;
- Integração com qualquer tipo de dado;
- Organização física e lógica;
- Escalabilidade em termos de armazenamento e processamento.

Por fim, Sawadogo e Darmont (2021) concluíram que um Data Lake nada mais é do que um sistema de armazenamento e análise escalonável para dados de qualquer tipo, retidos em seu formato nativo e usado principalmente por especialistas em dados para extração do conhecimento.

2.2.1 Arquitetura de Data Lake

Nos últimos anos, Data Lakes (DL) emergiram como uma das soluções de Big Data para gerenciar grandes volumes de dados heterogenêos. No entanto, implementar um Data Lake na prática mostra-se algo bastante desafiador. Pois, apesar dos diversos trabalhos existentes na literatura e publicações comerciais, ainda é difícil encontrar uma arquitetura abrangente, visto que a maioria trata apenas aspectos parciais de arquitetura (GIEBLER et al., 2021).

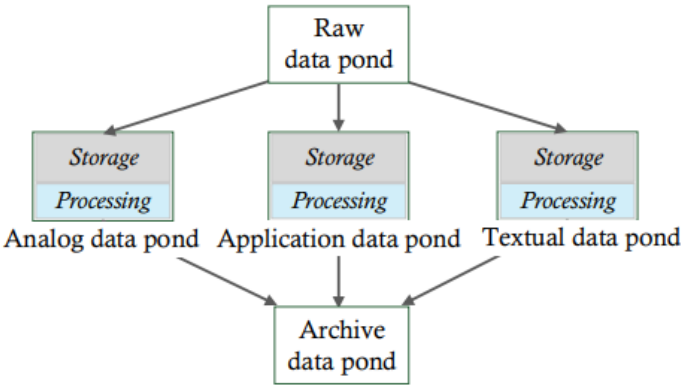
Sawadogo e Darmont (2021) mencionam dois tipos de arquitetura para DL, a arquitetura de lagoas (do inglês, *Ponds Architecture*) representada na Figura 2, e a de zonas (*Zone Architecture*) representada na Figura 3. Na arquitetura de lagoas, podemos considerar que um Data Lake é formado por um conjunto de *data ponds*, no qual cada um pode ser visto como um subconjunto que trata de um tipo especifico de dado. Já na chamada arquitetura de zonas, os dados são atribuídos a uma zona de acordo com seu grau de refinamento.

Na arquitetura de lagoas, são elencadas 5 subdivisões, cada uma delas é associada a um sistema de armazenamento, a algum tipo de processamento específico nos dados e a um serviço de análise relevante. Na lagoa de dados brutos (*Raw data pond*), são armazenados e processados os dados brutos recém coletados. Também funciona como uma zona transiente, uma vez que, após um condicionamento, os dados são transferidos para outra lagoa de dados. Os dados armazenados na lagoa de dados analógica (*Analog data pond*), são caracterizados por uma alta frequência de medições, como por exemplo, dados de IoT (*Internet of Things*). Já na lagoa de dados de aplicação (*Application data pond*), são ingeridos dados estruturados provenientes de aplicações de software e banco de dados relacionais. NA lagoa de dados textuais (*Textual data pond*), são gerenciados dados não estruturados. E e por fim, na lagoa de arquivamento (*Archive data pond*), são armazenados os dados que não são usados ativamente mas que ainda podem ser necessários no futuro. A figura 2 demonstra como ocorre o fluxo de dados nessa arquitetura.

Na arquitetura de zonas apresentada na Figura 3, a zona transiente de carregamento dos dados (*Transient loading zone*), lida com todo o processo de ingestão. Pela zona de dados brutos (*Raw data zone*), percorrem todos os dados provenientes da zona transiente em seu

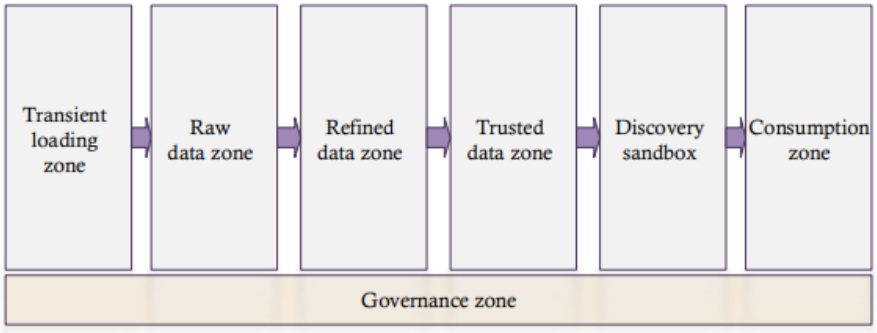
formato nativo. Na zona de refinamento (*Refined data zone*), os dados são tratados e enriquecidos. Na zona confiável (*Trusted loading zone*), os dados já encontram-se em agregados, em nível maior de granularidade, e posteriormente são movidos para a zona de descoberta (*Discovery sandbox*). Nesta última, especialistas em dados iniciam análises para descoberta de conhecimento e, na zona de consumo (*Consumption zone*), os dados estão disponíveis para uso tanto por novos usuários quanto aplicações. Por último, na zona de governança (*Governance zone*), os metadados são gerenciados, assim como o controle de qualidade, segurança, ciclo de vida e catálogo dos dados.

Figura 2 – Fluxo de dados em uma arquitetura de lagoa.



Fonte: (SAWADOGO; DARMONT, 2021)

Figura 3 – Arquitetura de zonas.

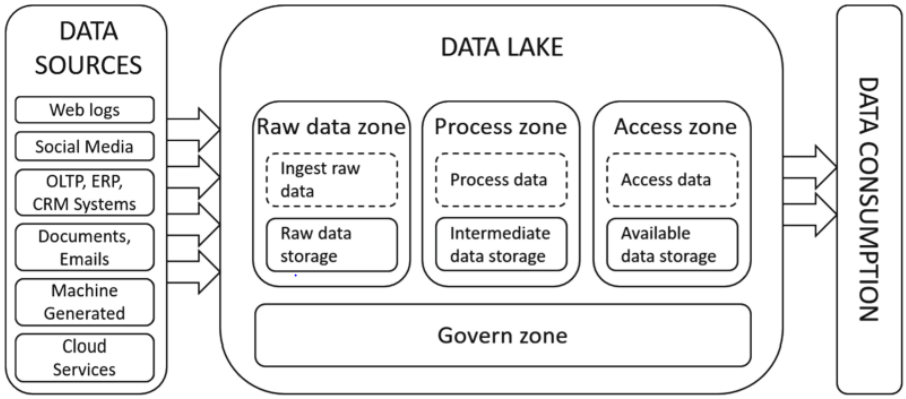


Fonte: (SAWADOGO; DARMONT, 2021)

Em ambas arquiteturas, de lagoa e zonas, os dados são pré-processados. A distinção entre elas não é tão nítida, a arquitetura de lagoa pode ser considerada uma variante da arquitetura de zonas, visto que a localização dos dados também depende do seu grau de refinamento. Uma das desvantagens no uso da arquitetura de lagoas é a possível perda de dados entre as transferências dos dados armazenados na lagoa de dados brutos para as demais. Em contrapartida, na arquitetura de zonas, o fluxo de dados na horizontal passando pelas seis zonas, pode ocasionar várias cópias do mesmo dado, e consequentemente, dificuldades no controle da linhagem dos dados.

Outras propostas de arquitetura foram apresentadas (RAVAT; ZHAO, 2019a), (GIEBLER et al., 2021), (CHIHOUB et al., 2020). Todas elas seguem a prática da divisão do Data Lake em zonas para manter um controle mais adequado de governança e organização dos dados. Ravat e Zhao (2019a) introduzem uma arquitetura funcional, distinguindo-a precisamente de uma arquitetura técnica, pois segundo eles, uma arquitetura funcional diz respeito a perspectiva de uso e pode ser implementada por diferentes soluções técnicas. Essa arquitetura, como mostra a Figura 4, é composta por quatro zonas consideradas essenciais, onde cada uma delas, exceto a zona de governança, possui uma área de tratamento e armazenamento de dados que armazena o resultado dos processos.

Figura 4 – Arquitetura funcional de Data Lake.



Fonte: (RAVAT; ZHAO, 2019a)

Assim como descrito em Sawadogo e Darmont (2021), a zona de dados brutos (*Raw data zone*) armazena todos os tipos de dados em seu formato original. A ingestão desses dados pode ocorrer no modo *batch*, *real-time* ou híbrido, combinando os dois, e o formato de armazenamento do dado nesta zona pode ser diferente do formato obtido na origem. Na zona de processamento (*Process zone*), os usuários podem transformar os dados, aplicando operações de refinamento e agregações para suas análises. A zona de acesso (*Access zone*) disponibiliza todos os dados prontos para consumo, permitindo o acesso a usuários e aplicações *self-service*. Por último, a zona de governança (*Govern zone*) se estende a todas as outras zonas, seu objetivo é garantir a segurança, qualidade dos dados, ciclo de vida, acesso e gerenciamento de metadados.

Arquiteturas funcionais possuem a vantagem de destacar claramente as funções a serem implementadas em um DL, o que ajuda a combinar mais precisamente as tecnologias necessárias. Na proposta de Ravat e Zhao (2019a), o acesso aos dados parece ser possível apenas para dados já refinados, armazenados na última zona, restringindo o acesso aos dados armazenados na zona de processamento que também podem ser úteis para análises, e consequentemente, limitando a descoberta de novas informações.

Todas as arquiteturas abordadas anteriormente visam ser genéricas, no entanto, imple-

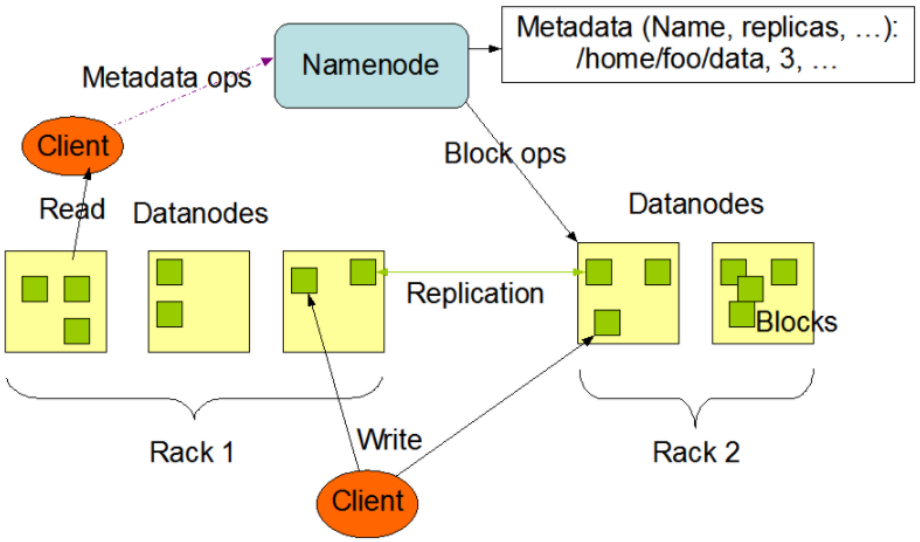
mentar a ingestão de dados, o armazenamento e a organização dos dados não é suficiente para um Data Lake funcionar efetivamente. Outros aspectos importantes como modelagem e gerenciamento de metadados também devem ser considerados (GIEBLER et al., 2021).

2.2.2 Soluções de Data Lake

A maioria das implementações de Data Lake são baseadas na suíte Apache Hadoop[1] (SAWADOGO; DARMONT, 2021). O Hadoop oferece uma estrutura de processamento escalonável e distribuído via MapReduce ou Spark que aumenta consideravelmente a performance de processamento de um grande volume de dados. Além disso, também fornece armazenamento com o Hadoop Distributed File System (HDFS), um sistema de arquivos altamente tolerante a falhas, projetado para ser implantado em hardwares de baixo custo (HUKKERI; KANORIA; SHETTY, 2020).

A base de arquitetura do HDFS funciona como mestre - escravo (FOUNDATION, 2020) e é representada na Figura 5. Isto é, existe um nó mestre e existem vários nós escravos, o mestre gerencia, mantém e monitora os nós escravos, enquanto que estes são os nós reais que executam o trabalho. Um cluster HDFS consiste em um único NameNode, há também vários DataNodes que gerenciam o armazenamento conectado aos nós em que são executados. Internamente, um arquivo é dividido em um ou mais blocos que são armazenados em um conjunto de DataNodes. Esses blocos se replicam de acordo com o fator de replicação, garantindo que mesmo após uma falha no cluster, cópias serão mantidas para recuperação(FOUNDATION, 2020). A Figura 5 ilustra o funcionamento da arquitetura HDFS.

Figura 5 – Arquitetura HDFS.



Fonte: (FOUNDATION, 2020)

No entanto, Hadoop não é a única tecnologia adequada para implementar um Data Lake.

[1] https://hadoop.apache.org/

A AWS[2] também desenvolveu uma arquitetura de DL que permite o desenvolvimento de uma solução simples e robusta por meio do Amazon Simple Storage Service (S3)[3] e outros serviços de suporte que fornecem uma gama de recursos especializados para minimizar a complexidade de implementação.

A arquitetura para Data Lake da AWS dispõe virtualmente de escalabilidade ilimitada. Seu modelo de negócio se caracteriza por pagamento mediante o uso, facilitando nas mudanças de infraestrutura de armazenamento. Além do que, sua durabilidade (confiabilidade) de dados calculada em 99,999999999% o posiciona muito a frente de outros provedores concorrentes (HUKKERI; KANORIA; SHETTY, 2020).

Na Figura 6, são apresentados outros serviços essenciais que compõe a arquitetura AWS. O AWS Glue[4] atua na catalogação dos conjuntos de dados existentes no Data Lake. O serviço AWS Lambda[5] executa aplicações utilizando o modelo *serveless*, permitindo o desenvolvimento de funções em várias linguagens de programação, como Python[6], sem a preoucupação de servidores.

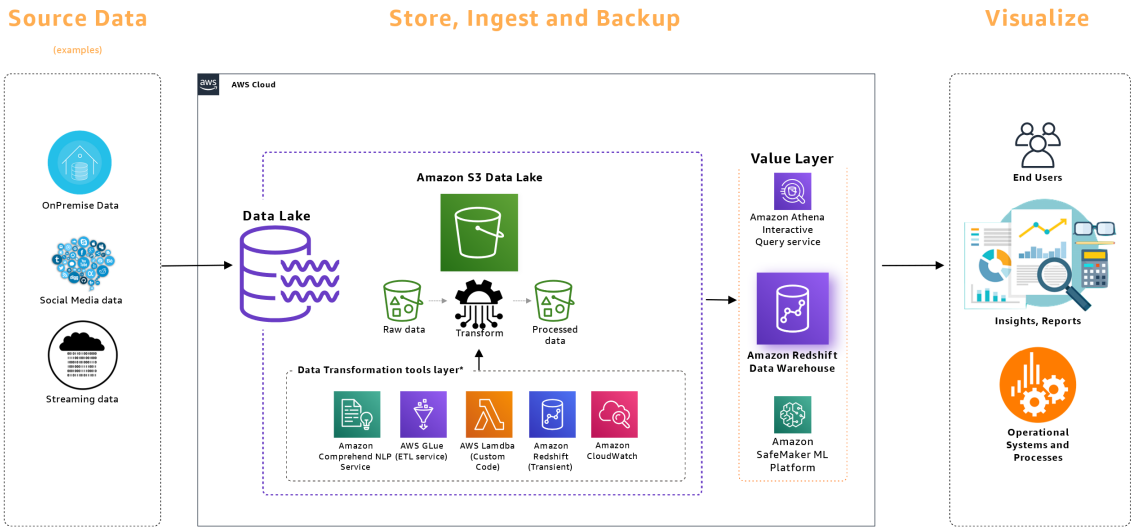
Com o AWS Athena[7], é possível executar consultas SQL padrão em dados armazenados no S3 ou no data warehouse Amazon Redshift[8]. Para monitoramento e observabilidade do DL, o AWS CloudWatch[9] dispõe dados e *insights* que ajudam na otimização e utilização dos recursos.

Paralelamente, a Microsoft[10] apresentou o Azure Data Lake Store (ADLS)[11] e o Azure Data Lake Analytics (ADLA)[12] que juntos formam a solução de DL oferecida pela empresa. Semelhante a arquitetura da AWS, a arquitetura do ADLS, representada na Figura 7, possui alta escalabilidade, modularidade, e é baseada em Microsserviços.

Também adota o mesmo modelo de negócio da AWS, isto é, o pagamento ocorre mediante o uso. O ADLA é o serviço de análise distribuído que permite desenvolver e executar pipelines de dados massivos e inclui U-SQL, uma linguagem de consulta distribuída que combina a simplicidade de SQL com alto poder de processamento (HUKKERI; KANORIA; SHETTY, 2020).

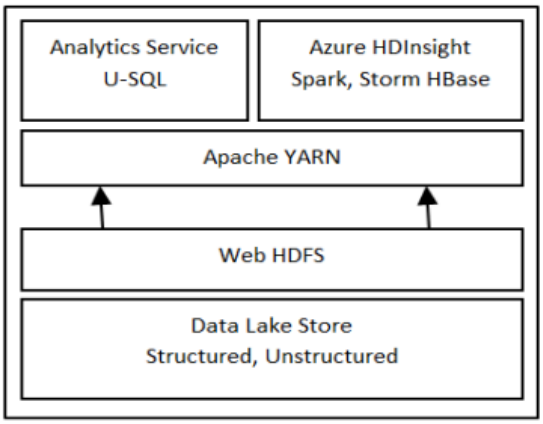
[2] https://aws.amazon.com/
[3] https://aws.amazon.com/pt/s3/
[4] https://aws.amazon.com/pt/glue/
[5] https://aws.amazon.com/pt/lambda/
[6] https://www.python.org/
[7] https://aws.amazon.com/pt/athena/
[8] https://aws.amazon.com/pt/redshift/
[9] https://aws.amazon.com/pt/cloudwatch/
[10] https://www.microsoft.com/pt-br
[11] https://azure.microsoft.com/pt-br/services/storage/data-lake-storage/
[12] https://azure.microsoft.com/pt-br/services/data-lake-analytics/

Figura 6 – Arquitetura AWS para Implementação de Data Lakehouse.



Fonte: (NAMBIAR; PRAVEEN; SUDARSHAN, 2021)

Figura 7 – Azure Data Lake.



Fonte: (HUKKERI; KANORIA; SHETTY, 2020)

A demanda por armazenamento e processamento de dados eficiente em grande escala é cada vez mais nítida. As soluções empresariais modernas discutidas anteriormente focam em atender a essa questão de maneira flexível, permitindo escalabilidade e inclusão de novos serviços de acordo com a necessidade. Um dos principais desafios relacionados ao uso do Hadoop é o nível de complexidade associado a plataforma (HUKKERI; KANORIA; SHETTY, 2020). As ferramentas do Hadoop podem exigir um nível técnico de conhecimento para manuseio, limitando parcialmente a praticidade. Por outro lado, é um projeto de código aberto em constante evolução, adotado por grandes empresas, flexível e que pode ser utilizado em hardware de custo acessível (FOUNDATION, 2020).

Ao invés de implementar um DL ponta a ponta, as empresas tem optado por adquirir serviços que abstraem a complexidade de implantação e infraestrutura local, como é o caso da AWS e Microsoft (HUKKERI; KANORIA; SHETTY, 2020). Na AWS, além da redução de

complexidade, outras vantagens também são ofertadas, como por exemplo, versionamento de *buckets* do S3 e proteção contra perda de dados. Do mesmo modo, a Microsoft projetou sua solução para ser flexível e amigável, disponibilizando vários microsserviços especializados. No entanto, soluções mais robustas tendem a ter um custo elevado, especialmente para ambientes que recebem cargas de trabalho muito críticas, tanto de processamento quanto análise de dados e requerem alta disponibilidade. Logo, para se manterem viável financeiramente, necessitam regularmente de fiscalização e controle sobre os componentes utilizados.

2.3 LEI GERAL DE PROTEÇÃO DE DADOS (LGPD)

A Lei 13.709/2018, conhecida como Lei Geral de Proteção de Dados Pessoais (LGPD), surgiu como um mecanismo para regular o tratamento de dados pessoais em território nacional permitindo a garantia dos direitos fundamentais de liberdade e privacidade de qualquer pessoa natural cujos dados aqui sejam tratados (MORTE et al., 2020). Segundo a LPGD, dados pessoais é toda e qualquer informação relacionada a pessoa natural identificada ou identificável (LEI..., 2018). Essa iniciativa adéqua a legislação brasileira aos regulamentos já aplicados em outros países, como é o caso da GDPR [13] (*General Data Protection Regulation*), lei sancionada na União Europeia e na qual a LGPD se baseia.

A LGPD define como tratamento de dados qualquer operação realizada com dados pessoais, tais como: Coleta, produção, recepção, classificação, utilização, acesso, reprodução, transmissão, distribuição, processamento, arquivamento, armazenamento, eliminação, avaliação, modificação, comunicação, transferência, difusão ou extração (LEI..., 2018). O GDPR possui definição semelhante, de forma que um tratamento em conformidade com a GDPR também estará em conformidade com a LGPD (MORTE et al., 2020).

Na LGPD são considerados alguns perfis. O **Titular dos dados** é a pessoa natural a quem se referem os dados pessoais que são objetos de tratamento. O **Controlador**, é o responsável por tomar as decisões acerca dos tratamentos, e o **Operador** é quem realiza os tratamentos em nome do Controlador. Ambos também são referenciados como agentes de tratamento. Há também a figura da **Autoridade Nacional de Proteção de Dados (ANPD)**, um órgão governamental, que dentre outras atribuições, possui o dever de fiscalizar o cumprimento da lei e zelar pela proteção dos dados pessoais (LEI..., 2018).

Perante a LGPD, os dados pessoais qualificados como sensíveis são aqueles que possuem origem étnica ou racial, que expressam convicção religiosa ou opinião política, filiação a sindicato de caráter religioso, filosófico ou político, dados referentes a saúde, vida sexual e dados biométricos ou genéticos quando vinculados a uma pessoa natural (LEI..., 2018). Os tratamentos aplicados especialmente nesta categoria de dados só poderão ocorrer mediante consentimento com finalidades específicas cedido pelo Titular, ou quando for indispensável para: a) cumprimento de obrigação legal do Controlador; b) tratamento compartilhado de da-

[13] https://gdpr-info.eu/

dos necessários à execução, pela administração pública; c) realização de estudos por órgãos de pesquisa; d) exercício regular de direitos; e) proteção de vida; f) tutela de saude; e g) garantia da prevenção à fraude e segurança do Titular (LEI..., 2018).

Tendo em vista que, pelo tipo de natureza de informação que trazem, estes dados quando tratados, podem ocasionar discriminação de seu Titular. Logo, devem todos os cuidados já previstos para o tratamento dos dados serem aplicados de forma ainda mais intensa, uma vez que, para os dados sensíveis se espera um padrão ainda mais rigoroso de proteção.

2.3.1 Direitos dos Titulares de Dados

Além de regulamentar as diretrizes sobre como as organizações devem lidar com dados pessoais, a LGPD também assegura o direito dos Titulares dos dados. No Artº 18 da Lei, são apresentadas as informações as quais o Titular tem direito a obter do Controlador, em relação aos dados por ele tratados, a qualquer momento mediante requisição. São elas:

- **Confirmação da existência de tratamento**: Confirmação de que existe um ou mais tratamento de dados sendo realizado;
- **Acesso aos dados**: Quando o Titular dos dados envia uma solicitação buscando exercer seu direito de acesso, o Controlador dos dados deve recuperar todos os dados associados ao Titular e todo processamento que tenha sido feito para enviar ao Titular;
- **Correção de dados**: A qualquer momento o Titular tem o direito de corrigir dados incompletos, inexatos ou desatualizados;
- **Anonimização**: O Titular dos dados pode solicitar a anonimização, bloqueio ou eliminação de dados desnecessários;
- **Portabilidade**: Também é direito do Titular solicitar a portabilidade dos dados a ele associados para outro fornecedor de serviço ou produto, observados os segredos comerciais e industrial. Esse terceiro deverá contactar o Controlador, que por sua vez, terá que recuperar todos os dados e os processamentos aplicados sobre eles para envio ao terceiro. Após o envio, o terceiro confirma o recebimento dos dados e finaliza a portabilidade;
- **Eliminação**: O Titular poderá solicitar eliminação de seus dados. Exceto quando o tratamento é legal, mesmo que sem o consentimento do Titular;
- **Informação de compartilhamento**: Da mesma forma, o Titular poderá solicitar informações sobre as entidades públicas e privadas com as quais o Controlador compartilhou seus dados;
- **Não consentimento**: É permitido solicitar informações sobre a possibilidade de não fornecer consentimento, ou seja, não autorizar o tratamento e as consequências negativas;

- **Revogação do consentimento**: Por fim, o Titular poderá solicitar revogação do consentimento. Assim como se opor ao tratamento realizado nos dados ou peticionar contra o Controlador junto à autoridade nacional.

Todos estes direitos que podem ser exercidos pelo Titular dos dados implica em uma lista de obrigações que devem ser cumpridas pelo Controlador para adequação a legislação. Ao capturar dados pessoais, o Controlador deverá solicitar consentimento ao Titular e prover informações sobre a finalidade do tratamento e como utilizará o dados.

Ao ser solicitado o acesso, correção ou portabilidade dos dados, o Controlador deverá recuperar todos os dados associados ao Titular junto com os tratamentos realizados para satisfazer a estas solicitações. Se o Titular solicitar que seus dados sejam eliminados, o Controlador tem a obrigação de parar de utilizar aqueles dados e removê-los. E em casos de violação, o Controlador tem até 72 horas para comunicar o ocorrido as autoridades e ao Titular, como também as ações que serão realizadas para minimizar o dano.

Todas as solicitações requerem a busca pelos dados do Titular e devolutiva de informações por parte do Controlador. Uma pesquisa realizada por (BUFALIERI et al., 2020) examinou mais de 300 Controladores de dados que estão sob regulamentação da GDPR, solicitando dados pessoais a cada um deles. Em 50,4% dos Controladores foram descobertas falhas nos procedimentos de identificação do Titular e na fase de envio de dados, expondo os usuários a novas ameaças. Todas as requisições foram feitas online e apenas 19,6% dos examinados apresentaram políticas de privacidade em suas páginas web de acordo com a GDPR. O tempo de resposta dos Controladores variou entre segundos e mais de 90 dias, tendo apenas 17,45% deles atendendo as requisições no mesmo dia da solicitação. Mais da metade dos avaliados (52,7%) responderam aos solicitantes em formato estruturado como JSON, CSV, XLS ou XML. Os demais atenderam enviando capturas de tela de seus sistemas, anexos no formato PDF e páginas HTML. Para 36% dos Controladores, não foi possível obter qualquer dado.

Com isso, é possível constatar que, sem uma estrutura de governança de dados que gerencie adequadamente os metadados, cumprir estas obrigações pode ser algo impraticável para o Controlador de dados. Tendo em vista que, para atender as solicitações de direitos do Titular, é preciso dispor de informações sobre toda linhagem dos dados.

2.4 METADADOS

As novas tecnologias como Big Data, IoT, Data Lakes e Inteligência Artificial estão criando um cenário de mais complexidade, variedade, volume de dados e uma necessidade ainda maior de governar e compreender estes dados (FLECKENSTEIN; FELLOWS, 2018).

Metadados tem se mostrado relevantes na gestão de dados porque fornecem contexto aos conjuntos de dados, permitindo que os usuários entendam os dados a curto, médio ou longo prazo. Eles são um ponto chave para garantir que as informações sobreviverão e continuarão disponíveis no futuro. Portanto, são um recurso importante para as empresas porque melhoram

ou permitem funções, como por exemplo, organização e seleção de recursos de informações, interoperabilidade, integração, identificação e proteção dos dados (ZGOLLI; COLLET; MADERA, 2020).

Metadados são frequentemente chamados de dados sobre dados ou informações sobre informações (ZGOLLI; COLLET; MADERA, 2020). No entanto, há outras definições mais enriquecidas, como em (FLECKENSTEIN; FELLOWS, 2018), na qual metadados são descritos como dados que descrevem várias facetas de um ativo de informação para melhorar sua usabilidade ao longo de seu ciclo de vida. Podendo ser compreendidos como uma categoria de informação que identifica, descreve, explica, fornece conteúdo, contexto, estrutura e classificações pertencentes aos ativos de dados de uma organização, assim como permitem a recuperação, uso e gerenciamento eficazes desses ativos.

A importância dos metadados em soluções de Data Lake tem sido enfatizada em muitos artigos (SAWADOGO; DARMONT, 2021), (ZGOLLI; COLLET; MADERA, 2020), (RAVAT; ZHAO, 2019a), (SAWADOGO; KIBATA; DARMONT, 2019), (SURIARACHCHI; PLALE, 2016). Todos eles respeitam a ideia de que os dados ingeridos em Data Lakes não possuem esquema explícito, e DLs que armazenam muitos conjuntos de dados sem modelos ou informações descritivas, podem facilmente se tornar incompreensíveis, de forma que é obrigatório implementar um sistema de gerenciamento de metadados.

Sawadogo e Darmont (2021) identificaram 6 aspectos importantes que um sistema de metadados para Data Lake idealmente deveria implementar de modo a impedir que o DL se torne inoperável. O primeiro deles é o Enriquecimento Semântico (*Semantic Enrichment*), que envolve adicionar informações como título, tags e descrições para tornar os dados compreensíveis. Metadados semânticos podem ser a base para a geração de *links* entre os dados. O segundo ponto elencado é a indexação de dados (*Data Indexing*), comumente utilizada na área de banco de dados e recuperação da informação para localizar rapidamente um objeto. Em DLs, a indexação pode servir para recuperação simples, baseada em palavras-chaves ou consultas mais complexas utilizando outro padrão. Todos os dados estruturados, semi-estruturados ou não estruturados, se beneficiam da indexação.

O terceiro aspecto, denominado como geração de *links* (*Link Generation*), consiste em identificar e integrar links entre os dados do DL. Estes *links* podem ser usados para detectar automaticamente clusters de dados fortemente vinculados. O Polimorfismo de dados (*Data Polymorphism*), quarto item elencado, atua no gerenciamento simultâneo de várias representações de dados no DL. Uma representação de dados em um documento textual, por exemplo, pode ser uma nuvem de etiquetas ou um vetor de frequência de termos. A penúltima característica, controle de versão de dados (*Data Versioning*) expressa a capacidade de um sistema de metadados em gerenciar operações de atualização ao mesmo tempo que mantém os estados de dados anteriores. Essa funcionalidade é relevante pois permite a evolução dos dados no DL. Por fim, o rastreamento de uso (*Usage Tracking*), sexto item proposto, consiste no gerenciamento de informações sobre as interações de usuários com o DL. Essas interações

são as operações, tais como criação, leitura e atualização. Esse recurso permite acompanhar a evolução dos objetos de dados, assim como pode servir para a segurança dos dados, explicando inconsistências por meio de detecções.

O gerenciamento de metadados busca fornecer aos usuários comerciais e técnicos acesso mais fácil a metadados integrados e de alta qualidade (ZGOLLI; COLLET; MADERA, 2020). No entanto, existem muitos tipos diferentes de metadados que precisam ser organizados e classificados para satisfazer as necessidades do negócio. Na próxima Seção, serão apresentadas as classificações propostas na literatura para metadados em Data Lake.

2.4.1 Classificação de Metadados para Data Lake

(SAWADOGO; DARMONT, 2021) apresentam duas topologias de classificação de metadados para Data Lakes, na qual a primeira distingue os metadados caracterizados como funcionais, enquanto a segunda classifica-os em relação aos tipos de metadados estruturais. A primeira topologia proposta por (DIAMANTINI et al., 2018), subdivide os metadados em três categorias, sendo elas:

1. Metadados de negócio (*Business*) incluem descrições que tornam os dados mais compreensíveis e definem regras de negócios, como por exemplo, restrições de integridade.

2. Metadados operacionais (*Operational*) incluem informações geradas automaticamente durante o processamento dos dados, tais como localização, tamanho do arquivo e número de linhas processadas.

3. Metadados técnicos (*Technical*) expressam como os dados são representados, incluindo seus formatos, estrutura ou esquemas.

Além disso, estas três classificações de metadados podem se cruzar, conforme o modelo apresentado na Figura 8.

Na segunda topologia, os metadados são classificados em outras três subdivisões: *Global, Intra* e *Inter-Object-Metadata*. O conceito de objeto (*Object*) refere-se ao conceito de conjunto de dados, podendo ser uma tabela no contexto de dados estruturados ou um documento textual, no contexto de dados semi ou não estruturados (SAWADOGO; DARMONT, 2021). Intra-metadados agrupam um conjunto de características que permitem os usuários entenderem os conjuntos de dados em termos de propriedades, versionamento e semântica, enquanto que a categoria Inter, diz respeito aos relacionamentos entre os conjuntos de dados por meio de *links*, que podem ser de agrupamento, similaridade ou de paternidade. Já os metadados da categoria Global fornecem contexto que facilitam o processamento e análise dos conjuntos de dados, abrangendo informações semânticas, de indexação, e *logs* de execução.

Figura 8 – Classificação de Metadados

Fonte: (DIAMANTINI et al., 2018)

(RAVAT; ZHAO, 2019b) adaptaram as categorias de Inter e Intra-metadados incluindo novos aspectos que devem ser capturados. No que diz respeito a classe Inter-metadados, foram integrados os conceitos de:

- *Dataset containment*: Significa que um conjunto de dados pode estar contido em outro conjunto de dados.
- *Partial Overlap*: Quer dizer que alguns atributos com dados correspondentes em conjuntos de dados diferentes se sobrepõem.
- *Provenance*: Indicando que um conjunto de dados é origem de outro conjunto de dados.
- *Logical Clusters*: Para representar que diferentes conjuntos de dados estão no mesmo domínio.
- *Content Similarity*: Que significa que diferentes conjuntos de dados compartilham os mesmos atributos.

Para a categoria de Intra-metadados, foram retidas as definições de:

- *Data characteristics*: São atributos que descrevem um conjunto de dados, como identificação, nome, tamanho e data de criação.
- *Definitional*: São metadados que auxiliam os usuários a compreenderem os conjuntos de dados, descrevendo seu significado por meio de vocabulários ou conjunto de palavras-chaves.

- *Navigational*: Fornecem informações sobre a localização dos conjuntos de dados, como caminhos de arquivos ou conexões.
- *Lineage*: São informações que dizem respeito ao ciclo de vida dos dados.
- *Quality*: Metadados sobre consistência e completude dos conjuntos de dados.
- *Security*: São metadados que contribuem para a segurança dos conjuntos de dados, provendo informações sobre níveis de acesso e sensibilidade dos dados.

Apesar da estruturação em categorias, as topologias anteriormente descritas podem soar confusas, uma vez que há interseções entre as classes e similaridade de metadados, como por exemplo, os metadados de localização na categoria operacional de (DIAMANTINI et al., 2018) são equivalentes aos metadados de navegação presentes na divisão de Intra-metadados propostos por (RAVAT; ZHAO, 2019b). Além disso, não há aprofundamento na descrição dos metadados que compõe as categorias, apenas exemplos pontuais para entendimento do contexto.

2.5 RASTREABILIDADE DE DADOS

A evolução das tecnologias de Big Data e seu impacto na vida cotidiana das pessoas promovem uma busca cada vez mais crescente por informações confiáveis. A medida que estes ecossistemas expandem sua capacidade em capturar dados de diferentes fontes, se faz necessário uma rastreabilidade confiável acerca de suas origens e uso.

Nas indústrias manufatureiras, a rastreabilidade tem sido prática padrão desde o final do século XX para rastrear o produto através de seu ciclo de vida, desde sua origem como matéria-prima, até os itens finais já disponíveis em loja (BARCLAY et al., 2019). Similarmente, no ramo agroalimentar, o objetivo da rastreabilidade é permitir o monitoramento completo de um produto na cadeia de suprimentos e traçar todo o histórico dele desde o produtor até o consumidor, atuando como um instrumento preventivo de gestão da qualidade e segurança alimentar (MIRABELLI et al., 2012).

Ainda no setor alimentício, a rastreabilidade é definida como a capacidade de acompanhar o movimento de um alimento através de fases específicas da produção, rastreando o histórico desde a origem de peças e materiais, até o processamento e distribuição do produto (VIOLINO et al., 2019). Em (KELEPOURIS; PRAMATARI; DOUKIDIS, 2007), é definida como a capacidade de rastrear a história, aplicação ou localização de uma entidade por meio de identificações registradas. Do ponto de vista de software, (NUNES; CAPPELLI; RALHA, 2017) definem como a qualidade de seguir, descobrir ou verificar o curso de desenvolvimento de algo. Já (AZEVEDO; JINO, 2019) resumem as definições anteriores, afirmando que a rastreabilidade nada mais é do que a habilidade de rastrear elementos de um sistema ao longo de seu ciclo de vida.

Tecnologias de Identificação por Radiofrequência (*Radio-Frequency Identification*) - RFID, Comunicação de Campo Próximo (*Near Field Communication*) - NFC e Resposta Rápida (*Quick Response*) - QR *code*, são mencionadas na literatura (COSTA et al., 2013), (VIOLINO

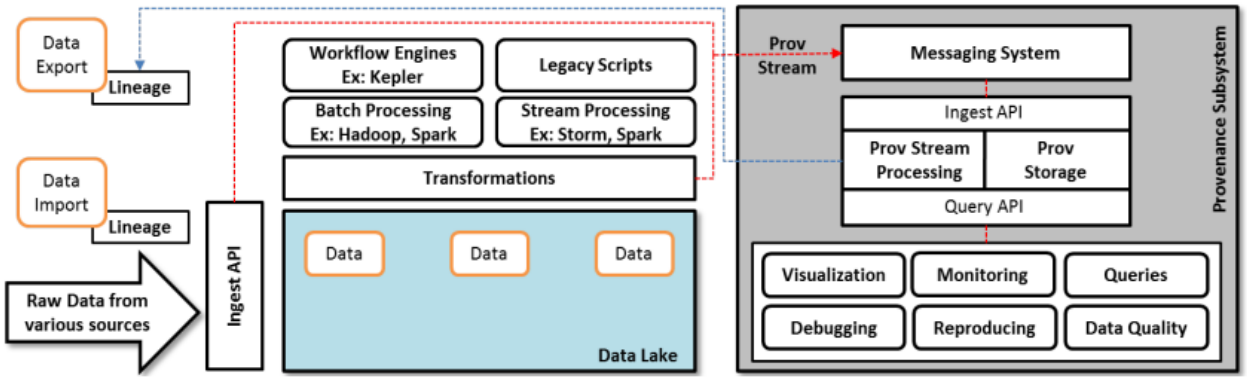
et al., 2019), (MIRABELLI et al., 2012) como soluções que permitem a rastreabilidade parcial de produtos, pois geralmente são inseridas nas embalagens, possibilitando o rastreamento apenas desse ponto da produção em diante. Outro recurso também utilizado neste contexto é o *Blockchain*, tecnologia baseada em sistemas distribuídos que utiliza blocos criptografados para armazenar os dados ao longo de uma cadeia. Uma das principais vantagens no uso do *Blockchain* é que após a aprovação da transação, as informações gravadas no bloco não podem ser alteradas. Além disso, estas transações possuem registros de data e hora e são verificadas pela comunidade da rede, transformando o bloco em um registro imutável de atividades passadas que podem ser rastreadas (HUANG; LI; THÜRER, 2019).

No cenário de Big Data, obter a rastreabilidade dos dados é um desafio. Para garantir qualidade, confiabilidade e transparência em sistemas, muitos trabalhos propõem soluções que capturam proveniência, focando em pontos específicos que dizem respeito a origem dos dados (DEZANI-CIANCAGLINI; HORNE; SASSONE, 2012), (KAKU; LOMOTEY; DETERS, 2016), (TRISOVIC et al., 2020). No entanto, para atender a questões de conformidade com as leis de proteção, informações apenas sobre a origem dos dados não são suficientes, é necessário também, capturar elementos que esclareçam como ocorre o processamento e uso dos dados. Mais uma vez, metadados podem ser úteis para compor uma solução capaz de atenuar este problema.

Referindo-se a rastreabilidade de dados em Data Lake, a literatura é restrita, tendo poucos trabalhos direcionados para DL e menos ainda para a captura de metadados que promovam a rastreabilidade em DL. O mais próximo relacionado é apresentado por (SURIARACHCHI; PLALE, 2016). Nele, os autores abordam a questão de que os itens de dados em um Data Lake podem existir em diferentes estágios durante seu ciclo de vida, reforçando que estas complicações aumentam a necessidade de mecanismos de rastreabilidade adequados. A primeira contribuição do trabalho consiste em uma arquitetura de referência baseada em um subsistema de proveniência central, como mostra a Figura 9. Este subsistema é responsável por armazenar e processar eventos de todos os sistemas do DL a ele conectados.

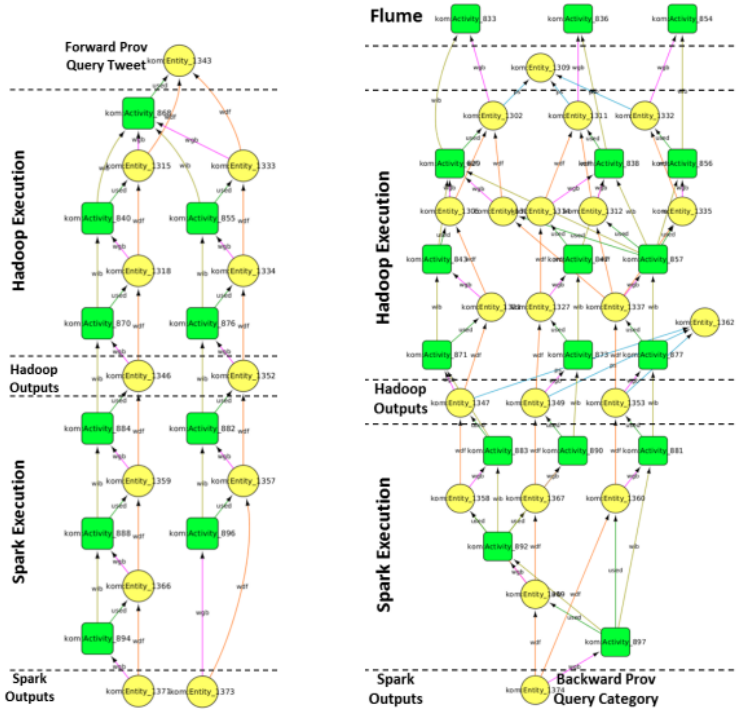
Tendo em vista que Data Lakes utilizam inúmeros *frameworks* de processamento e análise, e alguns destes mecanismos não produzem informações de proveniência por padrão, a coleta de dados de proveniência torna-se ainda mais desafiadora. Para solucionar este ponto, os autores propuseram como segunda contribuição uma técnica para captura de proveniência capaz de coletar metadados de *workflows* distintos e integrá-los. Dessa forma, é possível visualizar o fluxo completo dos dados durante seu ciclo de vida. Na Figura 10, é apresentado um exemplo de *workflow* que utiliza mais de uma ferramenta e como a proveniência coletada é integrada durante o processamento no Data Lake.

Figura 9 – Proveniência para Data Lakes: Arquitetura de referência.



Fonte: (SURIARACHCHI; PLALE, 2016)

Figura 10 – Proveniência Integrada.



Fonte: (SURIARACHCHI; PLALE, 2016)

No fluxo apresentado como resultado, é possível rastrear a origem dos itens de dados e a formação de novos a medida que eles passam por ferramentas distintas. No entanto, não é possível identificar com clareza se há outros tratamentos sendo aplicados nos dados, além da criação. Do mesmo modo, não há informações sobre quem executa as operações ou como os dados estão sendo efetivamente utilizados.

O subsistema no qual a arquitetura se baseia não leva em consideração nenhuma das arquiteturas de Data Lakes discutidas na seção 2.2.1 ou alguma das leis de proteção de dados

mencionadas na seção 2.3. Também não há descrições acerca dos metadados de proveniência coletados no processo.

Apesar das relevantes contribuições, a pesquisa discutida anteriormente apresenta lacunas a serem preenchidas. Sendo a principal delas, a descrição adequada dos metadados coletados e como estes podem ser utilizados para fins de conformidade com as leis de proteção de dados.

2.6 CONSIDERAÇÕES FINAIS

Neste Capítulo, foi apresentada uma visão geral sobre como Data Lakes emergiram e a necessidade imposta pelas leis de proteção de dados, de mecanismos de rastreabilidade para acompanhamento do ciclo de vida dos dados armazenados em DLs. Em seguida, descrevemos as características de Data Lakes, que envolvem arquitetura e soluções atualmente disponíveis. Na seção 2.3, discutimos a LGPD e os direitos que podem ser exercidos pelos Titulares dos dados. Na sequência, apresentamos definições para metadados e as topologias de classificações existentes na literatura. Por fim, finalizamos o Capítulo conceituando rastreabilidade e exemplificando sua aplicabilidade no contexto de Data Lake.

3 TRABALHOS RELACIONADOS

Neste Capítulo, iremos apresentar os trabalhos que mais se assemelham a proposta dessa dissertação. Inicialmente, na Seção 3.1, apresentamos uma visão geral dos modelos. Na Seção 3.2, abordaremos alguns modelos que representam o regulamento GDPR e na Seção 3.3 são mostrados alguns exemplos de modelos desenvolvidos no contexto de Data Lakes. Por fim, na Seção 3.4 são apresentadas as Considerações Finais do Capítulo.

3.1 VISÃO GERAL

Metadados desempenham um papel fundamental ao favorecer a interoperabilidade entre fontes de dados heterogêneas. Com o advento dos Data Lakes este papel se tornou ainda mais relevante, pois neste ambiente, metadados representam a única possibilidade de garantir associação entre as fontes de dados e evitar que o DL se torne inacessível e inoperável. (DIAMANTINI et al., 2018).

O gerenciamento de metadados está fortemente relacionado ao gerenciamento de modelos (QUIX; HAI; VATOV, 2016). Modelos de metadados podem ser utilizados como um mecanismo para estruturar e representar o conhecimento sobre um determinado domínio. Quando empregado em um sentido mais específico, como fazemos neste trabalho, o modelo pode viabilizar uma ponte de comunicação entre especialistas em TI e especialistas do domínio, com especialistas jurídicos que possuem pouca experiência em tecnologias de Big Data.

No contexto das leis de proteção de dados, alguns modelos foram propostos para representação da GDPR (*e.g.* (UJCICH; BATES; SANDERS, 2018), (TORRE et al., 2019), (TORRE et al., 2020)) com o intuito de servir como base para desenvolvimento de futuros métodos automatizados para avaliação de *compliance* com o regulamento. Para Data Lakes, também foram apresentadas outras propostas que tem como objetivo coletar metadados para o gerenciamento do conhecimento sobre os dados. Diante disso, nas próximas seções serão abordados alguns modelos para GDPR e alguns modelos de metadados para Data Lakes.

3.2 MODELOS PARA GDPR

GDPR é considerado o regulamento de proteção e privacidade de dados mais abrangente e tecnicamente já estabelecido. Embora seja benéfico para os indivíduos, na prática ele apresenta desafios significativos para as organizações que monitoram ou armazenam dados pessoais (TORRE et al., 2019). Seu surgimento trouxe a tona a necessidade de soluções automatizadas para verificação de conformidade. No entanto, ocorre que as organizações não tem utilizado outra abordagem para solucionar este ponto a não ser a realização de auditorias manuais e custosas.

Neste sentido, alguns esforços foram direcionados para representar o regulamento por meio de modelos conceituais visando a utilização deles para o desenvolvimento de soluções que estejam em conformidade com os requisitos regulatórios. A seguir, serão discutidos alguns trabalhos sobre modelagem no contexto da GDPR que influenciaram na construção da proposta dessa dissertação.

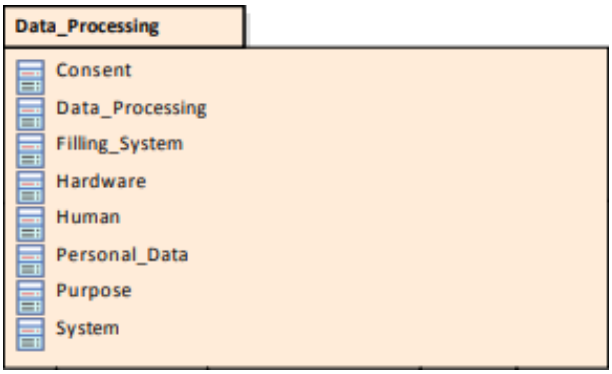
3.2.1 Torre et al. 2019

O trabalho apresentado por (TORRE et al., 2019) propõe um modelo conceitual utilizando a notação UML[1] para representar os principais conceitos da GDPR. Além disso, também são apresentados exemplos de verificação de restrições com OCL[2]. O principal objetivo deste trabalho consiste em prover uma solução de modelagem genérica que pode ser instanciada em diferentes contextos para satisfazer aos requisitos de conformidade do regulamento.

Para chegar a esta representação genérica, o processo de desenvolvimento do modelo passou por iterações com profissionais especializados no domínio da GDPR. Com base nas informações extraídas, alguns artefatos foram modelados totalizando nove pacotes, nos quais cada objeto representa um ou mais conceitos do regulamento. Cada pacote possui um modelo próprio. Discutiremos apenas o pacote de processamento dos dados (*Data Processing*).

No pacote apresentado na Figura 11, são listadas as informações que devem ser coletadas acerca do processamento dos dados de modo a alcançar a conformidade. Para algumas dessas informações, não é possível associar o termo a nenhum artigo da GDPR, como é o caso de *Hardware* e *Human*.

Figura 11 – Pacote de Processamento de Dados.



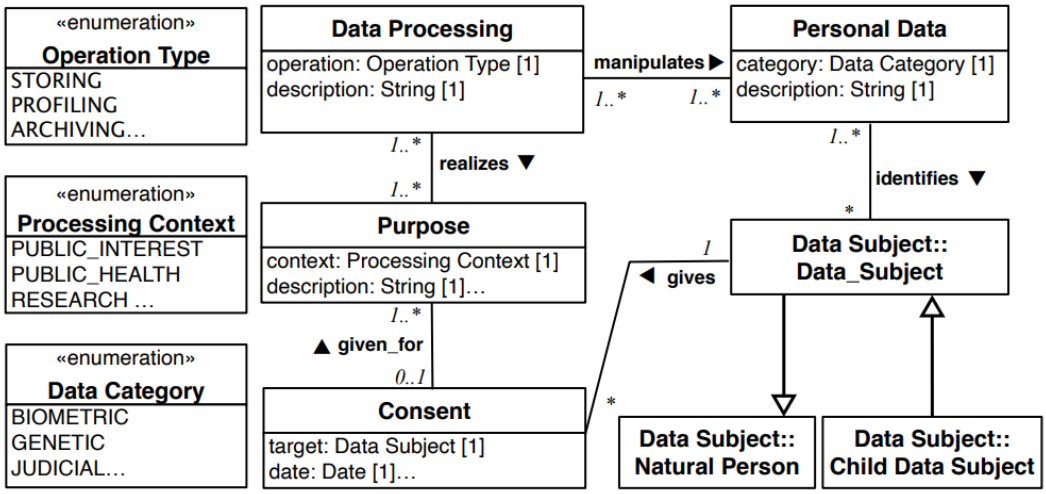
Fonte: (TORRE et al., 2019)

Na Figura 12, é possível visualizar como os termos apresentados no pacote de processamento dos dados se relacionam. Todo processamento de dados (*Data Processing*) manipula dados pessoais (*Personal Data*) e estes dados pertencem a categorias (*Data Category*) que distinguem dados pessoais de dados pessoais sensíveis. Os dados pessoais identificam o titular

[1] https://www.omg.org/spec/UML/2.5.1/About-UML/
[2] https://projects.eclipse.org/projects/modeling.mdt.ocl

dos dados (*Data Subject*), que por sua vez, consente (*Consent*) o tratamento nos seus dados para uma finalidade (*Purpose*). O processamento dos dados é realizado com base nesta finalidade e pode ser classificado em vários tipos de operação (*Operation Type*). Além disso, a finalidade a qual o processamento de dados deve atender possui um contexto (*Processing Context*) que descreve o cenário para qual o dado será processado.

Figura 12 – Modelo do Pacote de Processamento de Dados.



Fonte: (TORRE et al., 2019)

De acordo com os autores, o modelo serve como um ponto de partida para o desenvolvimento de abordagens que verifiquem de forma automatizada a conformidade dos sistemas com a GDPR. Para validação da proposta, foram desenvolvidos códigos em OCL que realizam verificações nas regras de conformidade especificadas com base nos artigos do regulamento. Além disso, são apresentados dicionários que facilitam a compreensão do mapeamento entre as regras técnicas e os artigos.

No entanto, não é esclarecido como aplicar o modelo em um cenário mais real que envolve sistemas que processam dados pessoais, assim como alguns termos dos pacotes modelados não são bem contextualizados. A rastreabilidade mencionada refere-se apenas ao mapeamento dos elementos do modelo com os artigos da GDPR, e apesar dos exemplos práticos com OCL mostrarem a aplicabilidade da proposta, o uso do mecanismo não é considerado em um cenário de Big Data.

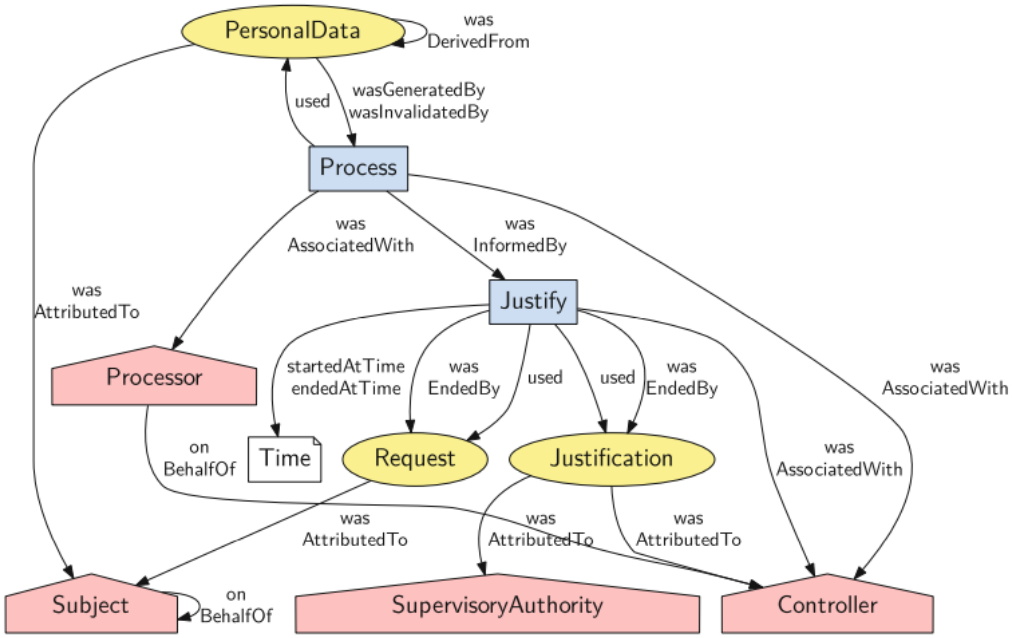
3.2.2 Ujcich et al. 2018

O trabalho de (UJCICH; BATES; SANDERS, 2018) foca em identificar os principais desafios no que diz respeito a conformidade com a GDPR e em como a proveniência de dados pode ser aplicada para resolução destes desafios. Para isso, os autores propuseram um modelo de proveniência de dados que representa quais informações devem ser coletadas pelos controladores e processadores de dados, para que estes estejam aptos a demonstrar as autoridades relevantes que armazenaram, processaram e compartilharam dados de maneira adequada.

A motivação para o modelo é baseada numa lista de conceitos que mapeiam alguns direitos e obrigações do regulamento para os quais a proveniência de dados é aplicável. Em relação ao direito a remoção dos dados, por exemplo, a proveniência pode auxiliar na identificação de quais dados serão afetados após a operação. Da mesma forma, a portabilidade dos dados pode ser rastreada entre controladores com o uso de um modelo de proveniência.

Na Figura 13, pode-se visualizar que o modelo é composto por classes que representam agentes, atividades, entidades e os relacionamentos entre elas. Os símbolos de coloração rosa, que remetem a imagem de uma casa, representam os agentes (a exemplo, o Titular dos dados, o Processador dos dados, o Controlador e a autoridade supervisionadora). Os retângulos representam as atividades e englobam o processamento dos dados e a justificativa. As elipses representam as entidades, caracterizadas pela solicitação do titular, seus dados pessoais, e a justificativa legal em qual o processamento se baseia. As setas refletem os relacionamentos e as notas descrevem as propriedades deles.

Figura 13 – Modelo de Proveniência de Dados para GDPR.



Fonte: (UJCICH; BATES; SANDERS, 2018)

Embora o modelo de proveniência descreva qual dado coletar, ele ainda não explica como utilizar tal proveniência. Visando atender este ponto, os autores introduziram um padrão de design que modeladores e profissionais podem utilizar para descrever eventos comuns. Baseado no modelo de alto nível descrito anteriormente, um caso de uso envolvendo um cenário de compras é desenvolvido para explicar como a proveniência de dados pode ser utilizada na verificação de conformidade com a GDPR.

Apesar de promissora, a coleta de proveniência dos dados não é suficiente para atender as questões de conformidade impostas pelo regulamento, pois é necessário capturar também,

além de dados sobre a origem, informações sobre o armazenamento e uso dos dados.

3.3 MODELOS PARA DATA LAKES

Data Lakes emergiram inicialmente como repositórios de dados e ao longo do tempo se tornaram uma das soluções para análise de Big Data mais populares. A principal questão dos Data Lakes é que eles podem facilmente se transformar em um pântano de dados (*data swamp*), ficando inacessível e invisível para os usuários.

Para atender esta lacuna, o gerenciamento de metadados é enfatizado em alguns trabalhos na literatura (*e.g* (ZHAO; MEGDICHE; RAVAT, 2021), (RAVAT; ZHAO, 2019b), (SCHOLLY et al., 2021), (EICHLER et al., 2021)). Todos eles seguem a abordagem de desenvolvimento de modelos para sistemas que devem coletar e gerenciar metadados automaticamente. A seguir, serão apresentados os modelos de metadados para Data Lake que mais se assemelham a proposta dessa dissertação.

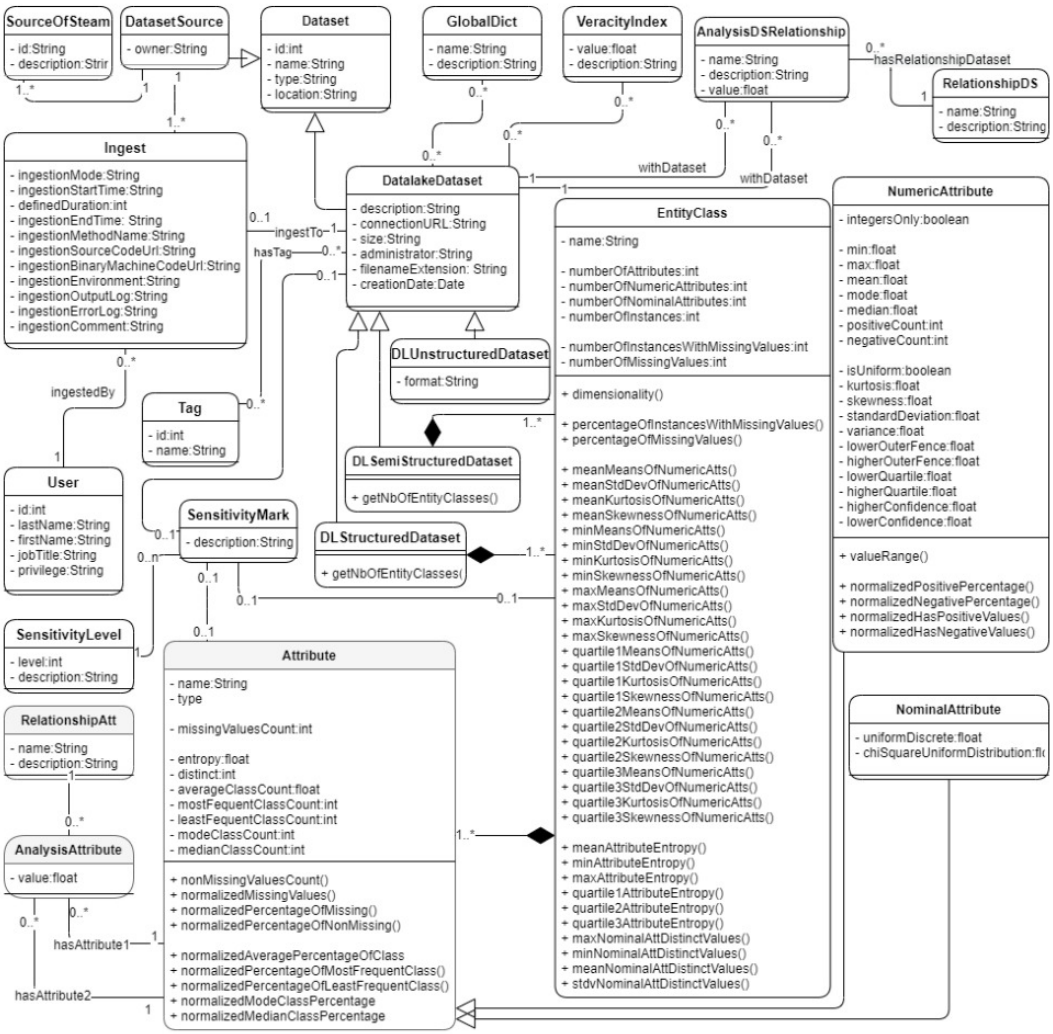
3.3.1 Zhao et al. 2021

Em (ZHAO; MEGDICHE; RAVAT, 2021), os autores propuseram um sistema para gerenciamento de metadados com foco no processo de ingestão de dados. Considerando que Data Lakes são caracterizados pela abordagem de *schema on read*, dados são ingeridos sem a necessidade de transformações e são processados apenas quando é necessário.

Diferentemente dos processos de ETL (*Extract, Transform, Load*) utilizados para tratamento e carga de dados estruturados em Data Warehouses, a ingestão de dados em Data Lakes lida com diversos tipos de dados, além de enfrentar desafios específicos, tais como: Integração com diferentes origens (banco de dados, servidores, IoT, etc.), integração de transformações em *small files* que envolvem modificação de formato, compressão e junção de múltiplos arquivos pequenos, e ingestão em múltiplos modos (*batch, real-time*).

Para lidar com estes desafios, uma das principais contribuições do trabalho consiste em um modelo de metadados que engloba diferentes categorias de metadados gerados durante a fase de ingestão. Este modelo estende uma versão anterior apresentada em (RAVAT; ZHAO, 2019b). Dentre as vantagens que o modelo oferece, destacam-se o fornecimento de: *(i)* informações sobre a proveniência dos dados, *(ii)* informações sobre o processo de ingestão, como código fonte, detalhes de execuções, dentre outros, e por último, *(iii)* informações descritivas dos dados, que envolvem as características básicas que ajudam os usuários a entenderem a estrutura dos dados e seu conteúdo. Na figura 14, é apresentado o modelo proposto.

Figura 14 – Modelo para Ingestão de Dados.



Fonte: (ZHAO; MEGDICHE; RAVAT, 2021)

O modelo é completo na representação de metadados que descrevem a origem de um conjunto de dados, quando e como e ele foi inserido no Data Lake, sua estrutura e conteúdo. Somado a isso, também é possível compreender seus relacionamentos após a ingestão no ambiente. No entanto, para atender a requisitos de conformidade, é preciso obter e gerenciar metadados que vão além da fase inicial de ingestão de dados. Isto é, que descrevam também os processamentos realizados no conjunto de dados, suas derivações e uso.

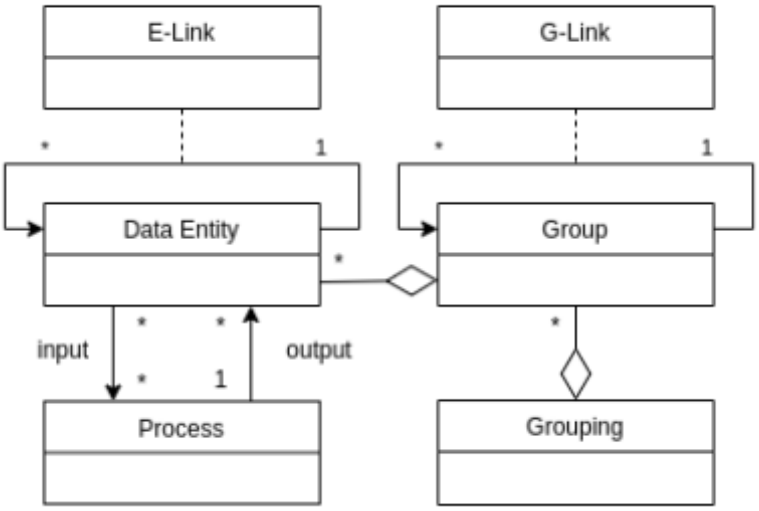
3.3.2 Scholly et al. 2021

No contexto de gerenciamento de metadados, metamodelos funcionam como uma abordagem genérica de caráter mais teórico, que tem como objetivo fornecer diretrizes detalhadas para projetos de sistemas de metadados de modo flexível. No trabalho de (SCHOLLY et al., 2021), é apresentado um metamodelo para gerenciamento de metadados, o qual tem como objetivo ser o mais genérico e adaptável para todos os tipos de Data Lakes.

O metamodelo nomeado de goldMEDAL, apresentado na Figura 15, engloba quatro conceitos essenciais: Entidade de dados (*Data Entity*), processos (*Process*), agrupamento (*Grouping*) e *Links*.

Entidade de dados é o conceito generalizado que representa um conjunto de dados em seus diferentes formatos e granularidade. Operações de transformação e atualização, que servem para rastrear a linhagem do conjunto de dados e seus versionamentos, são representadas pelo conceito de processos. Já o conceito de agrupamento é definido por um conjunto de grupos, no qual cada grupo reúne entidade de dados que possuem propriedades em comum. Por fim, os *links* representam a associação entre as entidades de dados e seus agrupamentos.

Figura 15 – Metamodelo goldMEDAL.



Fonte: (SCHOLLY et al., 2021)

Outra contribuição relevante do trabalho é a formalização do metamodelo para eliminação de ambiguidades. Além disso, os autores também apresentaram uma lista de *features* na qual o metamodelo se baseia para comportar todos os casos de uso possíveis no Data Lake. Essas funcionalidades são utilizadas como apoio para comparação do modelo proposto com outros existentes na literatura e avaliação da "genericidade". Do mesmo modo, também utilizamos algumas destas funcionalidades como base para a construção da proposta desta dissertação. A seguir, são descritas as funcionalidades base do metamodelo goldMEDAL.

- **Enriquecimento semântico**: Diz respeito a captura de informações que atribuam significado ao conjunto de dados no Data Lake.
- **Múltiplas zonas**: Remete a arquitetura do Data Lake, o qual pode ser dividido em zonas que armazenam dados em diferentes granularidades.
- **Versionamento**: Diz respeito a captura de metadados que descrevem a atualização do conjunto de dados.

- **Rastreabilidade**: Permite o rastreamento da linhagem dos dados no Data Lake.
- **Categorização**: Diz respeito a classificação dos metadados no Data Lake.
- **Links de similaridade**: Esta funcionalidade diz respeito a identificação de relacionamentos entre dados no Data Lake que estão fortemente vinculados.
- **Propriedades de metadados**: Diz respeito as características que definem os metadados.
- **Múltiplos níveis de granularidade**: Refere-se aos diferentes níveis de informação e formato que o conjunto de dados apresenta, como por exemplo, linha ou tabela.

No entanto, todas essas funcionalidades não estão evidentes no metamodelo, é preciso ler uma parte considerável do trabalho para conseguir relacionar os elementos visuais aos requisitos funcionais. De modo geral, o metamodelo define poucos elementos de metadados, os quais por sua vez, não apresentam nenhuma propriedade. Além disso, também não é possível distinguir com clareza quais são os elementos que representam as características de Data Lakes.

3.3.3 Eichler et al. 2021

O gerenciamento de metadados constitui todas as atividades que envolvem o gerenciamento do conhecimento de uma organização sobre seus dados. Sem este conhecimento, os dados podem não ser aplicáveis para a finalidade pretendida, devido a falta de qualidade ou confiança. Um aspecto importante no gerenciamento de metadados é a definição de um modelo que descreva as relações entre os dados e os elementos de metadados. Visando atender este ponto, em (EICHLER et al., 2021), os autores propuseram o HANDLE, um metamodelo genérico desenvolvido para refletir os metadados específicos de um elemento de dados ou de todo o Data Lake.

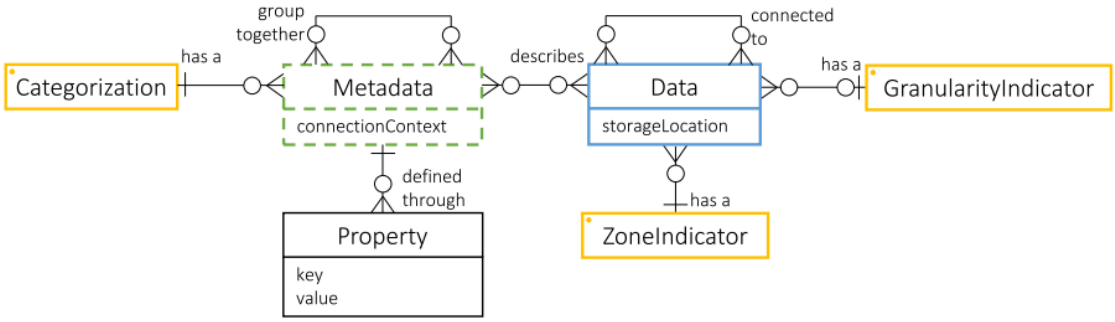
Semelhante ao GoldMEDAL (SCHOLLY et al., 2021) apresentado na seção anterior, o HANDLE (EICHLER et al., 2021) tem como objetivo ser genérico o suficiente para dar suporte a diferentes casos de uso de gerenciamento de metadados em um Data Lake. Os autores ainda pontuam que as abordagens de lista de *features* ou classificação de metadados não são apropriadas para o desenvolvimento de um modelo suficientemente flexível, devido a isso, a metodologia utilizada para conduzir o desenvolvimento do modelo proposto foi através de um caso de uso no contexto da Indústria 4.0.

Para abranger todos os cenários de uso de metadados no DL, os autores elencaram quatro requisitos que o metamodelo deve atender para ser classificado como genérico. O primeiro deles consiste em modelar os metadados da maneira mais flexível possível, e isso compreende armazenar metadados no formato de objetos de metadados, propriedades e relacionamentos. O segundo requisito está relacionado a habilidade de coletar metadados em diferentes granularidades, como por exemplo, tabelas, colunas, linhas ou arquivos.

Considerando as características do Data Lake, o terceiro requisito especifica que o metamodelo deve dar suporte ao conceito de zonas do DL, garantindo que os metadados sejam distinguíveis entre elas. Por fim, o último requisito elencado integra a categorização do metadado no formato de *label*.

Com base nesses requisitos, o metamodelo conceitual desenvolvido é composto por duas partes, o metamodelo principal e outras três extensões que podem ser adaptadas na implementação no Data Lake. O metamodelo central define todos os elementos e relacionamentos necessários para modelar um caso de uso de gerenciamento de metadados. Já as extensões abordam os tópicos de granularidade, zona e categorização com mais detalhes, respectivamente.

Figura 16 – HANDLE - Modelo Principal.



Fonte: (EICHLER et al., 2021)

No metamodelo central apresentado na Figura 16, um dos elementos principais é a entidade de Dados (*Data*) ilustrada pelo retângulo sólido na cor azul, que aponta para a localização (*storageLocation*) de um dado ou um conjunto de dados no Data Lake. Um elemento de dados possui duas entidades associadas a ele, que são o indicador de zona (*zoneIndicator*) e o indicador de granularidade (*granularityIndicator*).

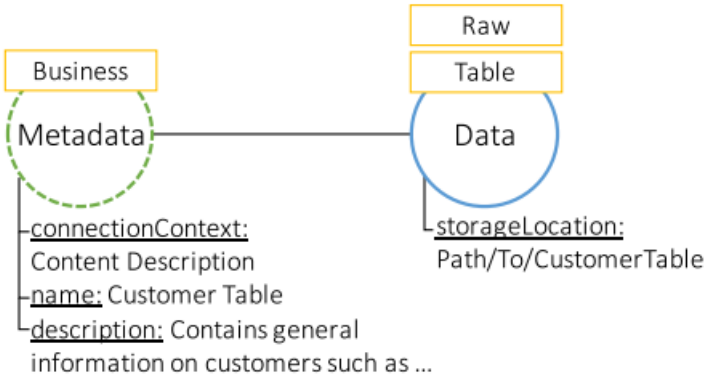
O outro elemento central do metamodelo é a entidade de metadados (*Metadata*), representada pelo retângulo pontilhado em verde. Como atributo, esta entidade possui o contexto de conexão (*connectionContext*) que descreve qual informação a entidade de metadados contém. Além disso, a entidade também é categorizada (*Categorization*) de acordo com seu conteúdo e definida por meio de propriedades (*Property*).

Além da flexibilidade, uma das principais vantagens do metamodelo HANDLE é que sua notação facilita visualmente na identificação dos principais elementos. No exemplo da Figura 17, podemos visualizar que o elemento de dados tem um atributo (*storageLocation*) que contém o caminho para os dados no Data Lake. Além disso, o indicador de granularidade (*granularityIndicator*) implica que o caminho aponta para uma tabela, e o indicador de zona (*zoneIndicator*) indica que o dado está armazenado na zona bruta.

O elemento de dados possui um elemento de metadados que descreve o conteúdo dos dados, de acordo o contexto de conexão (*connectionContext*). Somado a isso, o elemento de metadados possui as propriedades de nome (*name*) e descrição (*description*). Por fim, o

elemento é classificado como sendo de negócio, conforme especificado na entidade de categorização (*Business*).

Figura 17 – HANDLE - Modelo Principal Instanciado.



Fonte: (EICHLER et al., 2021)

Podemos concluir que os metamodelos goldMEDAL e HANDLE são genéricos e flexíveis. Ambos requerem conhecimento aprofundado sobre o domínio de implementação para adaptação e uso. No entanto, o goldMEDAL soa menos intuitivo devido ao seu alto nível de abstração, sem contar que a explicação do metamodelo é feita por meio de uma formalização, que requer um certo domínio matemático ou lógico para seu entendimento. Já o HANDLE, com seu esquema de cores e notação, facilita a interpretabilidade dos elementos, seus relacionamentos e propriedades. Além disso, as características do Data Lake são mais evidentes e a abordagem de explicação por meio de instanciação e caso de uso, torna mais didático o processo de aprendizado.

3.4 CONSIDERAÇÕES FINAIS

Neste Capítulo foram apresentados alguns modelos voltados para o gerenciamento de metadados. Esses modelos foram distribuídos entre modelos que representam a lei de proteção de dados da União Europeia GDPR, e modelos que direcionam a coleta de metadados em Data Lakes. Os modelos que abordam os conceitos principais do regulamento foram propostos com o intuito de servir como ponto de partida para o desenvolvimento de soluções automatizadas que verificam a conformidade com a lei, mas não abordam características de Data Lakes em si.

Já os modelos que são direcionados para a coleta de metadados em Data Lakes visam dar suporte ao conhecimento sobre os dados armazenados no DL para impedir que ele se torne um ambiente inoperável. É válido ressaltar que os modelos possuem caráter genérico e não focam na resolução de questões de conformidade com nenhuma das leis de proteção de dados, GDPR ou LGPD.

Sendo assim, podemos concluir que nenhum dos modelos apresentados cobre os pontos necessários para que um agente Controlador atenda a solicitação de acesso aos dados efetuada

pelo Titular dos dados, pois é necessário abordar aspectos legais e operacionais para responder a esta requisição.

4 MODELO DE SUPORTE PARA CONFORMIDADE DE DATA LAKE COM A LGPD

Como mostrado no Capítulo anterior, os modelos de metadados propostos na literatura geralmente não cobrem simultaneamente características de Data Lake e noções jurídicas a respeito das leis de proteção de dados. Também não foram encontrados modelos que representem a lei brasileira de proteção de dados pessoais. Somado a isso, os poucos modelos existentes possuem um alto nível de abstração e focam em ser genéricos. Dessa forma, neste trabalho propomos o Modelo de Suporte para Conformidade de Data Lake com a LGPD - DLCM, que descreve quais metadados devem ser coletados no Data Lake para que o Controlador dos dados consiga atender a uma requisição de acesso aos dados efetuada pelo Titular dos dados.

4.1 VISÃO GERAL DO DLCM

Como já mencionado, o *Modelo de Suporte para Conformidade de Data Lake com a LGPD* - DLCM - foi proposto com o objetivo de descrever quais metadados devem ser capturados acerca dos tratamentos realizados nos dados em um Data Lake. Esses metadados devem auxiliar o agente controlador no retorno à solicitação de acesso aos dados efetuada pelo Titular dos dados. Para isso, o modelo incorpora metadados que compreendem conceitos discutidos no Capítulo 2.

Para construção do DLCM, utilizamos alguns trabalhos como base para extração dos elementos de metadados, além daqueles apresentados no Capítulo 3. O procedimento de construção do DLCM foi composto por um processo iterativo e incremental, no qual as definições e relacionamentos foram aprimorados, refinados e validados. No Quadro 1 são mapeados os trabalhos contribuintes por categoria.

O modelo é dividido em duas partes. A primeira reúne todos os elementos de metadados necessários para atendimento a solicitação de acesso aos dados, e a segunda parte é composta pelos agrupamentos desses metadados por categorias. Cada categoria possui um modelo associado e para representá-lo utilizamos a notação do metamodelo HANDLE proposto por (EICHLER et al., 2021) e discutido na Seção 3.3.3.

Visando facilitar o entendimento do modelo, inicialmente introduziremos o agrupamento dos elementos de metadados nas categorias: Operacional, Técnico, Jurídico, Segurança e Negócios, representados na Figura 18. Na sequência, iremos apresentar os modelos de cada categoria, descrevendo seus elementos, propriedades e relacionamentos com mais detalhes. Por fim, será apresentado o DCLM, consolidando todas as categorias.

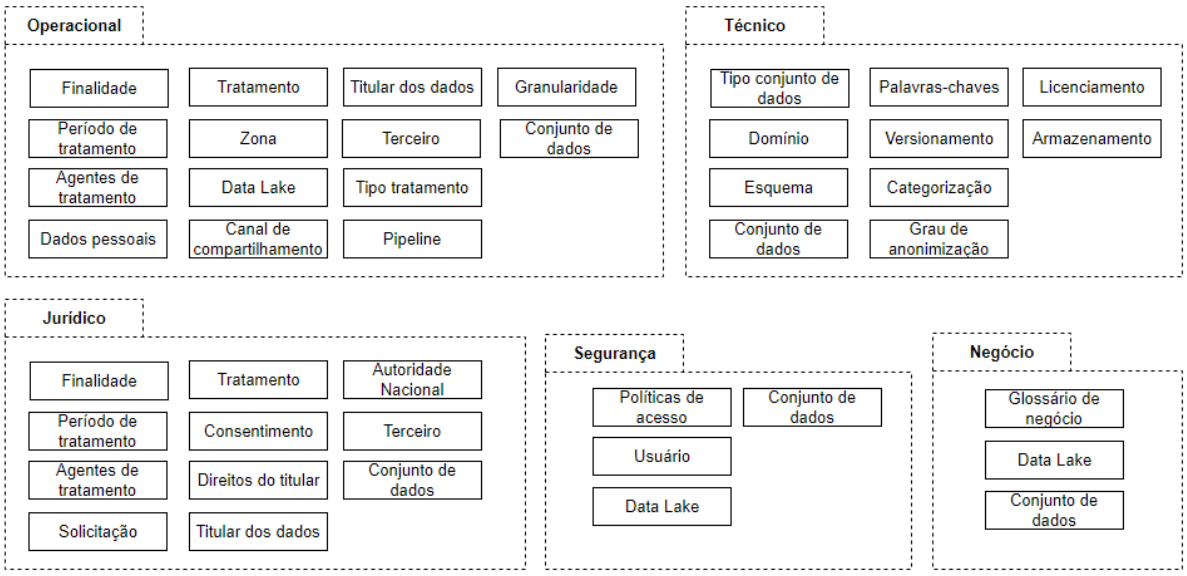
4.2 CATEGORIZAÇÃO DOS METADADOS

Seguindo a mesma ideia de classificação proposta por (DIAMANTINI et al., 2018) descrita na Seção 2.4.1, agrupamos os metadados em cinco categorias, sendo elas: Técnico, Operacional, Jurídico, Negócio e Segurança. Cada categoria reúne um conjunto de elementos do modelo conceitual e são representadas por pacotes, conforme mostra a Figura 18. A classificação dos metadados em categorias tem como objetivo facilitar a compreensão do modelo, tornando possível o mapeamento de papéis responsáveis por sua coleta, como também quais os processos para coleta e uso mais eficiente.

Os elementos que compõem a categoria de metadados Jurídico dizem respeito as noções legislativas da LGPD. Na categoria de metadados Operacional, estão agrupados os elementos que refletem as operações de tratamento aplicadas nos dados. Já na categoria Técnico, estão presentes os elementos que representam o conjunto de dados e sua infraestrutura de armazenamento. Por fim, no pacote de Segurança estão alocados os itens que identificam o usuário e seus acessos, enquanto que na categoria Negócio, estão reunidos os elementos que abordam as regras de negócio da organização, os quais são responsáveis pelo alinhamento entre o setor de negócio e o setor da TI.

Somado a isso, é possível que alguns metadados compartilhem mais de uma categoria, como é o caso dos metadados descritivos de um conjunto de dado e daqueles associados aos tratamentos. Cada pacote possui um modelo associado, que detalha as propriedades dos elementos e seus relacionamentos. A seguir, serão apresentados os modelos de cada categoria.

Figura 18 – Categorização dos Metadados.



Fonte: a autora, 2022

4.2.1 Operacional

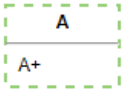
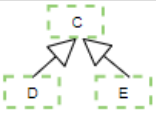
De acordo com (ZGOLLI; COLLET; MADERA, 2020), metadados operacionais descrevem as integrações de aplicativos de dados e seus fluxos de execução. Eles fornecem recurso para rastrear o processamento de dados entre sistemas, auxiliando na solução de problemas relacionados aos dados em si e sua qualidade.

Os elementos do pacote Operacional representam as operações de tratamento realizadas em um conjunto de dados no Data Lake. O intuito principal na coleta destes metadados consiste em auxiliar o Controlador dos Dados a dispor de recursos para atender a uma solicitação de acesso aos dados, identificando quais dados ele possui de um determinado individuo e rastreando todas as alterações aplicadas sobre eles em uma zona do Data Lake.

Utilizando a notação do HANDLE (EICHLER et al., 2021), o DCLM comporta os elementos da categoria Operacional e seus subtipos. Os símbolos retangulares pontilhados em verde representam objetos de metadados, que por sua vez, descrevem os dados, representados pelos símbolos retangulares sólidos na cor azul. E em amarelo, são representados os objetos de metadado que indicam a zona do DL e a granularidade do elemento de dado. Um resumo dos símbolos do metamodelo é apresentado na Figura 19.

Alguns objetos são definidos por propriedades. Os valores atribuídos a essas propriedades são relevantes para responder a solicitação de acesso aos dados, como por exemplo, o nome do Operador na entidade Tratamento, identifica quem é o agente responsável pelo tratamento, da mesma forma, a propriedade Finalidade explica o propósito no qual o tratamento se baseia. Com estas informações é possível tanto verificar se os dados estão sendo tratados de acordo com o consentimento do titular ou mesmo responsabilizar um operador pelo mau uso dos dados.

Figura 19 – Notação do Metamodelo HANDLE.

Símbolos	Descrição
	Representa um objeto de metadado.
	Representa os objetos de metadados que indicam a zona do Data Lake e a granularidade do objeto de dado.
	Representa um objeto de dado.
	A+ é uma propriedade que define o metadado A.
	D e E são subtipos de C e herdam todas as propriedades de C.

Fonte: a autora, 2022

No modelo apresentado na Figura 20, um conjunto de dados pode receber um ou vários tratamentos. Ele também é caracterizado por sua granularidade, que indica se ele é formado por uma tabela, um arquivo, ou apenas uma linha. Cada tratamento é executado por um *Pipeline* e obedece a uma finalidade específica dentro de um período. Como subtipos de tratamento temos:

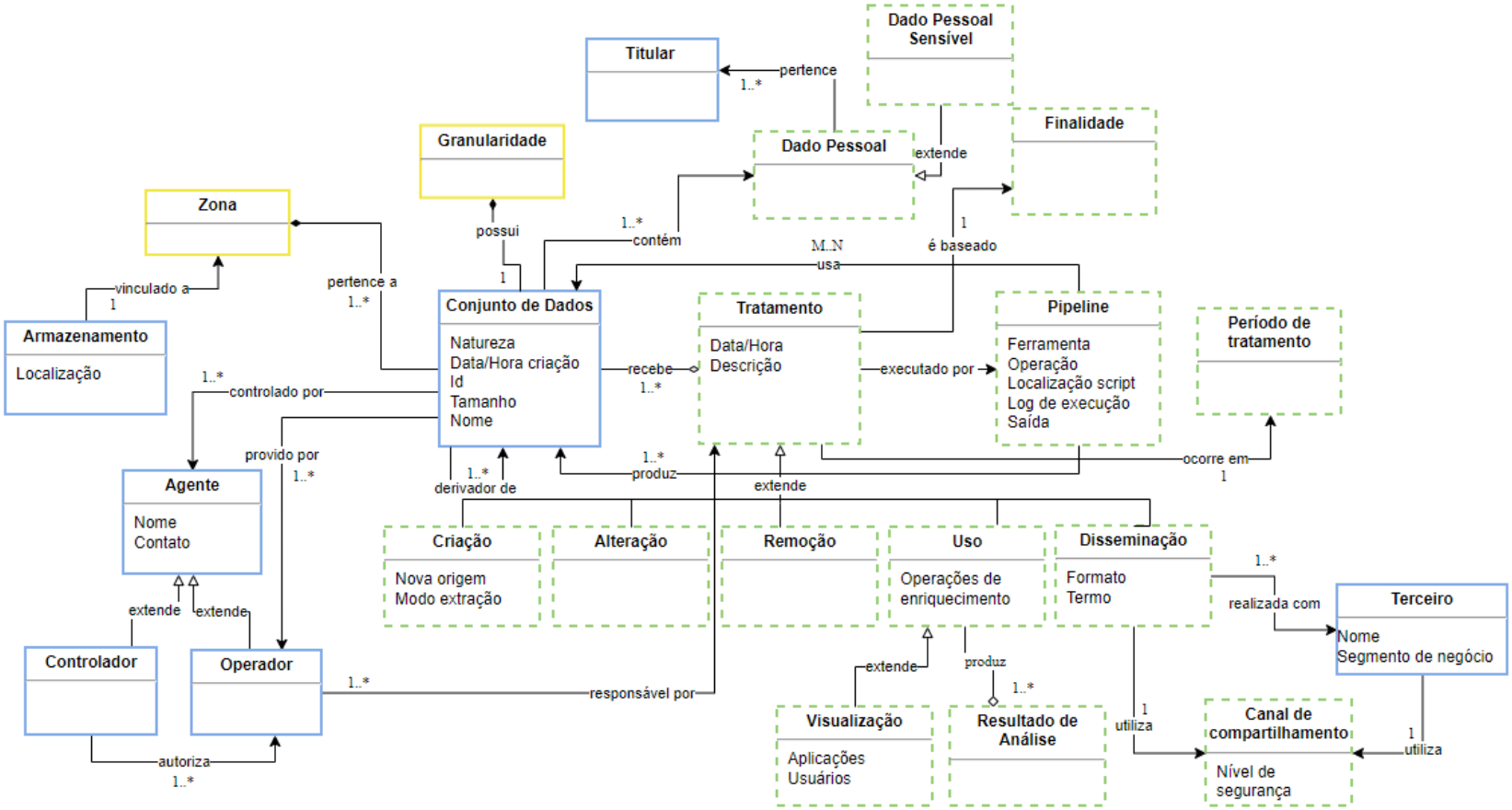
- *Criação*: Tratamento responsável por criar e extrair novos conjuntos de dados. Suas propriedades armazenam valores sobre a origem e o modo de extração.
- *Alteração*: Tratamento que realiza operações de formatação, agregação, consolidação e junção dos dados.
- *Remoção*: Tratamento responsável por deletar um dado ou um conjunto de dados.
- *Uso*: Tratamento que compreende por operações de análise e/ou enriquecimento, como por exemplo, padronização, filtragem, remoção de *outliers*, limpeza, dentre outros. Além disso, pode produzir resultados de análise, como relatórios e *dashboards* analíticos.

  Visualização: A visualização é um subtipo de uso que pode ser executado por usuários e aplicações.
- *Disseminação*: Tratamento responsável por compartilhar ou transferir o conjunto de dados com terceiros por meio de um canal de compartilhamento.

É importante ressaltar que o conjunto de dados está vinculado a uma zona do Data Lake. Logo, todo tratamento aplicado será mantido apenas no conjunto de dados da zona em questão e não replicado para as demais. Porém, um conjunto de dados pode ser usado como input para derivação ou criação de um novo conjunto de dados. Essa informação também é registrada nos metadados do Pipeline.

Com estes metadados podemos rastrear um conjunto de dados e todos os tratamentos pelos quais ele passou em uma zona do Data Lake. Somado a isso, também é possível identificar o Operador responsável pela execução, quando e como ela ocorreu. Tendo em vista que o elemento tratamento armazena valores que descrevem o tratamento e o seu propósito, o controlador dos dados terá como comprovar que o tratamento aplicado corresponde a finalidade concedida pelo Titular dos dados. Sendo assim, é possível gerenciar não só os tratamentos realizados mas também o ciclo de vida de um conjunto de dados no Data Lake.

Figura 20 – DLCM - Modelo de Metadados Operacional.



Fonte: a autora, 2022

4.2.2 Técnico

Metadados técnicos expressam como os dados são representados, incluindo seu formato, estrutura ou esquema e como eles podem ser acessados (SAWADOGO; DARMONT, 2021). Segundo (LÓSCIO; BURLE; CALEGARI, 2016), fornecer explicitamente informações descritivas do conjunto de dados permite a descoberta automática do objeto de dados e faz com que os humanos entendam sua natureza e distribuições.

No pacote de metadados Técnico estão reunidos os elementos que descrevem e contextualizam o conjunto de dados no Data Lake. Seguindo as boas práticas propostas por (LÓSCIO; BURLE; CALEGARI, 2016), todo conjunto de dados possui informações de licenciamento. A presença desses metadados é essencial para que os Consumidores de Dados avaliem quais os limites de uso que um usuário pode ter em relação a um conjunto de dados. Somado a isso, é uma informação útil para identificar as restrições de compartilhamento e reutilização do conjunto de dados.

Em Data Lakes, conjuntos de dados podem ser modificados para correção ou atualização de dados. Conjuntos de Dados também podem ser modificados na transição de uma zona para outra. Para lidar com essas mudanças, é importante o versionamento apropriado dos conjuntos de dados. No DLCM, cada nova versão é definida por um número que representa a versão anterior, outro número que representa a versão atual, data e hora da última atualização e uma descrição sobre as principais mudanças que ocorreram e diferenciam aquela nova versão da anterior. Esses metadados permitem acompanhar o histórico de mudanças.

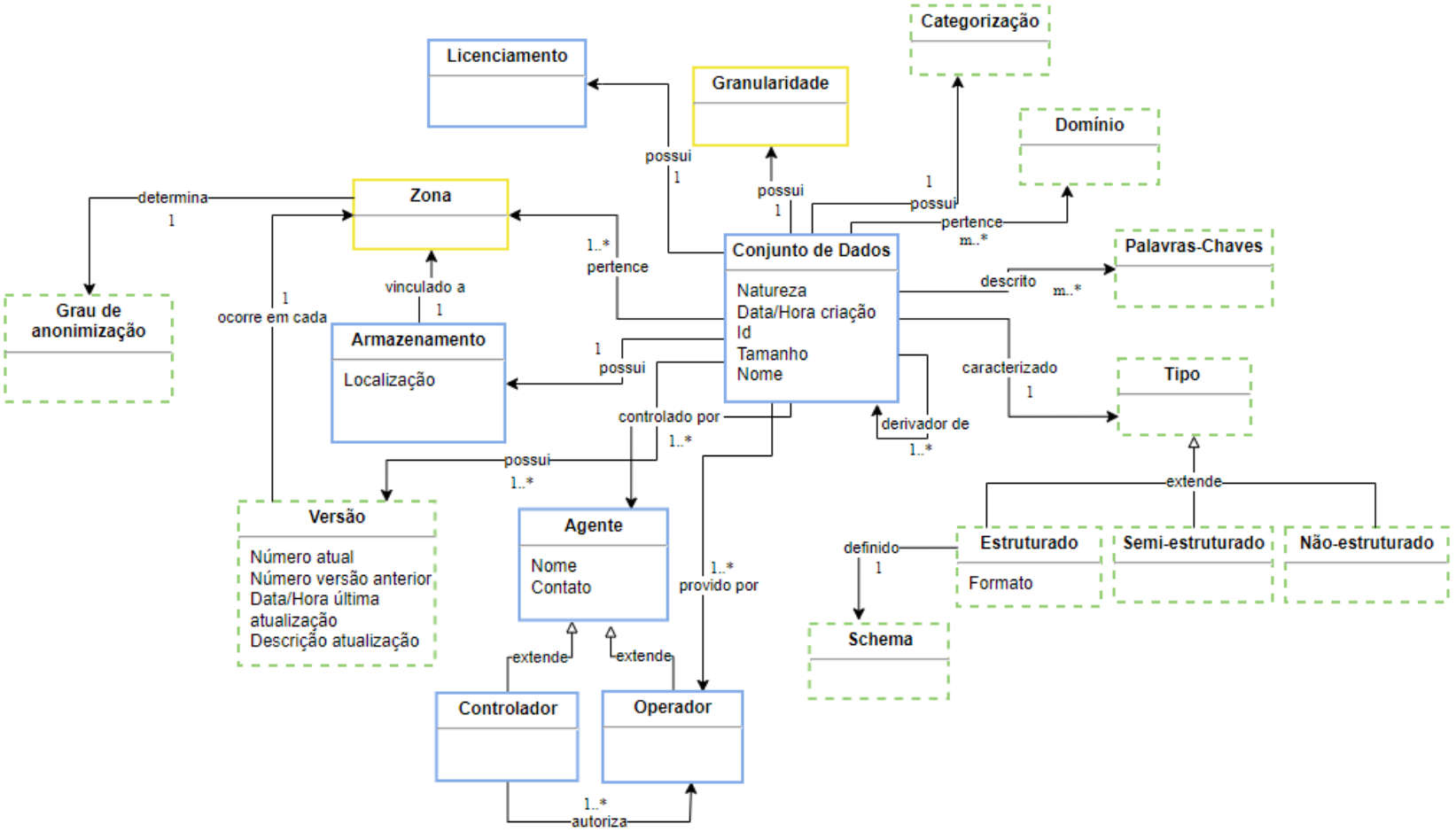
Diferentemente dos tradicionais Data Warehouses, que lidam apenas com a ingestão de dados estruturados, Data Lakes integram conjuntos de dados de diferentes tipos, sendo eles: Estruturado, semi-estruturado ou não-estruturado. Um conjunto de dados estruturado é definido por um formato, que pode ser um arquivo CSV, XML, JSON, tabela ou uma estrutura multi-dimensional, e associado a estes formatos haverá também um esquema representando os tipos de atributos existentes no conjunto de dados.

As zonas que dividem o Data Lake podem armazenar conjuntos de dados em diferentes granularidades. No exemplo demonstrando em (EICHLER et al., 2021), o mesmo conjunto de dados é armazenado duas vezes, primeiro na zona de dados brutos de forma mais detalhada e depois na zona confiável como uma tabela mais resumida. No DCLM, capturamos essa informação utilizando o objeto de metadados *Zona*. Considerando que os dados podem ser anonimizados, essa operação também irá depender da zona na qual o conjunto de dados está localizado. De acordo com (PANDIT; LEWIS, 2017), o dado pode ser completamente anonimizado, sendo impossível reverter a operação, pseudo-anonimizado, permitindo alguma flexibilidade para reversão, ou não anonimizado, sem nenhuma camuflagem associada as informações.

Por fim, temos o conjunto de dados associado a um ou mais Domínios, e o recurso de Palavras-chaves que o descrevem abreviadamente. Como propriedades que compõe o conjunto de dados temos a natureza, data e hora de sua criação ou disponibilização, código identificador, tamanho e nome. É importante ressaltar que o conjunto de dados também pode ser derivador

de novos conjuntos de dados. Ademais, como responsáveis por prover e controlar o conjunto de dados, existem os Agentes de tratamento, subdivididos em Controlador e Operador dos dados. Na Figura 21 é apresentado o modelo da categoria de metadados Técnico.

Figura 21 – DLCM - Modelo de Metadados Técnico.



Fonte: a autora, 2022

4.2.3 Jurídico

As leis de proteção de dados, LGPD e GDPR, introduziram importantes modificações na forma como os dados são obtidos e processados por organizações, sejam elas públicas ou privadas. Uma das principais mudanças envolve o consentimento, que fornece informações para o Titular dos dados sobre os dados coletados e qualquer uso pretendido, incluindo armazenamento e compartilhamento com terceiros (PANDIT; LEWIS, 2017). O pacote de metadados Jurídico engloba todos os aspectos legais que representam a LGPD. Seu principal objetivo é descrever os conceitos da legislação, como por exemplo, os Agentes de tratamento e os direitos dos Titulares de dados.

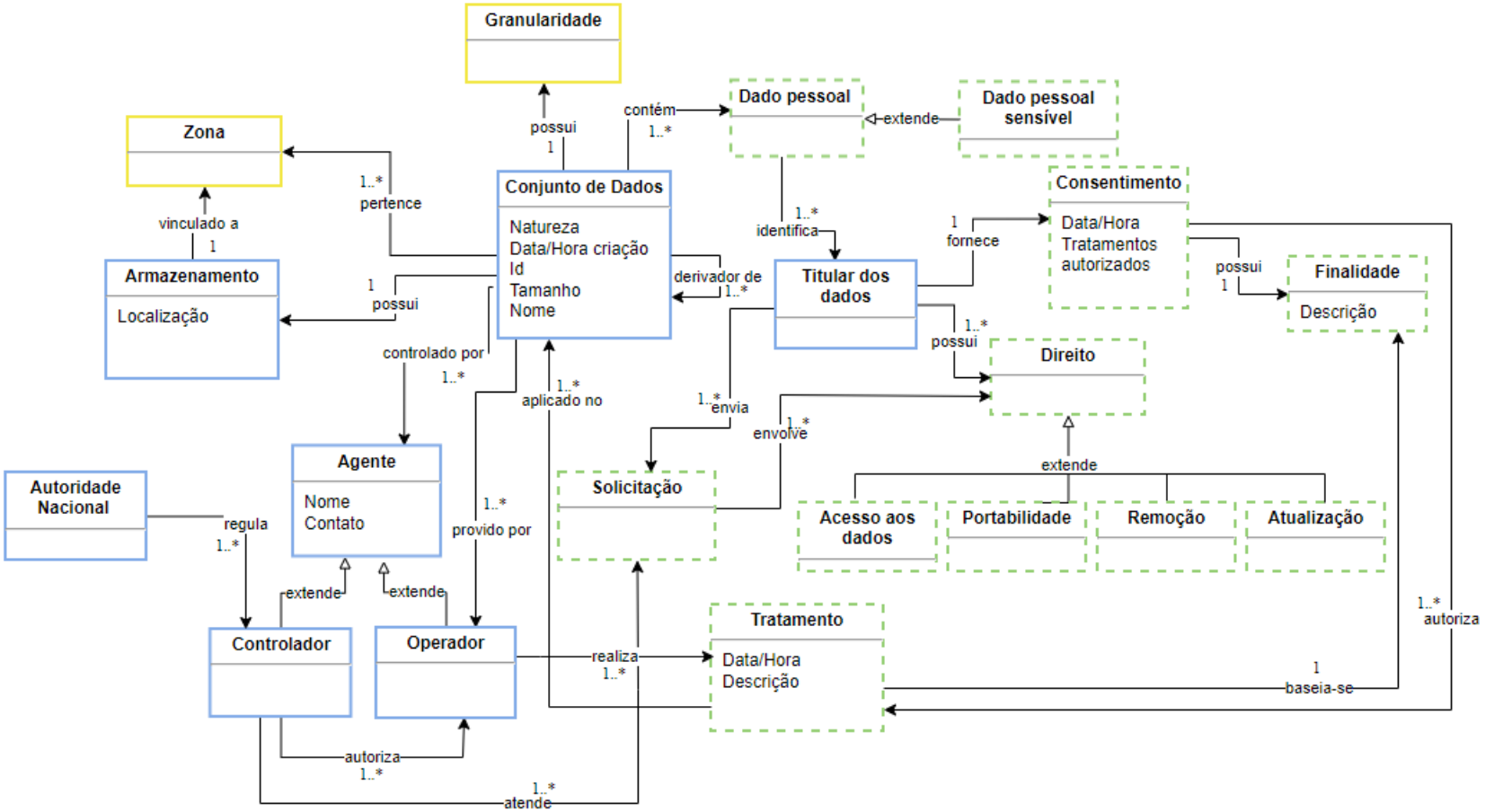
Na Figura 22, temos novamente o elemento central Conjunto de Dados, atrelados a ele estão os objetos de metadados que indicam a zona na qual conjunto de dados está armazenado, e sua granularidade. De acordo com a (LEI..., 2018), os Agentes de tratamento são classificados em Controlador e Operador. Além de prover e controlar o conjunto de dados no DL, estes agentes são os responsáveis por realizar os tratamentos no conjunto de dados descritos no modelo Operacional.

O Controlador dos dados representa a organização, seja ela pública ou privada, que possui o Data Lake e trata dados pessoais. Ele é regulado pela Autoridade Nacional de Proteção de Dados - ANPD, que por sua vez, tem como objetivo assegurar o cumprimento da legislação. Já o Operador dos dados é o ator quem opera em nome do Controlador, ambos são definidos pelas propriedades descritas na entidade Agente. Estas propriedades armazenam informações que identificam os agentes e seus respectivos contatos.

Como parte de suas atribuições, o Controlador atende a uma ou mais solicitação. Esta requisição parte do Titular dos dados, que é a pessoa natural cujos dados são os objetos de tratamento. O elemento Direito representa especificamente os artigos 18 e 20 da LGPD, onde são informados os direitos do Titular dos dados. Como subelementos de direito, temos o acesso aos dados, portabilidade, remoção e atualização. Na Seção 2.3.1 é fornecida uma descrição mais enriquecida sobre eles.

Consentimento é um dos elementos mais importantes do modelo, pois é a partir dele que o tratamento é autorizado. Ele também é a base para verificação da conformidade, tendo em vista que suas propriedades armazenam valores que refletem o desejo do Titular dos dados, isto é, os tratamentos autorizados. Além disso, o consentimento está vinculado a uma finalidade que determina o tratamento. Essa finalidade é definida por uma descrição que informa o objetivo do tratamento. Completando os metadados de caráter Jurídico, temos os elementos dado pessoal e sua extensão dado pessoal sensível, que podem ser considerados como objetos de proteção da LGPD.

Figura 22 – DLCM - Modelo de Metadados Jurídico.



Fonte: a autora, 2022

4.2.4 Segurança

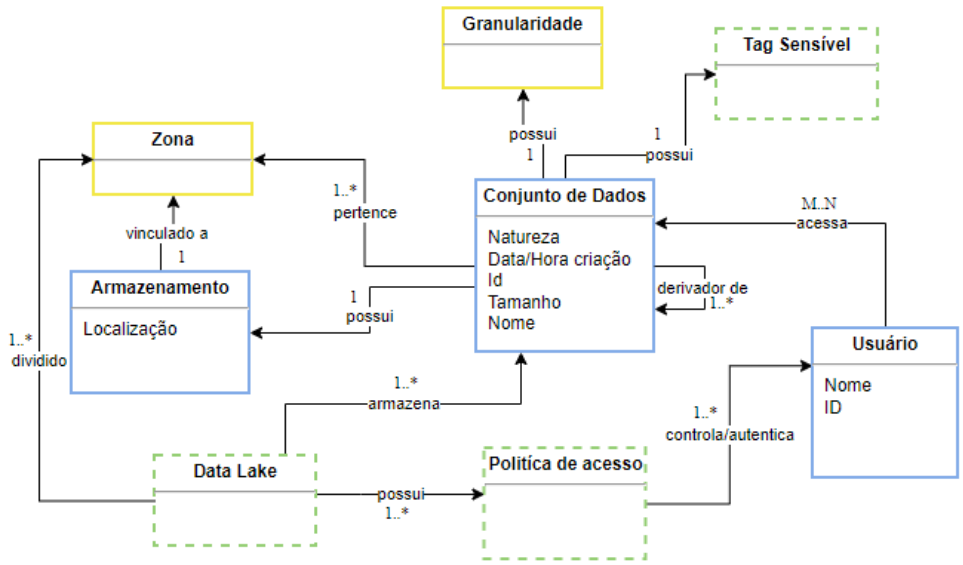
Os elementos que compõem o modelo de Segurança tem como objetivo representar os metadados que dizem respeito aos logins dos usuários no Data Lake, e consequentemente, seus acessos aos conjunto de dados. Considerando que o foco da proposta é dar subsídios para que o Controlador dos dados consiga atender a solicitação de acesso aos dados efetuada pelo Titular, o modelo apresentado na Figura 23 incorpora essencialmente informações sobre os mecanismos de controle de acesso dos usuários no Data Lake.

Como parte dos requisitos legais e de segurança, todo conjunto de dados deve indicar se armazena dados sensíveis. No DLCM, além dos indicadores de zona e granularidade, o conjunto de dados também estabelece relacionamento com a entidade Tag Sensível para cumprir com este propósito.

O elemento Política de acesso representa todas as informações de restrições de acesso ao conjuntos de dados do Data Lake. Tais políticas são responsáveis por autenticar os usuários bem como controlar seus acessos aos dados. Todo usuário é definido por propriedades que armazenam seu nome e código único de identificação (ID), respectivamente.

Com estes metadados é possível identificar o acesso de um ou vários usuários, e consequentemente, gerenciar o uso de um conjunto de dados no Data Lake. Somado a isso, também é factível distinguir se o conjunto de dados contém dados sensíveis ou não, bem como sua localização e granularidade.

Figura 23 – DLCM - Modelo de Metadados de Segurança.



Fonte: a autora, 2022

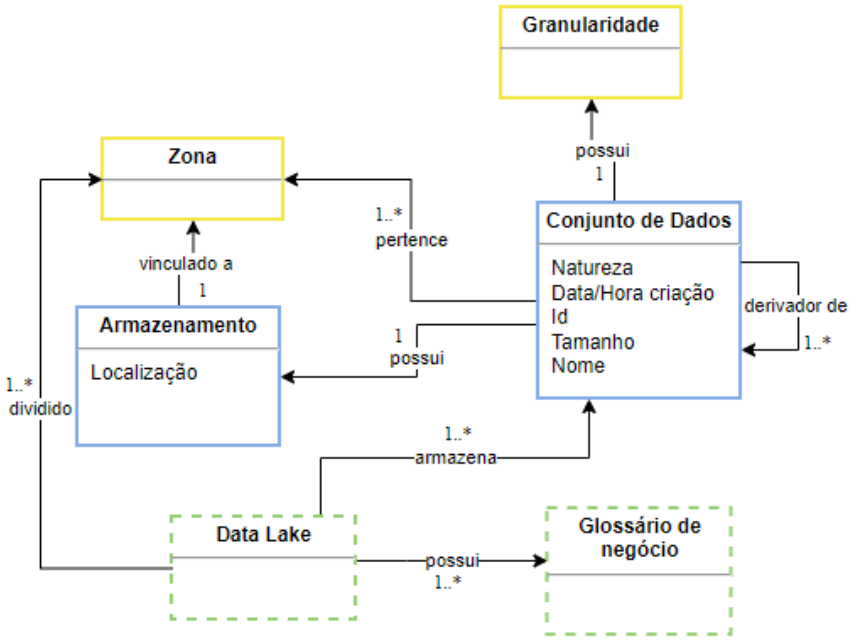
4.2.5 Negócio

Metadados de negócio constituem um aspecto importante em qualquer programa de governança de informações bem sucedido (ZGOLLI; COLLET; MADERA, 2020). Em Data Lakes, eles são relevantes para fornecer contexto sobre os dados armazenados. Pois com eles, os usuários finais podem compreender melhor os conjuntos de dados disponíveis e seu conteúdo.

Em nosso modelo, apresentado na Figura 24, metadados de negócio são representados pelo elemento Glossário de Negócio. Em Data Lakes, dados são ingeridos sem esquemas ou objetivos de negócio pré-definidos, devido a isso, o elemento atua apenas como um objeto genérico, que deve ser definido de acordo com as particularidades de cada Data Lake. Este elemento também pode representar metadados sobre os processos de negócio, descrevendo as atividades, *stakeholders* envolvidos, modelos BPMN, e todos os objetos que compõe um projeto.

O elemento Data Lake também é caracterizado como uma entidade genérica sem propriedades definidas, que deve representar essencialmente o ambiente no qual os tratamentos ocorrem. Resumidamente, um Data Lake é dividido por zonas, que armazenam conjuntos de dados em diferentes granularidades. Como parte de sua estrutura, ele possui um ou mais Glossário de negócio, que além de prover informações sobre regras e processos de negócio, pode atuar como referência dos conjuntos de dados.

Figura 24 – DLCM - Modelo de Metadados de Negócio.



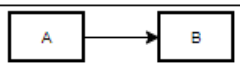
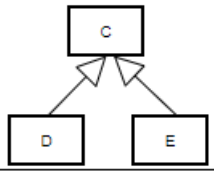
Fonte: a autora, 2022

4.3 MODELO CONCEITUAL - DLCM

Anteriormente, foram apresentados os modelos que agrupam os metadados por categorias. A primeira categoria, Operacional, propõe-se a descrever os tratamentos realizados nos conjuntos de dados do Data Lake. A segunda categoria nomeada de Técnico, descreve as características dos conjuntos de dados e suas derivações. A terceira categoria, que representa os metadados de perfil Jurídico, objetiva relacionar os conceitos da LGPD com os dados do Data Lake, enquanto que o grupo de Segurança, especifica as políticas de acesso que gerenciam os usuários do DL. Por último, a categoria Negócio, descreve os aspectos de negócios que fornecem contexto para os usuários finais do Data Lake.

A consolidação de todos os elementos das categorias descritas na Seção 4.2 resultam no modelo conceitual DLCM. Para representá-lo utilizamos a notação UML, no qual as caixas retangulares simbolizam os elementos, e as setas reproduzem as relações entre eles. É possível que um elemento tenha subelementos, representamos essa classificação utilizando o conceito de herança em UML. Um breve guia para os símbolos da notação UML usados no modelo é mostrado na Figura 25.

Figura 25 – Símbolos UML.

Símbolos notacionais	**Descrição**
	A é um elemento do modelo conceitual.
	A seta indica um relacionamento entre os elementos.
	O elemento A tem relação com o elemento B.
	Representa uma herança
	D e E são subtipos de C.

Fonte: a autora, 2022

No modelo, os elementos centrais com maior influência sobre os outros elementos são: *Conjunto de Dados* e *Tratamento*. Um *Conjunto de Dados* é caracterizado por um *Tipo* e pode receber um ou mais tratamentos. Todo *Tratamento* baseia-se em uma *Finalidade*, que é expressa no *Consentimento* fornecido pelo *Titular dos dados* e ocorre em um *Período de tratamento*. A execução de um tratamento é realizada através de uma *Pipeline*, que pode fazer uso de um conjunto de dados existente como entrada, do mesmo modo que produz um novo conjunto de dados como saída.

O conjunto de dados contém *Dados pessoais* que identificam o Titular dos dados. Atrelado a ele, há outras informações, como *Licenciamento*, *Versionamento*, *Esquema* e *Domínio*. Além

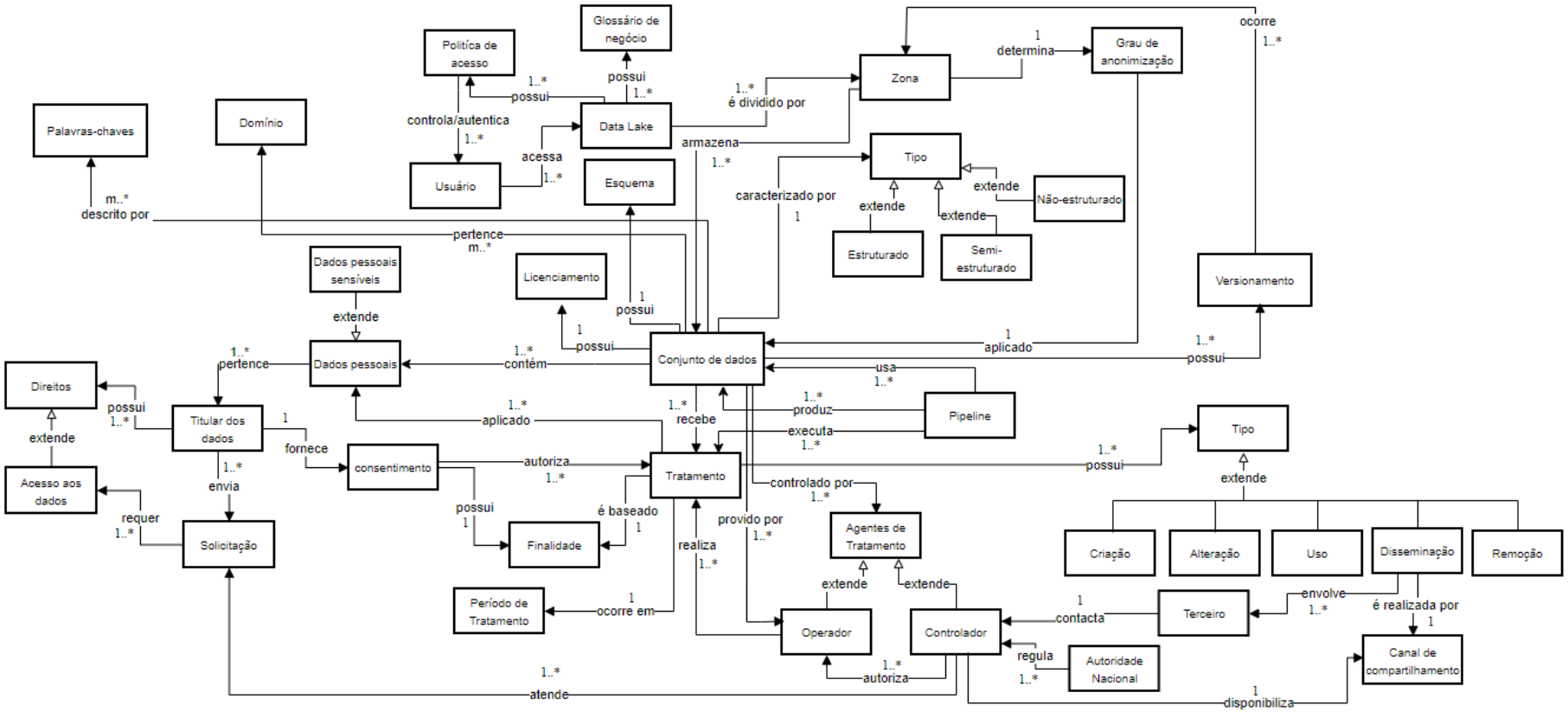
disso, pode ser descrito por *Palavras-chaves* e armazenado em uma *Zona* do *Data Lake*. Tal zona determina o *Grau de anonimização* do conjunto de dados.

Os *Agentes de tratamento* são os responsáveis por controlar o conjunto de dados e se subdividem em *Controlador* e *Operador*. O Operador cria o conjunto de dados e executa os tratamentos em nome do Controlador. O controlador por sua vez, tem a obrigação de atender a *Solicitação* do Titular dos dados, quando este exercer seu *Direito* de *Acesso aos dados*.

Criação, *Alteração*, *Uso*, *Disseminação* e *Remoção* são subtipos de Tratamento. A Disseminação do conjunto de dados é realizada com um *Terceiro* através de um *Canal de compartilhamento* disponibilizado pelo Controlador. Também há a figura da *Autoridade Nacional* que regula o Controlador dos dados.

Por fim, temos o *Glossário de negócio* que contém informações relevantes do Data Lake, como por exemplo, regras de negócio e restrições. E para controle e segurança, o Data Lake possui *Políticas de acesso* responsáveis pela autenticação dos *Usuários* no ambiente. Na Figura 26 é apresentado o DLCM.

Figura 26 – Modelo Conceitual para Conformidade de Data Lake com a LGPD.



Fonte: a autora, 2022

4.4 CONSIDERAÇÕES FINAIS

Este Capítulo apresentou uma proposta de Modelo de Suporte para Conformidade de Data Lake com a LGPD, chamado DLCM. Esse modelo foi construído a partir do estado da arte de modelos de metadados propostos na literatura. Além disso, seu desenvolvimento foi constituído por um processo interativo e incremental, onde os elementos de metadados foram identificados, aprimorados, refinados e validados até chegarmos a versão disponibilizada no presente trabalho.

O DLCM é dividido em duas partes, a primeira agrupa os metadados em cinco categorias: Operacional, Técnico, Jurídico, Segurança e Negócio. Cada Categoria possui um modelo associado que melhor detalha seus metadados, propriedades e relacionamentos. Já a segunda parte, consiste em um modelo conceitual que relaciona todos os elementos das cinco categorias, fornecendo uma visão unificada dos metadados que darão suporte ao Controlador dos dados, na evolução em termos de conformidade com a LGPD.

Em contraste aos modelos apresentados no Capítulo 3, nosso modelo se diferencia por além de definir e relacionar metadados sobre conceitos técnicos, operacionais e jurídicos, auxilia na rastreabilidade dos conjuntos de dados armazenados no Data Lake. Para melhor compreensão de como os trabalhos existentes na literatura contribuíram com o DLCM, elaboramos o Quadro 1, mapeando por categoria os artigos e seus respectivos autores.

Quadro 1 – Contribuição dos Trabalhos da Literatura para o DLCM

Categorias	**Trabalhos**					
Operacional	Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies Pandit e Lewis (2017)	A Big Data Provenance Model for Data Security Based on Prov-DM Model Gao, Chen e Du (2020)	Metadata Management on Data Processing in Data Lakes Megdiche, Ravat e Zhao (2021)	DCAT Consortium et al. (2014)	Metadata in Data Lake Ecossystems Zgolli, Collet e Madera (2020)	Data Lake Ingestion Management Zhao, Megdiche e Ravat (2021)
Técnico	Metadata in Data Lake Ecossystems Zgolli, Collet e Madera (2020)	Data on the Web Best Practices Lóscio, Burle e Calegari (2016)	Data Lake Ingestion Management Zhao, Megdiche e Ravat (2021)	DCAT Consortium et al. (2014)	Schema.org Guha, Brickley e Macbeth (2016)	
Jurídico	Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies Pandit e Lewis (2017)	A Provenance Model for the European Union General Data Protection Regulation Ujcich, Bates e Sanders (2018)				
Segurança	Data Lake Ingestion Management Zhao, Megdiche e Ravat (2021)	Metadata Management for Data Lakes Ravat e Zhao (2019b)	A Zone-Based Data Lake Architecture for IoT, Small and Big Data Zhao et al. (2021)			
Negócios	Metadata in Data Lake Ecossystems Zgolli, Collet e Madera (2020)					

Fonte: a autora, 2022

5 AVALIAÇÃO

Neste Capítulo, é discutida a avaliação do Modelo de Suporte para Conformidade de Data Lake com a LGPD proposto neste trabalho. Para essa avaliação, foi utilizado o método AIMQ (*Assessment Information Methodology Quality*). Dessa forma, a Seção 5.1 apresenta o que constitui o método de avaliação, enquanto que a Seção 5.2 descreve os resultados obtidos. Inicialmente, na Seção 5.2.1 caracterizamos os participantes, na sequência descrevemos a avaliação do modelo conceitual (Seção 5.2.2). Na Seção 5.2.3 são apresentados os resultados do modelo de categoria Operacional. Por fim, na Seção 5.3 são apresentadas as considerações finais do Capítulo.

5.1 MÉTODO DE AVALIAÇÃO

A etapa de avaliação tem como objetivo avaliar o modelo desenvolvido. Usualmente, uma avaliação de modelo implica fazer um julgamento técnico a respeito de sua exatidão. Também é importante avaliar sua praticidade e utilidade, bem como aspectos de qualidade, como por exemplo, completude e compreensibilidade. Neste sentido, a avaliação do DCLM teve os seguintes objetivos: (i) Validar a completude dos metadados; (ii) Avaliar a compreensividade, objetividade e relevância do modelo proposto; (iii) Refinar o modelo proposto.

Em (OLIVEIRA et al., 2018), é utilizada uma alternativa promissora que avalia a qualidade de um metamodelo com base na avaliação da Qualidade das Informações (QI) derivadas do metamodelo. Essa metodologia, amplamente referenciada por diversos trabalhos, é denominada *Assessment Information Methodology Quality* (AIMQ) e foi proposta por (LEE et al., 2002). Ela é constituída por um modelo de Qualidade da Informação, um questionário para medir as dimensões da Qualidade da Informação e técnicas para métricas de interpretação.

Em (LEE et al., 2002), os autores explicam que o modelo, primeiro componente da metodologia, possui quatro quadrantes e introduz os principais conceitos de Qualidade da Informação e seu significado para consumidores e gerentes de informação. Cada quadrante representa um agrupamento de dimensões por categoria, são elas: QI intrínseco, QI contextual, QI representacional e QI de acessibilidade. Já o questionário, segundo componente, pode ser utilizado para medir a QI ao longo das dimensões. O terceiro componente do AIMQ consiste em duas técnicas de análise para interpretar as avaliações captadas pelo questionário. Essas duas técnicas ajudam as organizações a concentrar seus esforços de melhoria de QI com base no resultado da análise de suas avaliações de QI.

(LEE et al., 2002) ainda completam afirmando que cada componente da metodologia AIMQ tem um mérito em si, sendo possível aplicá-los separadamente. O uso adequado dos componentes produz uma avaliação eficaz da Qualidade da Informação em ambientes organizacionais onde as decisões devem ser tomadas para priorizar tarefas e alocar recursos para a melhoria

da QI (LEE et al., 2002). Por esta razão, ao invés de utilizar qualquer questionário, optamos por desenvolver um questionário centrado nas dimensões do AIMQ para avaliar o DLCM, considerando a Qualidade da Informação que pode ser reunida a partir da aplicação do modelo.

Porém, em função da dificuldade de reunir avaliadores para o questionário, optamos por selecionar apenas uma parte das dimensões definidas pelo AIMQ. A ideia era diminuir o quantitativo de questões a serem respondidas, de forma a, consequentemente, reduzir o tempo final para conclusão do questionário. Quanto menor o tempo, maior a taxa de sucesso na resposta de questionários.

Assim, o questionário usado na nossa avaliação centrou-se em avaliar os aspectos essenciais de Qualidade da Informação. Considerando que o DLCM engloba o modelo conceitual e os modelos por categoria, utilizamos 4 dimensões para avaliação do modelo geral, e 2 dimensões para avaliar um dos modelos de categoria.

Considerando as dimensões da Qualidade da Informação do AIMQ, o modelo conceitual geral foi analisado de acordo com as seguintes dimensões:

- **Completude**: Indica se a percepção dos avaliadores do modelo contém todos os conceitos sobre LGPD e Data Lakes (LEE et al., 2002);
- **Objetividade**: Refere-se à medida em que o metamodelo é imparcial e isento de preconceitos (LEE et al., 2002);
- **Relevância**: Refere-se à facilidade com que o modelo pode lidar com o propósito da atividade de modelagem (LEE et al., 2002);
- **Compreensão**: Definida como a capacidade de compreender o significado do conhecimento (OLIVEIRA et al., 2018).

O modelo de categoria escolhido para avaliação foi o Operacional por representar os metadados que permitem acompanhar a rastreabilidade e linhagem dos dados. Estes metadados são essenciais para auxiliar no atendimento da maioria dos direitos dos Titulares de Dados definidos pela LGPD. Sua análise foi com base nas seguintes dimensões:

- **Facilidade de uso**: Significa a facilidade com que o modelo pode derivar outro modelo (LEE et al., 2002);
- **Interpretabilidade**: Refere-se à extensão em que o modelo é representado em linguagens, símbolos e unidades apropriadas, e se as definições são claras (LEE et al., 2002).

A ideia aqui é verificar se o metamodelo ajudou na interpretação dos metadados. Isto é, se os elementos estéticos definidos pelo HANDLE contribui positivamente para interpretação dos metadados.

A coleta dos dados da avaliação foi realizada por meio de um formulário eletrônico. Optamos em fazer uso do questionário devido a sua praticidade e flexibilidade, e também pelas limitações impostas pelo distanciamento social ainda vigentes em função da pandemia.

O questionário foi composto basicamente por questões fechadas adaptadas do AIQM, que tinham como objetivo avaliar a Qualidade das Informações derivadas do modelo proposto. Em resumo, as perguntas de caracterização dos participantes em sua maioria eram fechadas, tendo apenas uma em aberto. As dimensões escolhidas do AIMQ para avaliação eram compostas por afirmações que tinham uma escala variável de 1 a 10, na qual o número 1 corresponde a "Não concordo com a afirmação" e o número 10 corresponde a "Concordo com a afirmação". Adicionamos também perguntas objetivas de "Sim"ou "Não"para questões comparativas a respeito da notação dos modelos, e no final do questionário, deixamos um espaço em aberto para os participantes enviar suas críticas e sugestões de melhoria. Na próxima Seção, serão apresentados os resultados obtidos.

5.2 ANÁLISE DOS DADOS

Esta Seção apresenta a análise dos dados coletados no formulário eletrônico, discutindo detalhadamente cada dimensão e apontando algumas questões que devem ser levadas em consideração.

A avaliação foi segmentada em três partes, a primeira caracteriza os participantes, enquanto que a segunda analisa o modelo conceitual geral, e a terceira parte, analisa um dos modelos de categoria, especificamente o Operacional. A seguir, apresentaremos os resultados do DLCM, iniciando pela descrição dos participantes que responderam o questionário eletrônico.

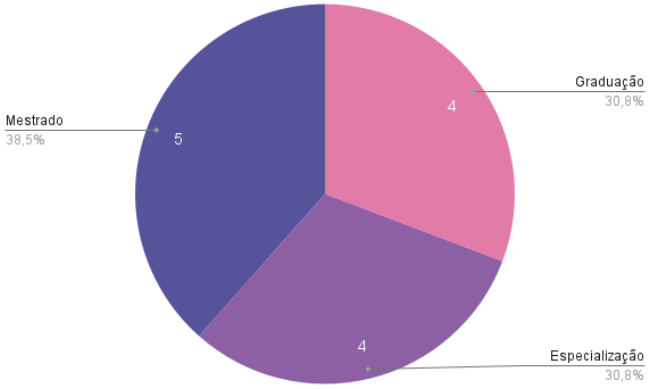
5.2.1 Caracterização dos Participantes

Inicialmente, entramos em contato com os participantes através de mensagem por telefone, explicando brevemente sobre o que se tratava a avaliação, o tempo médio necessário para preenchimento, e por fim, solicitando a sua participação. Também solicitamos aos candidatos algum e-mail para envio do formulário de avaliação e um resumo explicativo sobre a contextualização do problema de pesquisa e como os modelos podem contribuir na resolução deste. Após o envio para todos os participantes, o questionário ficou disponível por um período de cinco dias. Não houve a divulgação em outros meios, pois a avaliação exigia um certo grau de conhecimento sobre tópicos como Data Lake e LGPD. Sendo assim, optamos por selecionar os participantes manualmente. No final, solicitamos a participação de 14 candidatos, dentre eles 13 responderam o questionário.

No que diz respeito ao cargo/função que eles exercem atualmente, obtivemos como resposta 2 docentes do ensino técnico e superior, 2 gerentes de projeto, 1 analista de sistemas, 1 analista de *Business Intelligence*, 3 engenheiros de dados, 3 cientistas de dados e 1 desenvolvedor de software nível sênior. Na Figura 27 é possível visualizar que, dentre os participantes,

30,8% possuem graduação completa, 38,5% possuem mestrado e os demais que representam 30,8%, possuem especialização.

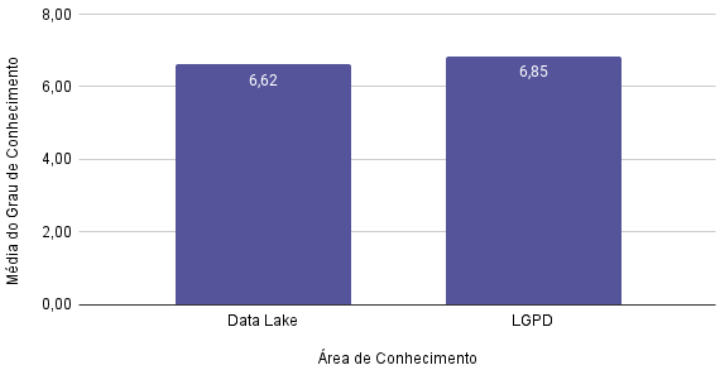
Figura 27 – Nível de Formação dos Participantes.



Fonte: a autora, 2022

Considerando o conhecimento sobre Data Lake e a LGPD, numa escala que vai de 0 a 10, na qual 0 significa nenhum conhecimento sobre os conceitos e 10 representa um conhecimento avançado, a média de conhecimento dos participantes sobre Data Lake ficou em 6,62, já a média de conhecimento acerca da LGPD ficou em 6,85, afirmando que os respondentes conhecem um pouco mais sobre a lei do que sobre a tecnologia de Data Lake. Na Figura 28 é possível visualizar estes resultados.

Figura 28 – Média do Grau de Conhecimento em Data Lake e LGPD.



Fonte: a autora, 2022

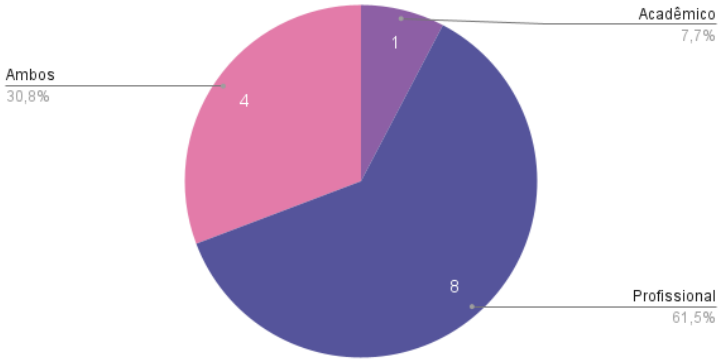
Nove dos respondentes informaram que tem participado direta ou indiretamente de atividades que envolvem Data Lakes. 4 dos respondentes atuam há menos de 2 anos, 3 atuam de 2 há menos de 3 anos, e os últimos 2, atuam de 3 há menos de 4 anos.

Eles ainda caracterizaram sua participação nessas atividades. Cinco responderam que atuam apenas como consumidores de dados do Data Lake, enquanto que os 4 restantes, responderam

que atuam como consumidor de dados e também como operador responsável por tratar os dados no Data Lake.

Acerca do contexto no qual essas atividades são desenvolvidas, o gráfico da Figura 29 apresenta que o maior percentual dos participantes atuam no contexto profissional, seguido do profissional e acadêmico, tendo apenas uma pequena parcela atuando exclusivamente no meio acadêmico.

Figura 29 – Meio em que os Participantes Desenvolvem as Atividades.



Fonte: a autora, 2022

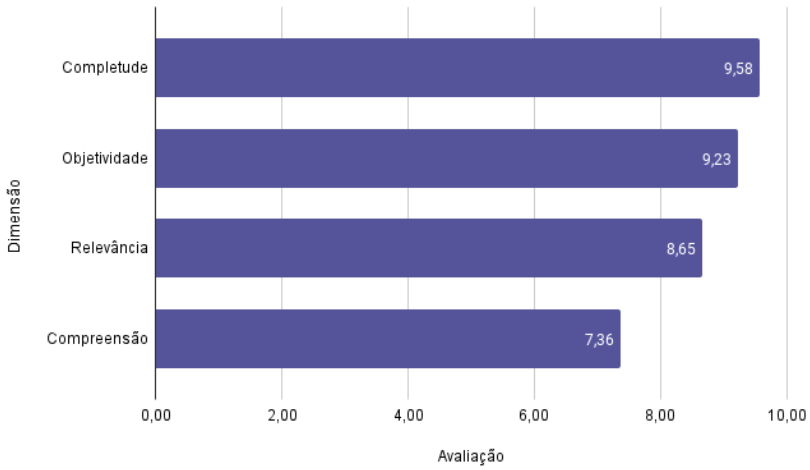
5.2.2 Avaliação de Completude, Objetividade, Relevância e Compreensão do DLCM

A avaliação do modelo conceitual buscou analisar as informações derivadas do modelo de acordo com um conjunto de dimensões de Qualidade da Informação. Essas dimensões representam a identificação de atributos considerados fundamentais para a compreensão do modelo e sua utilidade. Cada dimensão foi avaliada usando um conjunto de declarações.

Os participantes responderam as afirmações utilizando uma escala com valores que variam de 1 a 10, na qual o número 1 corresponde a "Não concordo com a afirmação" e o número 10 corresponde a "Concordo com a afirmação". A Tabela 2 apresenta os resultados por dimensão, detalhando a média obtida por declaração.

De acordo com o gráfico da Figura 30, as dimensões Relevância e Compreensão obtiveram as notas mais baixas na avaliação. Isso parece refletir as restrições relacionadas ao grau de conhecimento sobre modelagem e os conceitos de Data Lake e LGPD para alguns dos participantes.

Figura 30 – Média de Avaliação Geral das Dimensões.



Fonte: a autora, 2022

Algumas declarações dessas dimensões estão escritas na forma negativa, de acordo com (LEE et al., 2002), como por exemplo, a declaração "*Os significados dos conceitos do modelo são difíceis de entender*", utilizada na dimensão Compreensão. No caso em que as afirmações estão no sentido negativo, quanto menor a nota, mais positivo é o resultado, pois compreende-se que o participante considera o oposto da sentença.

De modo a não prejudicar a média geral da avaliação, as notas dessas sentenças escritas no sentido negativo foram normalizadas. Isto é, para cálculo da média, utiliza-se o valor que falta para alcance da nota máxima. Na declaração "*Os significados dos conceitos do modelo são difíceis de entender*"presente na dimensão Compreensão, a média calculada com as notas atribuídas originalmente pelos participantes é **3,85**, normalizando a mesma, obtemos o valor **6,15**, conforme apresenta o Quadro 2.

Em geral, esses resultados indicam que o modelo foi capaz de derivar conhecimento relevante para os participantes e apresentá-lo adequadamente. Somado a isso, os participantes também responderam a seguinte pergunta objetiva a respeito da completude do metadados do modelo: **O modelo apresentado dispõe dos metadados necessários que devem ser coletados no Data Lake para atender ao objetivo proposto?** 12, dentre os 13 participantes afirmaram que **Sim, os metadados são suficientes** para atender ao objetivo proposto.

5.2.3 Avaliação de Facilidade de Uso e Interpretabilidade do DLCM

A terceira etapa da avaliação buscou analisar a notação utilizada nos modelos de categoria. Dentre os 5 modelos propostos, o escolhido para compor a avaliação foi o Operacional, por representar os metadados que permitem acompanhar a rastreabilidade e linhagem dos dados. Conforme dito anteriormente na Seção 5.1, estes metadados são essenciais para auxiliar no atendimento da maioria dos direitos dos Titulares de Dados definidos pela LGPD. Para isto, os participantes responderam as afirmações das dimensões Facilidade de Uso e Interpretabilidade.

Quadro 2 – Média detalhada dos resultados da avaliação do questionário - avaliação da qualidade da informação

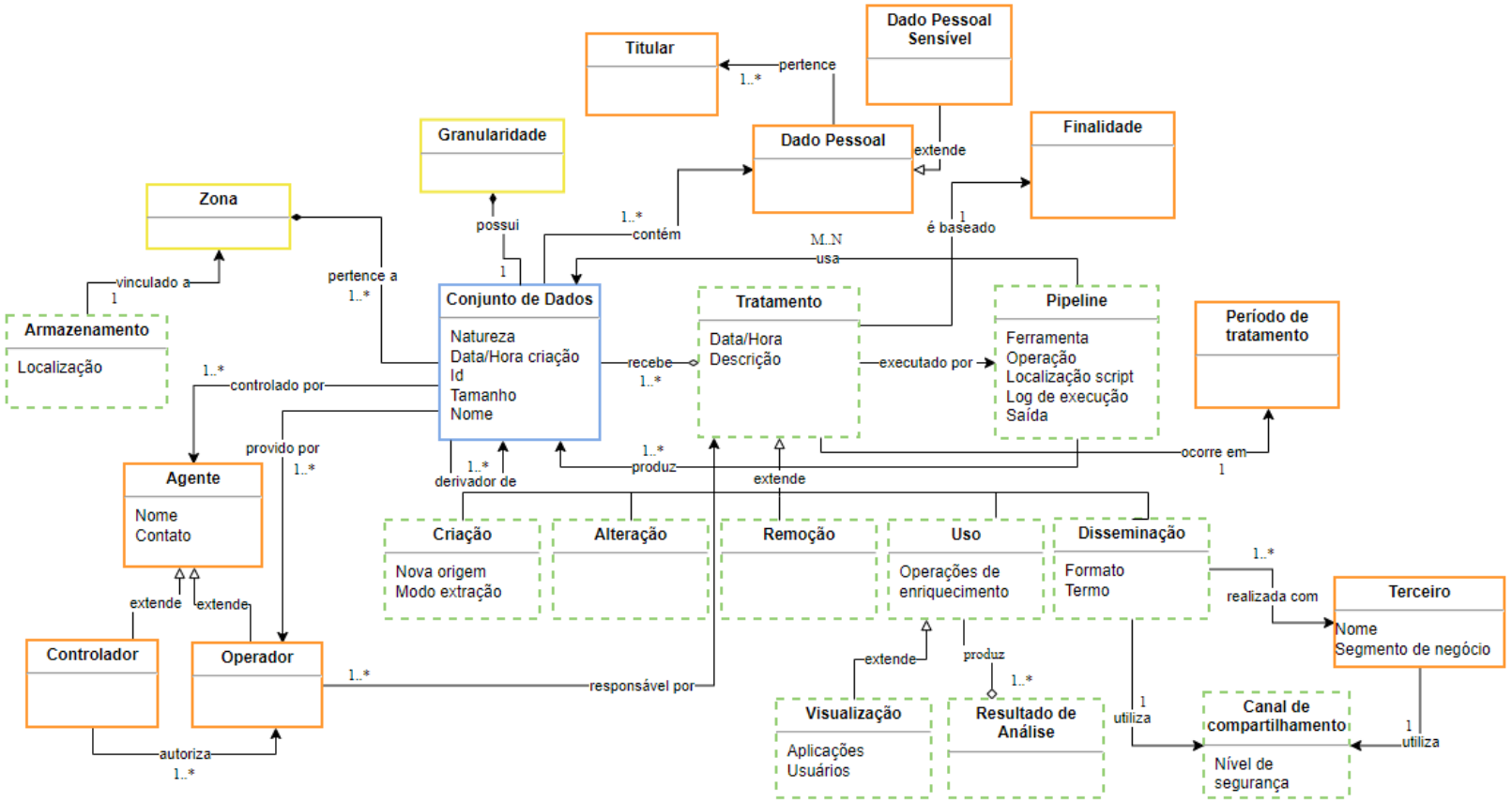
Dimensão	Declarações	Média Avaliação
Completude	O modelo é suficientemente completo para suas necessidades	9,54
	O modelo cobre as necessidades de suas tarefas	9,62
Objetividade	O modelo é baseado em fatos	9,54
	O modelo é objetivo	9,31
	O modelo apresenta uma visão imparcial	8,85
Relevância	O modelo é útil para seu trabalho	9,54
	O modelo não é relevante para seu trabalho	7,46
	O modelo é apropriado para seu trabalho	8,85
	O modelo é aplicável a seu trabalho	8,77
Compreensão	O modelo é fácil de entender	7,31
	Os significados dos conceitos do modelo são difíceis de entender	6,15
	O modelo é de fácil apreensão	7,69
	Os significados dos conceitos do modelo são fáceis de compreender	8,31

Fonte: a autora, 2022

Conforme apresentado no Capítulo 4, os modelos de categoria fazem uso da notação do metamodelo HANDLE (EICHLER et al., 2021) apresentado na Seção 3.3.3. O modelo da categoria Operacional representa metadados que podem ser classificados em diferentes contextos. Por exemplo, ao coletar metadados sobre a finalidade na qual o tratamento se baseia, o modelo agrega informações de caráter jurídico, que podem ser utilizadas para verificar se o tratamento está de acordo com o consentimento fornecido pelo Titular. Sendo assim, propusemos uma alteração estética na notação com o intuito de facilitar para o usuário esta classificação, através de uma identificação visual mais sugestiva.

Na Figura 31 os retângulos sólidos laranja classificam os metadados que fazem parte de um contexto jurídico, enquanto que os objetos retangulares pontilhados em verde, representam metadados de operações. Mantendo a classificação original, em amarelo estão os objetos que identificam a zona do Data Lake e a granularidade do conjunto de dados, que por sua vez, é representado pela cor azul.

Figura 31 – DLCM - Modelo de Metadados Operacional - Versão 2, com proposta de alteração.



Fonte: a autora, 2022

Sendo assim, como forma de testar o benefício dessa alteração, solicitamos aos participantes que avaliassem a interpretabilidade e facilidade de uso do modelo nas duas formas, a primeira utilizando a notação original do HANDLE, apresentada no Capitulo 4 na Seção 4.2.1, e a segunda com o novo esquema de cores, apresentada na Figura 31. O resultado da avaliação dos dois modelos é apresentado detalhadamente na Tabela 3.

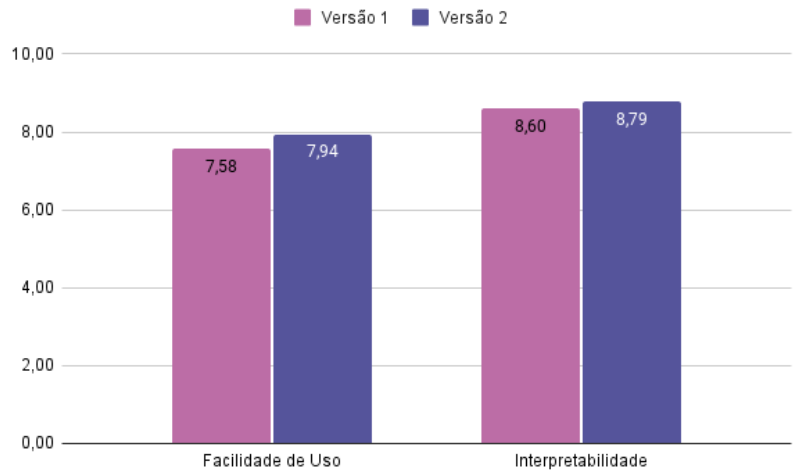
Quadro 3 – Média detalhada dos resultados da avaliação dos Modelos de Categoria Operacional

Dimensão	Declarações	Média Avaliação (Versão 1)	Média Avaliação (Versão 2)
Facilidade de Uso	O modelo é fácil de manipular para atender às nossas necessidades	8,85	8,77
	O modelo é difícil de manipular para atender às nossas necessidades	7,38	7,69
	O modelo é difícil de agregar	6,69	6,92
	O modelo é fácil de combinar com outras informações e modelos	7,78	8,38
Interpretabilidade	É fácil interpretar o que os conceitos do modelo significam	9	9,15
	O modelo é difícil de interpretar	7,54	7,85
	O modelo é fácil de interpretar	8,77	8,92
	Os conceitos usados no modelo são claros	9,08	9,23

Fonte: a autora, 2022

De acordo com o gráfico da Figura 32, a média entre os dois modelos ficaram semelhantes, comprovando que o novo esquema de cores não causou nenhum impacto na avaliação dos participantes. É importante ressaltar que algumas afirmações foram apresentadas no sentido negativo, como é o caso da sentença "O modelo é difícil de agregar", presente na dimensão facilidade de uso. Neste caso, quanto menor a nota, mais positivo é o resultado. Logo, para não comprometer a média da avaliação, os valores dessas sentenças foram normalizados. A Versão 1 corresponde ao modelo da Seção 4.2.1 e a Versão 2, corresponde ao modelo da Figura 31 que contém nossa proposta de alteração estética, .

Figura 32 – Média de Avaliação dos Modelos de Categoria Operacional.



Fonte: a autora, 2022

Além disso, todos os participantes confirmaram que a notação utilizada nos modelos de categoria é útil para representar os metadados. Apesar de todos eles também terem respondido "**Sim**"para a pergunta "**Se o novo esquema de cores é útil para identificar metadados de**

diferentes contextos", apenas 10 participantes confirmaram que a versão 2 é mais adequada do que a versão 1.

A última etapa da avaliação contou com uma pergunta discursiva aberta, para que os participantes dessem sugestões de melhoria. Obtivemos três respostas:

Participante 1 comentou: *Acredito que fornecer mais exemplos, com mais informações de contexto do exemplo, iram enriquecer bastante a visão do modelo para o usuário comum.*

Participante 2 comentou: *Não notei muita diferença entre as versões 2 e 3.*

Participante 3 comentou: *Não tenho sugestões, somente elogios. Considero esse trabalho bastante relevante pois está trazendo a LGPD, lei que entrou em vigência no Brasil recentemente, aplicada ao contexto de Data Lakes. Até então no país não havia regulamentação que para os dados pessoais coletados, e esse trabalho traz essa visão de forma abrangente com um modelo que pode ser aplicado a diferentes contexto, e principalmente o operacional. Parabéns pelo trabalho!*

5.3 CONSIDERAÇÕES FINAIS

Esta avaliação tinha como objetivo analisar o DLCM a nível de completude, objetividade, relevância, compreensibilidade, interpretabilidade e facilidade de uso. No geral, os participantes se mostraram satisfeitos com o modelo proposto e relataram a importância dele para o contexto de Data Lakes. Optamos por não incluir as alterações estéticas presentes na versão 2 do modelo de categoria Operacional, justificando o fato de que a média de avaliação ficou muito próxima a do modelo apresentado na Seção 4.2.1.

6 CONCLUSÃO

Neste Capítulo, descrevemos as considerações finais sobre esta Dissertação. Apresentamos as contribuições de pesquisas alcançadas (Seção 6.1) e as diretrizes de trabalhos futuros (Seção 6.3).

6.1 CONSIDERAÇÕES FINAIS

Neste trabalho, propomos um Modelo de Suporte para Conformidade de Data Lake com a LGPD (DLCM). O modelo proposto se mostra relevante para Controladores de dados que tratam dados pessoais e possuem Data Lakes, uma vez que apresenta os metadados que devem ser coletados para atendimento das solicitações legais efetuadas pelos Titulares de dados. Além disso, os metadados descritos também servirão como fonte de conhecimento sobre os dados armazenados no Data Lake, provendo informações que permitem a exploração e análise por usuários técnicos e de negócio.

Inicialmente, foi apresentada uma visão geral sobre os principais conceitos que fundamentam este trabalho, transcorrendo por temas como Data Lakes, LGPD, Metadados e Rastreabilidade de Dados. Em seguida, introduzimos um aspecto muito importante do gerenciamento de metadados, que são os modelos conceituais. Nesse capítulo, dissertamos sobre como os modelos executam um papel essencial, atuando como uma ponte de comunicação entre especialistas de áreas distintas. Além disso, também apresentamos alguns modelos que representam o regulamento GDPR e modelos de metadados para Data Lakes. O resultado da análise desse capítulo apontou a ausência de soluções para coleta de metadados que descrevam os conjuntos de dados e os tratamentos aplicados sobre eles no Data Lake, levando em consideração, a lei brasileira de proteção de dados LGPD.

Desse modo, propomos o DLCM, que tem como objetivo descrever os metadados que devem ser coletados acerca dos conjuntos de dados armazenados no Data Lake e dos tratamentos aplicados sobre eles. Somado a isso, ele visa apoiar no gerenciamento do Data Lake, fornecendo metadados que registram toda linhagem dos conjuntos de dados.

O modelo proposto é dividido em duas partes, a primeira é composta pelo agrupamento dos metadados em cinco categorias: Operacional, Técnico, Jurídico, Segurança e Negócio, na qual cada uma delas possui um modelo associado que melhor detalha seus elementos, propriedades e relacionamentos. A segunda parte reúne todos os elementos de metadados necessários para que o Controlador dos dados possa atender uma solicitação de acesso aos dados.

Para avaliarmos o DLCM, realizamos um questionário que contou com 13 participantes com experiência profissional e acadêmica no desenvolvimento de atividades que envolvem Data Lakes. Este questionário foi desenvolvido com base nas dimensões de qualidade da informação do AIMQ, e enviado para os participantes por meio de um formulário eletrônico. A primeira

etapa do questionário consistia em caracterizar os participantes, com perguntas a respeito de suas formações e atuações profissionais ou acadêmicas. A segunda etapa buscou avaliar o modelo conceitual em termos de completude, objetividade, relevância e compreensão. Já na terceira e última etapa, foi analisado se a notação utilizada nos modelos de categoria facilita a interpretação dos metadados.

Por meio da avaliação, coletamos evidências sobre a importância do Modelo de Suporte para Conformidade de Data Lake com a LGPD, uma vez que foi possível obter bons resultados.

6.2 LIMITAÇÕES

Durante o estudo algumas limitações puderam ser observadas. Primeiramente, não foi possível utilizar no trabalho a metodologia de grupo focal para avaliação do modelo, devido as restrições de distanciamento social impostas pela pandemia. Todas as discussões para desenvolvimento do modelo ocorreram em maior parte nas reuniões de orientação, sem a participação de especialistas, o que pode ocasionar em um certo viés na formação de opiniões.

Em relação ao método de pesquisa, as limitações são típicas de estudos empíricos, particularmente na generalização dos resultados. A extração e análise de dados pode ser influenciada pelas opiniões pessoais do pesquisador que executa o processo. Considerando que a avaliação ocorreu através de um formulário eletrônico, houve pouca interação com os participantes a respeito das perguntas, e também não é possível afirmar se houve outros fatores, como ambiente e tempo, por exemplo, que influenciaram na redução do contato.

Com relação ao modelo, mais refinamentos podem ser necessários, para minimizar a complexidade de entendimento e aplicação, bem como para adicionar novos elementos de metadados que eventualmente passaram despercebidos pelos participantes e pesquisadores.

6.3 TRABALHOS FUTUROS

Como trabalhos futuros, identificamos as seguintes questões:

- **Aplicar o modelo em um caso de uso real**: Para ter uma melhor visão da importância do modelo proposto é interessante realizar a sua aplicação em um cenário real. Visto que, em um caso de uso, outros elementos de metadados que não foram pensados durante a sua construção podem aparecer.
- **Desenvolver um processo para aplicação do modelo**: Além de apresentar quais metadados devem ser coletados, outra contribuição interessante seria apresentar como coletar tais metadados. Isto é, desenhar um processo que mostre onde o modelo pode ser utilizado para a coleta bem sucedida dos metadados.
- **Desenvolver uma solução de catalogação dos metadados coletados**: Outro trabalho importante seria disponibilizar estes metadados em forma de catálogo, para que

usuários analistas possam produzir novos dados, gerando mais conhecimento para a organização.

- **Avaliar os outros modelos de categoria**: Visando diminuir o tempo de preenchimento da avaliação para os participantes, optamos por avaliar apenas o modelo da categoria Operacional. No entanto, é importante avaliar os demais modelos de categoria (Técnico, Jurídico, Segurança e Negócios), para validar e completar os elementos de metadados propostos.

REFERÊNCIAS

AGOSTINELLI, S.; MAGGI, F. M.; MARRELLA, A.; SAPIO, F. Achieving gdpr compliance of bpmn process models. In: CAPPIELLO, C.; RUIZ, M. (Ed.). *Information Systems Engineering in Responsible Information Systems*. Cham: Springer International Publishing, 2019. p. 10–22. ISBN 978-3-030-21297-1.

AZEVEDO, B.; JINO, M. Modeling traceability in software development: A metamodel and a reference model for traceability. In: *Proceedings of the 14th International Conference on Evaluation of Novel Approaches to Software Engineering*. Setubal, PRT: SCITEPRESS - Science and Technology Publications, Lda, 2019. (ENASE 2019), p. 322–329. ISBN 9789897583759. Disponível em: <https://doi.org/10.5220/0007715103220329>.

BARCLAY, I.; PREECE, A.; TAYLOR, I.; VERMA, D. Towards traceability in data ecosystems using a bill of materials model. *arXiv preprint arXiv:1904.04253*, 2019.

BUFALIERI, L.; MORGIA, M. L.; MEI, A.; STEFA, J. Gdpr: when the right to access personal data becomes a threat. In: IEEE. *2020 IEEE International Conference on Web Services (ICWS)*. [S.l.], 2020. p. 75–83.

CHIHOUB, H.; MADERA, C.; QUIX, C.; HAI, R. Architecture of data lakes. In: _____. *Data Lakes*. John Wiley Sons, Ltd, 2020. cap. 2, p. 21–39. ISBN 9781119720430. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119720430.ch2>.

CONSORTIUM, W. W. W. et al. Data catalog vocabulary (dcat). World Wide Web Consortium, 2014.

COSTA, C.; ANTONUCCI, F.; PALLOTTINO, F.; AGUZZI, J.; SARRIÁ, D.; MENESATTI, P. A review on agri-food supply chain traceability by means of rfid technology. *Food and bioprocess technology*, Springer, v. 6, n. 2, p. 353–366, 2013.

COTTRELL, N. Compliance and gdpr. In: *MongoDB Topology Design*. [S.l.]: Springer, 2020. p. 75–98.

CRESWELL, J. W. Projeto de pesquisa métodos qualitativo, quantitativo e misto. In: *Projeto de pesquisa métodos qualitativo, quantitativo e misto*. Porto Alegre: Artmed, 2010.

DEZANI-CIANCAGLINI, M.; HORNE, R.; SASSONE, V. Tracing where and who provenance in linked data: A calculus. *Theoretical Computer Science*, Elsevier, v. 464, p. 113–129, 2012.

DIAMANTINI, C.; GIUDICE, P. L.; MUSARELLA, L.; POTENA, D.; STORTI, E.; URSINO, D. A new metadata model to uniformly handle heterogeneous data lake sources. In: SPRINGER. *European Conference on Advances in Databases and Information Systems*. [S.l.], 2018. p. 165–177.

EASTERBROOK, S.; SINGER, J.; STOREY, M.-A.; DAMIAN, D. Selecting empirical methods for software engineering research. Springer, p. 285–311, 2008.

EICHLER, R.; GIEBLER, C.; GRÖGER, C.; SCHWARZ, H.; MITSCHANG, B. Modeling metadata in data lakes—a generic model. *Data & Knowledge Engineering*, Elsevier, v. 136, p. 101931, 2021.

FANG, H. Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In: *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*. [S.l.: s.n.], 2015. p. 820–824.

FLECKENSTEIN, M.; FELLOWS, L. Metadata. In: *Modern Data Strategy*. [S.l.]: Springer, 2018. p. 179–193.

FOUNDATION, T. A. S. *Apache Hadoop*. 2020. <https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html#Introduction>. Acesso em 31 de dezembro de 2021.

GAO, Y.; CHEN, X.; DU, X. A big data provenance model for data security supervision based on prov-dm model. *IEEE Access*, IEEE, v. 8, p. 38742–38752, 2020.

GIEBLER, C.; GRöGER, C.; HOOS, E.; EICHLER, R.; SCHWARZ, H.; MITSCHANG, B. The data lake architecture framework: A foundation for building a comprehensive data lake architecture. In: . [S.l.: s.n.], 2021.

GUHA, R. V.; BRICKLEY, D.; MACBETH, S. Schema. org: evolution of structured data on the web. *Communications of the ACM*, ACM New York, NY, USA, v. 59, n. 2, p. 44–51, 2016.

HEVNER, A. R.; MARCH, S. T.; PARK, J.; RAM, S. Design science in information systems research. *Management Information Systems Quarterly*, v. 28, n. 1, p. 6, 2008.

HUANG, J.; LI, S.; THÜRER, M. On the use of blockchain in industrial product service systems: A critical review and analysis. *Procedia CIRP*, Elsevier, v. 83, p. 552–556, 2019.

HUKKERI, T. S.; KANORIA, V.; SHETTY, J. *A Study of Enterprise Data Lake Solutions*. [S.l.]: IRJET, 2020.

KAKU, E.; LOMOTEY, R. K.; DETERS, R. Using provenance and coap to track requests/responses in iot. *Procedia Computer Science*, Elsevier, v. 94, p. 144–151, 2016.

KELEPOURIS, T.; PRAMATARI, K.; DOUKIDIS, G. Rfid-enabled traceability in the food supply chain. *Industrial Management & data systems*, Emerald Group Publishing Limited, 2007.

LEE, Y. W.; STRONG, D. M.; KAHN, B. K.; WANG, R. Y. Aimq: a methodology for information quality assessment. *Information & management*, Elsevier, v. 40, n. 2, p. 133–146, 2002.

LEI Nº 13.709, de 14 de Agosto de 2018. 2018. <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em 05 de Janeiro de 2022.

LÓSCIO, B.; BURLE, C.; CALEGARI, N. Data on the web best practices: Challenges and benefits. In: *Open Data Reserach Symposium (ODRS 2016)*. [S.l.: s.n.], 2016.

MEGDICHE, I.; RAVAT, F.; ZHAO, Y. Metadata management on data processing in data lakes. In: SPRINGER. *International Conference on Current Trends in Theory and Practice of Informatics*. [S.l.], 2021. p. 553–562.

MILOSLAVSKAYA, N.; TOLSTOY, A. Big data, fast data and data lake concepts. *Procedia Computer Science*, Elsevier, v. 88, p. 300–305, 2016.

MIRABELLI, G.; PIZZUTI, T.; GONZÁLEZ, F. G.; BOBI, M. Á. S. Food traceability models: an overview of the state of the art. DIME-University of Genoa (Viena, Austria), 2012.

MORTE, A. B.; MEIRA, A.; COSTA, R.; MARIZ, D. Uma análise sobre o uso de dlts no tratamento de dados pessoais: Aderência aos princípios e direitos elencados na lgpd. In: *Anais do III Workshop em Blockchain: Teoria, Tecnologia e Aplicações*. Porto Alegre, RS, Brasil: SBC, 2020. p. 74–87. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/wblockchain/article/view/12435>.

NAMBIAR, V.; PRAVEEN, K.; SUDARSHAN, R. *Benefits of Modernizing On-premises Analytics with an AWS Lake House*. 2021. <https://aws.amazon.com/pt/blogs/architecture/benefits-of-modernizing-on-premise-analytics-with-an-aws-lake-house/>. Acesso em 31 de dezembro de 2021.

NARGESIAN, F.; ZHU, E.; MILLER, R. J.; PU, K. Q.; AROCENA, P. C. Data lake management: challenges and opportunities. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 12, n. 12, p. 1986–1989, 2019.

NUNES, V.; CAPPELLI, C.; RALHA, C. G. Transparency in information systems. *Sociedade Brasileira de Computação*, 2017.

OLIVEIRA, M. I. S.; OLIVEIRA, L. E. R.; BATISTA, M. G. R.; LÓSCIO, B. F. Towards a meta-model for data ecosystems. In: *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. [S.l.: s.n.], 2018. p. 1–10.

PANDIT, H. J.; LEWIS, D. Modelling provenance for gdpr compliance using linked open data vocabularies. In: . [s.n.], 2017. Disponível em: <http://ceur-ws.org/Vol-1951/#paper-06>.

QUIX, C.; HAI, R.; VATOV, I. Metadata extraction and management in data lakes with gemms. *Complex Systems Informatics and Modeling Quarterly*, n. 9, p. 67–83, 2016.

RAVAT, F.; ZHAO, Y. Data lakes: Trends and perspectives. In: HARTMANN, S.; KÜNG, J.; CHAKRAVARTHY, S.; ANDERST-KOTSIS, G.; TJOA, A. M.; KHALIL, I. (Ed.). *Database and Expert Systems Applications*. Cham: Springer International Publishing, 2019. p. 304–313. ISBN 978-3-030-27615-7.

RAVAT, F.; ZHAO, Y. Metadata management for data lakes. In: SPRINGER. *European Conference on Advances in Databases and Information Systems*. [S.l.], 2019. p. 37–44.

SAWADOGO, P.; DARMONT, J. On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, v. 56, n. 1, p. 97–120, Feb 2021. ISSN 1573-7675. Disponível em: <https://doi.org/10.1007/s10844-020-00608-7>.

SAWADOGO, P.; KIBATA, T.; DARMONT, J. Metadata management for textual documents in data lakes. *arXiv preprint arXiv:1905.04037*, 2019.

SCHOLLY, E.; SAWADOGO, P.; LIU, P.; ESPINOSA-OVIEDO, J. A.; FAVRE, C.; LOUDCHER, S.; DARMONT, J.; NOÛS, C. Coining goldmedal: a new contribution to data lake generic metadata modeling. *arXiv preprint arXiv:2103.13155*, 2021.

SURIARACHCHI, I.; PLALE, B. Crossing analytics systems: a case for integrated provenance in data lakes. In: IEEE. *2016 IEEE 12th International Conference on e-Science (e-Science)*. [S.l.], 2016. p. 349–354.

TORRE, D.; ALFEREZ, M.; SOLTANA, G.; SABETZADEH, M.; BRIAND, L. Model driven engineering for data protection and privacy: Application and experience with gdpr. *arXiv preprint arXiv:2007.12046*, 2020.

TORRE, D.; SOLTANA, G.; SABETZADEH, M.; BRIAND, L. C.; AUFFINGER, Y.; GOES, P. Using models to enable compliance checking against the gdpr: An experience report. In: *2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems (MODELS)*. [S.l.: s.n.], 2019. p. 1–11.

TRISOVIC, A.; JONES, C. R.; COUTURIER, B.; CLEMENCIC, M. Provenance tracking in the lhcb software. *Computing in Science & Engineering*, IEEE, v. 22, n. 2, p. 88–94, 2020.

UJCICH, B. E.; BATES, A.; SANDERS, W. H. A provenance model for the european union general data protection regulation. In: BELHAJJAME, K.; GEHANI, A.; ALPER, P. (Ed.). *Provenance and Annotation of Data and Processes*. Cham: Springer International Publishing, 2018. p. 45–57. ISBN 978-3-319-98379-0.

VIOLINO, S.; ANTONUCCI, F.; PALLOTTINO, F.; CECCHINI, C.; FIGORILLI, S.; COSTA, C. Food traceability: A term map analysis basic review. *European Food Research and Technology*, Springer, v. 245, n. 10, p. 2089–2099, 2019.

ZGOLLI, A.; COLLET, C.; MADERA, C. Metadata in data lake ecosystems. *Data Lakes*, Wiley Online Library, v. 2, p. 57–96, 2020.

ZHAO, Y.; MEGDICHE, I.; RAVAT, F. Data lake ingestion management. *ArXiv*, abs/2107.02885, 2021.

ZHAO, Y.; MEGDICHE, I.; RAVAT, F.; DANG, V.-n. A zone-based data lake architecture for iot, small and big data. In: *25th International Database Engineering & Applications Symposium*. [S.l.: s.n.], 2021. p. 94–102.

APÊNDICE A – FORMULÁRIO DE AVALIAÇÃO DO DLCM

Formulário de Avaliação do DLCM

Meu nome é Vitória Maria da Silva Maciel. Sou mestranda em Ciência da Computação no Centro de Informática (CIn) da Universidade Federal de Pernambuco (UFPE). Também sou orientanda da professora Bernadette Farias Lóscio. Gostaria de agradecer por colaborar no meu trabalho participando desta avaliação. Seu feedback é extremamente valioso para a conclusão da minha pesquisa de mestrado.

O questionário a seguir avalia um Modelo de Suporte para Conformidade de Data Lake com a LGPD - DCLM. Em particular, o modelo proposto é composto por um conjunto de elementos que representam metadados, os quais foram pensados com o intuito de descrever os dados e os tratamentos aplicados sobre eles em um Data Lake.

Suas respostas são anônimas, ou seja, você não precisa fornecer suas informações pessoais, além do seu e-mail, se quiser receber os resultados dessa pesquisa.

*Obrigatório

FASE 1 - CARACTERIZAÇÃO DO PARTICIPANTE

As informações abaixo servem meramente para caracterização do perfil dos respondentes. Em nenhum momento os respondentes serão identificados ou relacionados individualmente dentro do trabalho.

1. Qual o seu nível de formação (concluído)? *

Marcar apenas uma oval.

- ◯ Graduação
- ◯ Especialização
- ◯ Mestrado Acadêmico/Profissional
- ◯ Doutorado
- ◯ Pós-Doutorado

2. Qual o seu cargo/função atualmente? *

_____

3. Você desenvolve suas atividades em qual meio? *

Marcar apenas uma oval.

- ◯ Acadêmico
- ◯ Profissional
- ◯ Ambos

4. Como você classificaria o seu grau de conhecimento sobre Data Lakes de uma forma geral? *

Marcar apenas uma oval.

	0	1	2	3	4	5	6	7	8	9	10	
Nenhum	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Avançado

5. Como você classificaria o seu grau de conhecimento sobre a LGPD? *

Marcar apenas uma oval.

	0	1	2	3	4	5	6	7	8	9	10	
Nenhum	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Avançado

6. Você tem participado direta ou indiretamente de atividades que envolvem Data Lakes? *

Marcar apenas uma oval.

- ◯ Sim
- ◯ Não

7. Se sim, como você caracteriza sua participação?

Marque todas que se aplicam.

- ☐ Atuo como consumidor dos dados do Data Lake
- ☐ Atuo como operador responsável por inserir os dados no Data Lake
- ☐ Atuo como operador responsável por tratar os dados no Data Lake
- ☐ Atuo na gestão do Data Lake
- ☐ Atuo no desenvolvimento de infraestrutura do Data Lake
- ☐ Atuo na gestão da conformidade do Data Lake

8. Se sim, por quanto tempo atuou ou tem atuado em atividades com Data Lakes?

Marcar apenas uma oval.

- ◯ Menos de 2 anos
- ◯ De 2 a menos de 3 anos
- ◯ De 3 a menos de 4 anos
- ◯ De 4 a menos de 4 anos
- ◯ Mais de 5 anos

FASE 2 -Avaliação da Qualidade da Informação reunida pelo Modelo

Visão Geral

A Lei Geral de Proteção de Dados Pessoais (LGPD) surgiu como um mecanismo para regular o tratamento de dados pessoais em território nacional. Ela define tratamento de dados pessoais como qualquer operação realizada com dados pessoais, tais como: coleta, produção, transmissão, dentre outros. Além disso, são caracterizados três atores: (i) o Titular dos Dados, pessoa natural a quem se referem os dados pessoais que são objetos de tratamento; (ii) o Controlador dos Dados, agente responsável por tomar as decisões acerca dos tratamentos; e o (iii) Operador dos Dados, que é a figura encarregada de realizar os tratamentos em nome do Controlador.

Além de regulamentar as diretrizes sobre como as organizações devem lidar com dados pessoais, a LGPD também assegura o direito dos titulares dos dados. O direito de acesso aos dados é um dos tipos de solicitação que o controlador pode receber do titular. Para atender a esta demanda, ele deverá informar quais dados possui relacionados ao titular e quais tratamentos foram aplicados sobre eles em um prazo de 15 dias.

Em Data Lakes (DL), a ingestão dos dados ocorre inicialmente sem a preocupação de rotulagem ou auditorias. Não há informações sobre os fins para quais os dados foram extraídos, e além disso, estes dados podem passar por diversos workflows em ferramentas distintas e tratamentos para diferentes finalidades. Para alcançar a conformidade com a LGPD, é necessário que os Data Lakes possuam informações de toda linhagem dos dados armazenados e processados, comprovando que é possível identificar onde eles podem ser encontrados e como eles fluem dentro do ambiente.

Neste contexto, o Modelo de Suporte para Conformidade de Data Lake com a LGPD (DLCM) foi proposto com o objetivo de descrever quais metadados devem ser capturados acerca dos dados e dos tratamentos realizados sobre eles em um DL. Estes metadados devem auxiliar o Controlador no retorno à solicitação de acesso aos dados efetuada por um Titular.

Considerando a extensão e complexidade do modelo, decidimos agrupar os metadados em cinco categorias, são elas: Técnico, Operacional, Jurídico, Negócio e Segurança. Cada categoria possui um modelo associado, que melhor detalha os relacionamentos entre os metadados e seus atributos. O objetivo desse agrupamento consiste em facilitar o entendimento do modelo e identificar processos nos quais ele pode ser utilizado.

Para esta avaliação, utilizaremos como base apenas a categoria de metadados Operacional.

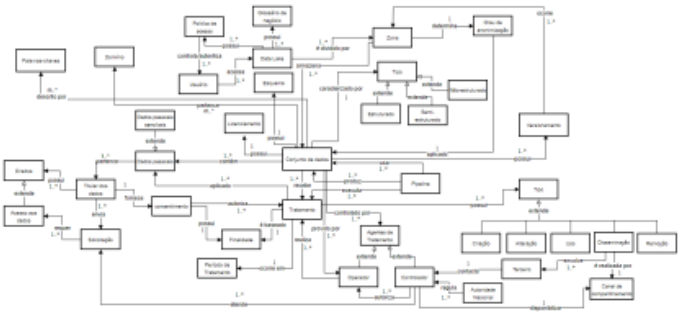
Modelo de Suporte para Conformidade de Data Lake com a LGPD (DCLM)

O DLCM tem como objetivo representar todos os metadados que devem ser coletados em um Data Lake, para dar suporte ao Controlador dos dados no atendimento de uma solicitação de acesso aos dados, efetuada pelo Titular.

Uma descrição completa de todos os elementos pode ser encontrada aqui:
https://drive.google.com/file/d/1npq0Jrd_CdKYU0HChtUmU-2Xwid_Hc4Z/view?usp=sharing

Figura de alta resolução: https://drive.google.com/file/d/18Ef2thCMYvsH-zbQv_pYiEU2wOWGmEiJ/view?usp=sharing

Figura 1: Modelo de Suporte para Conformidade de Data Lake com a LGPD.



9. O modelo apresentado dispõe dos metadados necessários que devem ser coletados no Data Lake para atender ao objetivo proposto? *

Marque todas que se aplicam.

- [ ] Não, um ou mais metadados precisam ser excluídos
- [ ] Não, um ou mais metadados precisam ser atualizados
- [ ] Não, um ou mais metadados precisam ser adicionados
- [ ] Sim, os metadados são suficientes

10. Quais são as suas sugestões em caso de necessidade de inclusão , exclusão ou atualização dos metadados ?

_____

_____

_____

_____

_____

Qualidade dos Metadados

Quanto as informações reunidas através do modelo de suporte para conformidade de Data Lake com a LGPD, leia as afirmações abaixo e responda de acordo com metadados representados no modelo proposto e a descrição da visão geral. Na escala à direita: o número 1 corresponde a "não concordo com a afirmação" e o número 10 corresponde a "concordo com a afirmação".

1) Completude

O modelo é composto de todos os metadados necessários para a descrição dos dados e os tratamentos aplicados sobre eles no Data Lake.

11. b) O modelo é suficientemente completo para suas necessidades. *

Marcar apenas uma oval.

	0	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

12. e) O modelo cobre as necessidades de suas tarefas. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

2) Objetividade

O modelo é claro e atende aos objetivos do trabalho.

13. a) O modelo é baseado em fatos. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

14. b) O modelo é objetivo. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

15. c) O modelo apresenta uma visão imparcial. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

3) Relevância

Modelo é aplicável, útil no trabalho, indispensável e importante para a execução do trabalho.

16. a) O modelo é útil para seu trabalho. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

17. b) O modelo não é relevante para seu trabalho. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

18. c) O modelo é apropriado para seu trabalho. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

19. d) O modelo é aplicável a seu trabalho. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

4) Compreensão

Modelo é fácil de ser compreendido, é de fácil entendimento.

20. a) O modelo é fácil de entender. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

21. b) Os significados dos conceitos do modelo são difíceis de entender. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

22. c) O modelo é de fácil apreensão. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

23. d) Os significados dos conceitos do modelo são fáceis de compreender. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

FASE 3 - Avaliação de Usabilidade do Modelo Operacional

Quanto as informações reunidas através do modelo que representa a categoria de metadados Operacional, leia as afirmações abaixo e responda de acordo com a descrição fornecida.
Na escala à direita: o número 1 corresponde a "não concordo com a afirmação" e o número 10 corresponde a "concordo com a afirmação".

1.2 DLCM - Operacional

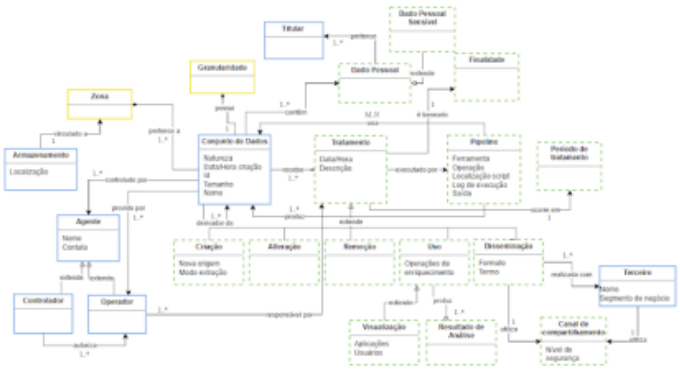
Visando facilitar o entendimento do modelo anterior, agrupamos os metadados em categorias. Cada categoria possui um modelo associado que melhor detalha as propriedades dos metadados e seus relacionamentos.

A categoria de metadados Operacional descreve as operações de tratamento realizadas em um conjunto de dados no Data Lake. Este modelo pode ser utilizado para identificar e gerenciar as modificações aplicadas nos dados durante seu ciclo de vida.

Uma descrição detalhada pode ser encontrada aqui:
https://drive.google.com/file/d/1npq0Jrd_CdKYU0HChtUmU-2Xwid_Hc4Z/view?usp=sharing

Figura em alta resolução: https://drive.google.com/file/d/1-S0hyOnVQAuEI2m1wAE1_dX9w0aJPMKn/view?usp=sharing

Figura 2: DLCM - Modelo Operacional



1) Facilidade de Uso

O modelo permite o uso sem causar dificuldades durante o seu manuseio.

24. a) O modelo é fácil de manipular para atender às nossas necessidades. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

25. b) O modelo é difícil de manipular para atender às nossas necessidades. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

26. c) O modelo é difícil de agregar. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

27. d) O modelo é fácil de combinar com outras informações e modelos. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

2) Interpretabilidade

O modelo possui visualização gráfica e semântica claras. É apresentada em linguagens, símbolos e unidades apropriados.

28. a) É fácil interpretar o que os conceitos do modelo significam. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

29. b) O modelo é difícil de interpretar. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

30. c) O modelo é fácil de interpretar. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

31. d) Os conceitos usados no modelo são claros. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

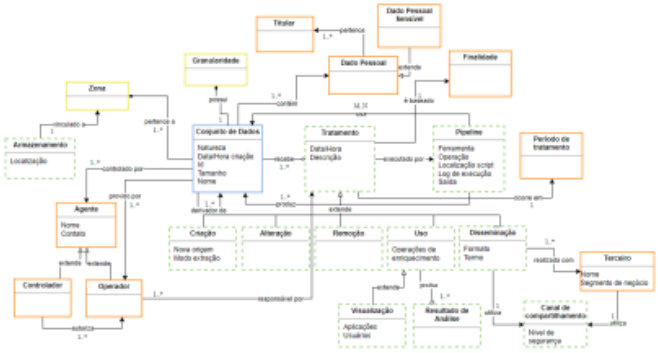
1.2 DLCM - Operacional - Versão 2

O modelo da categoria Operacional representa metadados que podem ser classificados em diferentes contextos. Por exemplo, ao coletar metadados sobre a finalidade na qual o tratamento se baseia, o modelo agrega informações de caráter jurídico. Sendo assim, introduzimos uma nova classificação nos metadados dentro da categoria Operacional, através de um novo esquema de cores.

Uma descrição detalhada pode ser encontrada aqui:
https://drive.google.com/file/d/1npg0Jrd_CdKYU0HChtUmU-2Xwid_Hc4Z/view?usp=sharing

Figura em alta resolução: https://drive.google.com/file/d/1YGwV0rR4iBzH0hxrRw2dUbGOg_ArCQa4/view?usp=sharing

Figura 3: DLCM - Modelo Operacional - Versão 2



1) Facilidade de Uso

O modelo permite o uso sem causar dificuldades durante o seu manuseio.

32. a) O modelo é fácil de manipular para atender às nossas necessidades. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

33. b) O modelo é difícil de manipular para atender às nossas necessidades. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

34. c) O modelo é difícil de agregar. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

35. d) O modelo é fácil de combinar com outras informações e modelos. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

2) Interpretabilidade

O modelo possui visualização gráfica e semântica claras. É apresentada em linguagens, símbolos e unidades apropriados.

36. a) É fácil interpretar o que os conceitos do modelo significam. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

37. b) O modelo é difícil de interpretar. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

38. c) O modelo é fácil de interpretar. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

39. d) Os conceitos usados no modelo são claros. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Não concordo com a afirmação	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo com a afirmação

FASE 3 - PERGUNTAS DISCURSIVAS

40. A notação/linguagem utilizada no modelo que representa a categoria de metadados Operacional é útil? *

Marcar apenas uma oval.

◯ Sim
◯ Não

41. Em caso de Não, qual seria sua sugestão de notação/linguagem para representar os metadados?

_____

42. A classificação introduzida na versão 2 (Figura 3) do modelo Operacional é útil para identificar metadados de diferentes contextos? *

Marcar apenas uma oval.

◯ Sim
◯ Não

43. Para você, a versão 2 do modelo Operacional (Figura 3) é mais adequada do que a anterior (Figura 2)? *

Marcar apenas uma oval.

◯ Sim
◯ Não

44. Demais sugestões para os modelos? Favor descrever aqui:

_____
_____
_____
_____
_____