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

ORGANIZAÇÃO DA INFORMAÇÃO DO SISTEMA ELETRÔNICO DE INFORMAÇÕES (SEI) PARA VIABILIZAR CONSULTAS ANALÍTICAS E AUDITORIA

Samuel Victor Cavalcante da Ponte

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Marcio de Carvalho Victorino

Brasília
2024

Ponte, Samuel Victor Cavalcante
Po ORGANIZAÇÃO DA INFORMAÇÃO DO SISTEMA ELETRÔNICO DE INFORMAÇÕES (SEI) PARA VIABILIZAR CONSULTAS ANALÍTICAS E AUDITORIA / Samuel Victor Cavalcante Ponte; orientador Marcio de Carvalho Victorino. -- Brasília, 2024.
100 p.

Dissertação(Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2024.

1. SEI - Sistemas Eletrônico de Informação. 2. Manipulação de LOG. 3. ETL. 4. Datawarehose. 5. Tumming de banco. I. Victorino, Marcio de Carvalho, orient. II. Título.

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

ORGANIZAÇÃO DA INFORMAÇÃO DO SISTEMA ELETRÔNICO DE INFORMAÇÕES (SEI) PARA VIABILIZAR CONSULTAS ANALÍTICAS E AUDITORIA

Samuel Victor Cavalcante da Ponte

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Marcio de Carvalho Victorino (Orientador)
FCI/UnB

Profa. Dra. Ana Paula Bernardi da Silva
UCB

Prof. Dr. Andrei Lima Queiroz
UnB

Prof. Dr. Edison Ishikawa
UnB

Prof. Dr. Gladston Luiz da Silva
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 24 de outubro de 2024

Dedicatória

À minha esposa e filhos

Agradecimentos

Agradecimentos à Secretaria de Tecnologia da Informação (STI), pela parceria, que disponibilizou o ambiente computacional e acesso aos dados necessários para os pesados processamentos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

Em um mundo de documentos eletrônicos o Sistemas Eletrônico de Informação (SEI) destaca-se no Brasil. Adotado por um grande número de entidades do estado, tornou-se estratégico para a Universidade de Brasília (UnB) desde a entrada em produção, em 2016. Em seu processo arquivístico, faz o registro de todas operações realizadas por usuários em documentos e também no próprio sistema. É o chamado *log* do sistema. No entanto, dado ao grande número de usuários e interações, gerando um elevando volume de registros, fez com que o tempo de resposta de consultas a ele fosse degradando ao longo do tempo, tornado-o intratável hoje na maioria dos casos. O objetivo deste trabalho é apresentar uma estrutura adjacente de dados, a qual recebe dados compilados do repositório de *log*. Esta estrutura é carregada com um subconjunto de informações estrategicamente escolhidos e tradados, os quais respondem, inclusive através de *dashboards*, a grande maioria das demandas dos usuários.

Palavras-chave: SEI, log de utilização, tunning

Abstract

Abstract

In a world of electronic documents, the Sistemas Eletrônico de Informação (SEI) stands out in Brazil. Adopted by a large number of state entities, it became strategic for the Universidade de Brasília (UnB) since its deployment in 2016. In its archival process, it records all operations performed by users in documents and also within the system itself. This is referred to as the system's *log*. However, due to the high number of users and interactions, resulting in a high volume of records, the query response time has degraded over time, rendering it unmanageable in most cases today. The aim of this work is to present an adjacent data structure that receives compiled data from the *log* repository. This structure is loaded with a subset of strategically chosen and processed information, which addresses the vast majority of user demands, including through dashboards.

Keywords: SEI, utilization log, tunning

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Códigos

Lista de Abreviaturas e Siglas

ACE Arquivo Central.

BI Business Intelligence.

CRUD criação, leitura, edição e exclusão.

CSV Comma Separated Values.

DBMS Sistema de Gerenciamento de Banco de Dados não Relacional.

DM Data Mart.

DPO Data Protection Officer.

DW Data Warehouse.

EDW Enterprise Data Warehouse.

ETL Extraction, Transformation and Loading.

GDPR General Data Protection Regulation.

HITECH Health Information Technology for Economic and Clinical Health.

HOLAP Hybrid Online Analytical Processing.

HTTP Hypertext Transfer Protocol.

JSON JavaScript Object Notation.

LAI Lei de Acesso à Informação.

LGPD Lei Geral de Proteção de Dados.

MOLAP Multidimentional Online Analytical Processing.

NoLAP Not Online Analytical Processing.

NoSQL Not SQL.

OLAP Online Analytical Processing.

OLTP Online Transaction Processing.

PHP Hypertext Preprocessor.

ROLAP Relacional Online Analytical Processing.

SAD Sistemas de Apoio à Decisão.

SAE Sistemas de Apoio Executivo.

SEI Sistemas Eletrônico de Informação.

SGBD Sistema Gerenciador de Banco de Dados.

SGBDR Sistema Gerenciador de Banco de Dados Relacional.

SIG Sistemas de Informações Gerenciais.

SPT Sistemas de Processamento de Transações.

SQL Strutured Query Language.

SQL Server Banco de Dados Relacional SQL Server.

SSD Sistemas de Suporte à Secisão.

STC Sistemas de Trabalhadores do Conhecimento.

STI Secretaria de Tecnologia da Informação.

TEMAC Teoria do Enfoque Meta Analítico Consolidado.

TJDFT Tribunal de Justiça do Distrito Federal e Territórios.

TRF4 Tribunal Regional Federal - 4ª Região.

TRF5 Tribunal Regional Federal Quinta Região.

UnB Universidade de Brasília.

VOSviewer Visualização de landscapes.

WoC Web Of Science.

Capítulo 1

Introdução

Este capítulo introduz o objeto de estudo deste trabalho, apresentando o SEI e os principais conceitos associados. A seção 1.2 descreve o ambiente em que o sistema está inserido, enquanto a seção 1.3 explora sua evolução e crescimento ao longo do tempo. Por fim, a seção 1.4 detalha a estrutura de armazenamento de dados utilizada.

1.1 Introdução

A sigla "SEI" designa o Sistema Eletrônico de Informação, uma plataforma voltada para a gestão eficiente de documentos e processos, concebida com o intuito primordial de otimizar os procedimentos de tramitação e, consequentemente, ao mesmo tempo, reduzir a dependência do meio físico, particularmente o papel.

Essa diminuição de produção de papel fez com que até mesmo lixeiras fossem retiradas das salas da Universidade de Brasília (UnB) no ano de 2023. Tal medida não se limita apenas à adoção de um regime de coleta mais criterioso, mas também encontra justificativa na crescente superfluidade das lixeiras, devido à diminuição substancial da produção de esboços e rascunhos ao longo dos últimos anos.

O SEI foi concebido, ainda, para simplificar o intricado processo de tramitação de documentos, graças à sua incorporação de processos predefinidos, o que proporciona maior clareza na definição do destino apropriado para diferentes tipos de solicitações, ao estabelecer mapas de processo delineados. Ademais, uma característica meritória é a sua capacidade de transpor barreiras geográficas, beneficiando-se do caráter eletrônico que permite o envio instantâneo de processos, eliminando, assim, a necessidade de logística baseada em papel e, por conseguinte, gerando uma economia expressiva.

Essa característica não apenas facilita o processo de tramitação entre unidades e setores separados, independente da proximidade geográfica, mas também viabiliza a criação, edição, compartilhamento e consulta de documentos de forma simplificada.

Destacadamente reconhecido pela sua inserção cotidiana na dinâmica da Universidade de Brasília, o SEI teve sua origem no âmbito do Tribunal Regional Federal - 4ª Região (TRF4), cuja jurisdição abrange os estados do Paraná, Rio Grande do Sul e Santa Catarina. Em um momento não precisamente identificado, o sistema distribuiu suas virtudes para além dos limites desse tribunal, sendo adotado por diversos órgãos e tribunais, incluindo o Tribunal de Justiça do Distrito Federal e Territórios (TJDFT) e o Tribunal Regional Federal Quinta Região (TRF5). Em Brasília, além do TJDFT, sua utilização também atende ao Governo do Distrito Federal.

Dessa maneira, o SEI se erigiu como um padrão nacional, notadamente impulsionado por sua gratuidade, o que o torna amplamente empregado como um sistema de protocolo tanto no âmbito do governo federal quanto em outras esferas administrativas. O sistema compreende funcionalidades que abarcam recursos de controle arquivístico, englobando a conservação de informações que delineiam o ciclo de vida de documentos e processos, englobando inclusive registros de atividades, tais como a criação, edição e encaminhamento.

Através deste mecanismo de controle, é efetuado um monitoramento minucioso do ciclo de vida da informação, principalmente no contexto do SEI, onde tal acompanhamento incide sobre a trajetória evolutiva dos processos e seus documentos.

Este sistema de controle pode ser segmentado em duas esferas distintas: a primeira, visível para os usuários, compreende funcionalidades de criação, leitura, edição e exclusão (CRUD), bem como a tramitação documental e a consulta textual. A segunda esfera abarca os metadados, que carecem de um interesse explicitamente aparente, tais como os registros de atividades.

Além disso, ao lidar com sistemas de informação que abrangem toda a estrutura da entidade, como é o caso em questão, é fundamental que as diretrizes estabelecidas pela Lei Geral de Proteção de Dados (LGPD) [6] e pela Lei de Acesso à Informação (LAI) [7] sejam rigorosamente observadas. A LGPD define regras claras e responsabilidades específicas para os agentes no que diz respeito à proteção de dados.

Já a LAI, aponta responsabilidades e compromissos quanto a transparência de dados abertos. Portanto, este trabalho deve obrigatoriamente incluir medidas de proteção para garantir que esses dados permaneçam inacessíveis a terceiros não autorizados sem prejuízo de acesso aos dados abertos. Ainda na LAI são apresentados os princípios de tratamento de dados, dentre os quais podemos destacar a qualidade dos dados e prestação de contas. É preciso garantir que os titulares dos dados tenham acesso a informações de maneira precisa e eficiente. A UnB é integrante do poder executivo, sendo pois sujeita a estas diretrizes, e desta forma a LAI, tornou este trabalho importante uma vez que o público tem direito a estes dados

Uma ressalva de extrema relevância é a de que todas as operações efetuadas pelos

usuários são devidamente registradas, abrangendo desde a gênese de documentos, emissão de despachos, encaminhamento de processos, até o arquivamento dos processos.

1.2 SEI na UnB

No mês de julho de 2023, o sistema contabilizava o cadastro de mais de 1600 unidades. Este escopo engloba todas as estruturas da universidade, desde a reitoria até as portarias, abarcando tanto as unidades em vigor quanto aquelas que compuseram o histórico da instituição.

O número de usuários totalizava quase 160 mil. Esse grupo é constituído por indivíduos tanto internos quanto externos à universidade, como estudantes, servidores e outros que tiveram interações de cunho formal com a instituição. Como exemplo, podemos citar um jornalista solicitando dados quantitativos para embasar a elaboração de uma reportagem.

Nesse mesmo período, como ilustrado na Figura 1.1, a base de dados continha mais de 1 mil e seiscentas unidades, mais de 157 mil usuários e, quase 10 milhões de documentos (9.184.938), e aproximadamente 3 terabytes de disco, refletindo a vasta quantidade de informações contidas no sistema.

Figura 1.1: Principais Quantitativos de Objetos do SEI. Julho 2023.

Para dar conta de todo esse volume, o espaço em disco alocado somava quase 3 terabytes. É válido salientar que essa quantia refere-se apenas ao espaço destinado ao Banco de Dados Relacional SQL Server (SQL Server). Além deste, o sistema também abriga um repositório de arquivos binários e dos HTMLs gerados nos documentos. Além disso, há também o espaço utilizado pelo SolaR, um banco de dados textual que faz parte da infraestrutura do sistema.

1.3 Evolução de Crescimento

A Figura 1.2 ilustra a trajetória do crescimento do SEI desde sua implantação em 2017 até julho de 2023, focando particularmente na ampliação do tamanho dos arquivos do banco de dados. O gráfico revela uma progressão constante e estável.

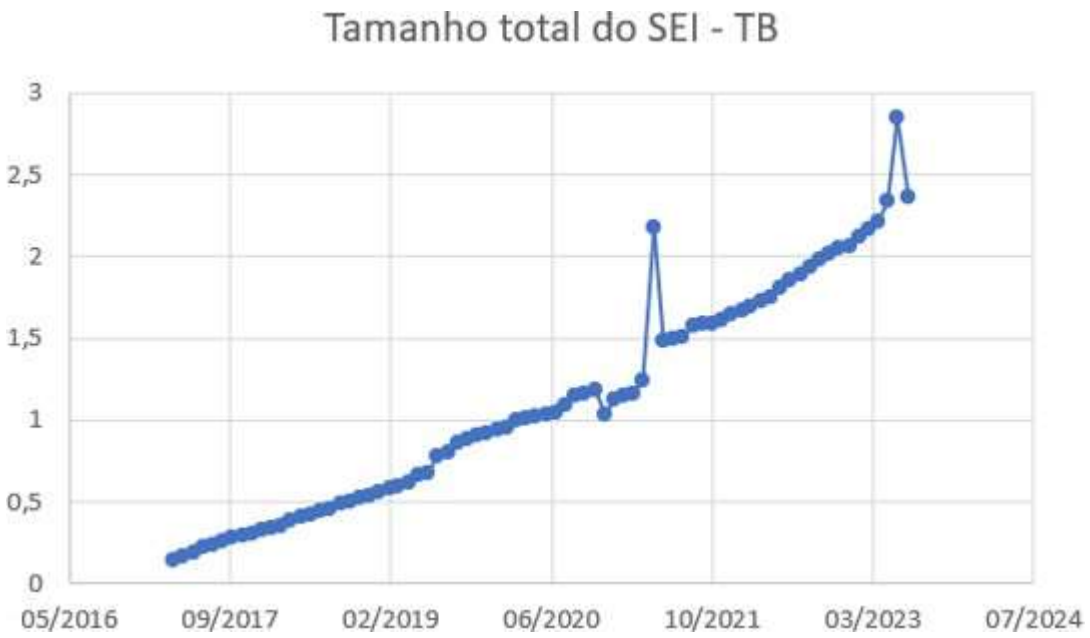

Figura 1.2: Evolução do tamanho total (em terabytes) dos arquivos de dados do SEI.

Dois picos notáveis surgem na evolução dessa alocação, marcando maio de 2021 e junho de 2023. Ambos correspondem a atividades de manutenção do sistema que resultaram em uma alocação extraordinária de espaço. Contudo, logo após essas situações, o espaço alocado foi prontamente otimizado.

O primeiro desses eventos coincide com a ação realizada na época para separar os arquivos de auditoria do sistema, gerando uma temporária alocação adicional de espaço para a importação. Esse espaço foi na sequencia liberado pelo banco de dados relacional SQL Server. O segundo acontecimento, mais recente, envolveu a otimização dos índices por meio de um procedimento de refatoração. Isso consumiu espaço temporário para processamento, o qual também foi posteriormente liberado.

A Figura 1.3 complementa a análise, a variação percentual mês a mês do tamanho do espaço em disco. A linha vermelha nesse gráfico representa a média deste percentual, que se mantém em torno de 3% mensais. Há uma leve tendência de queda nesse percentual, o que se explica matematicamente. Como o valor adicionado permanece fixo enquanto o total continua a aumentar, o percentual naturalmente tende a diminuir. Os dois pi-

cos nesse gráfico estão alinhados com os dois mencionados anteriormente (Figura 1.2), demonstrando sua conexão com as situações de manutenção no gráfico anterior.

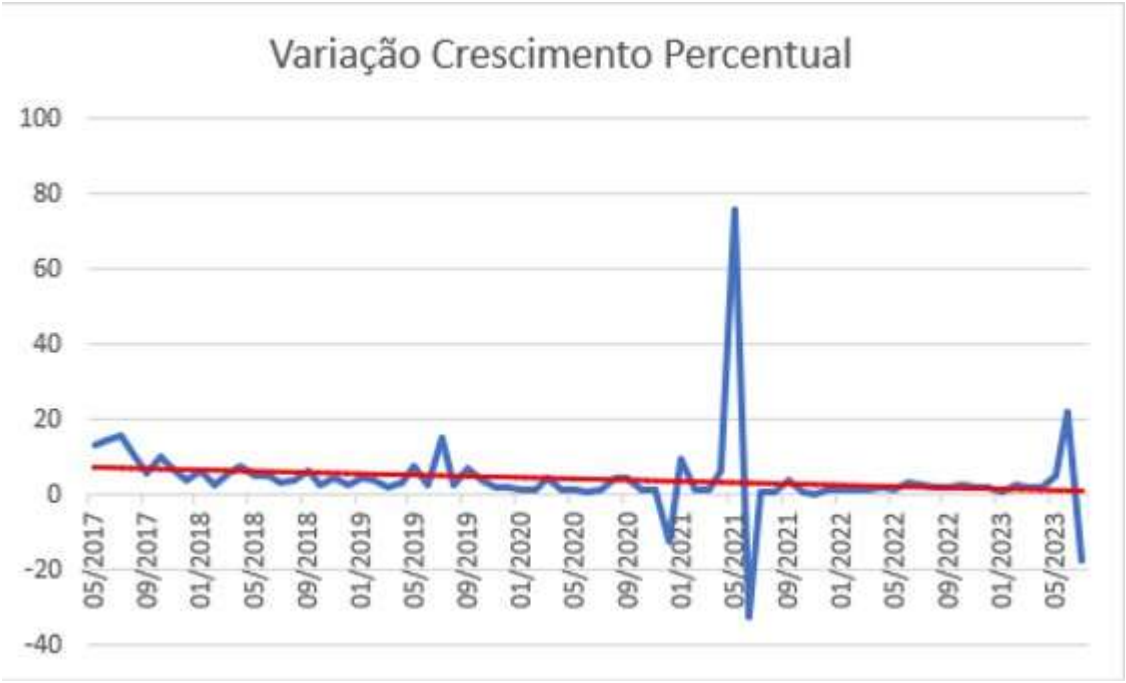

Figura 1.3: Percentual de crescimento do banco de dados mensal (azul) e tendência deste crescimento (vermelho).

1.4 A tabela infra_auditoria

Refinando a abordagem e focando no cerne deste trabalho, concentramo-nos na tabela denominada "infra_auditoria".

Esta tabela, de nome infra_auditoria, ostenta um total excedente de 350 milhões de registros (352.534.728).

A Figura 1.4 exibe um instantâneo de sua estrutura, capturado diretamente da interface SQL Server.

infra_auditoria

	Nome da Coluna	Tipo Condensado
🔑	id_infra_auditoria	bigint
	recurso	varchar(50)
🔑	dth_acesso	datetime
	ip	varchar(39)
	id_usuario	int
	sigla_usuario	varchar(100)
	nome_usuario	varchar(100)
	id_orgao_usuario	int
	sigla_orgao_usu...	varchar(30)
	id_usuario_emul...	int
	sigla_usuario_e...	varchar(100)
	nome_usuario_...	varchar(100)
	id_orgao_usuari...	int
	sigla_orgao_usu...	varchar(30)
	id_unidade	int
	sigla_unidade	varchar(30)
	descricao_unida...	varchar(250)
	id_orgao_unida...	int
	sigla_orgao_uni...	varchar(30)
	servidor	varchar(250)
	user_agent	varchar(MAX)
	requisicao	varchar(MAX)
	operacao	varchar(MAX)

Figura 1.4: Estrutura física da tabela de dados infra_auditoria.

Os atributos dessa tabela podem ser categorizados em duas partes distintas. A Figura 1.5 destaca em vermelho os campos que agregam informações acerca dos eventos, como endereço IP, usuário, unidade, data e outros com algum nível de interligação no banco de dados, notadamente através de relações de chave primária e estrangeira (como os campos usuário e unidade).

Figura 1.5: Classificação dos tipos de campos da tabela infra_auditoria.

É digno de nota o campo user_agente, o qual, apesar de ser do tipo *varchar(max)*, capaz de armazenar vastas quantidades de texto, é empregado aqui exclusivamente para registrar detalhes sobre o navegador utilizado pelo usuário.

Importante frisar que nenhum dos campos mencionados faz menção à natureza da atividade executada.

A segunda parte, destacada em verde, abarca o próprio conteúdo da atividade, os metadados. São as colunas requisição e operação.

Nessas colunas, como será explicado posteriormente, é registrado algo similar a um instantâneo das ações ou eventos que os documentos e processos enfrentam. Ambas as colunas são do tipo *varchar(max)*, com tamanho limitado apenas pela capacidade de armazenamento em disco.

É perceptível a presença de uma sobreposição considerável (duplicidade de informações) entre essas colunas. Contudo, nos casos analisados, a coluna operação tende a conter

detalhes mais ricos. Por tal razão, o presente trabalho focou exclusivamente nesta última, tendo em vista tanto a redundância quanto o espaço e tempo necessários para abordar todo o conjunto de informações.

Essa tabela desempenha um papel predominante no que diz respeito ao espaço ocupado em disco pelo SEI.

Nesse contexto, destaca-se que quase 70% do espaço é destinado à tabela infra_auditoria (Figura 1.6), com o versionamento de documentos e os próprios documentos ocupando as posições subsequentes.

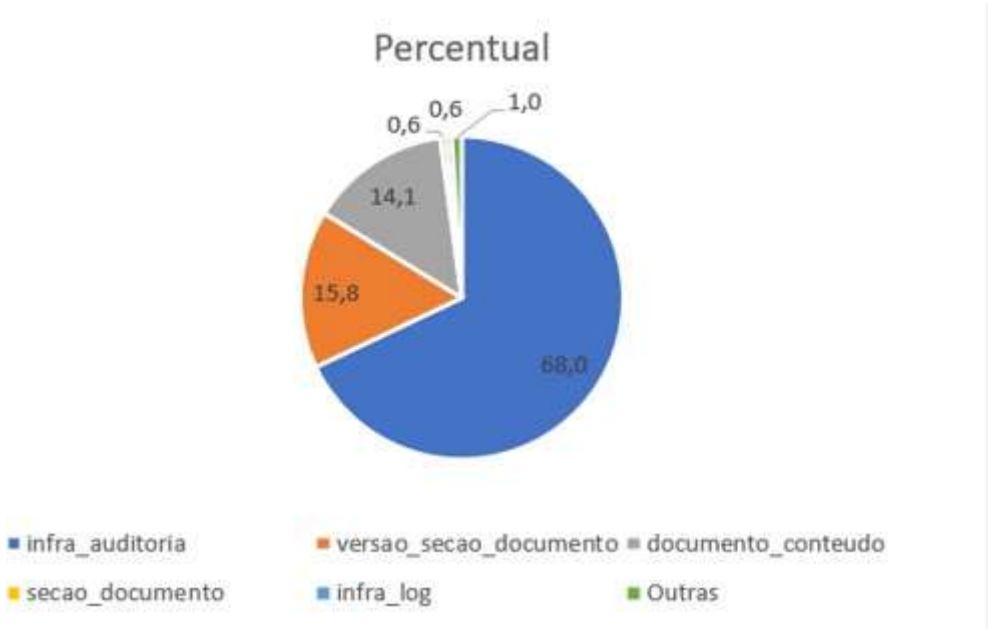

Figura 1.6: Distribuição do percentual de espaço utilizado no banco de dados pelas tabelas do sistema.

Um padrão uniforme de crescimento nos registros de log pode ser observado (Figura 1.7), que encontra justificação no modo de utilização regular do SEI, o qual não é afetado por flutuações sazonais. Vale ressaltar um aspecto importante: conforme mencionado anteriormente, em 2021 foi realizada uma tentativa fracassada de resolver essa questão por meio do particionamento dos arquivos por ano.

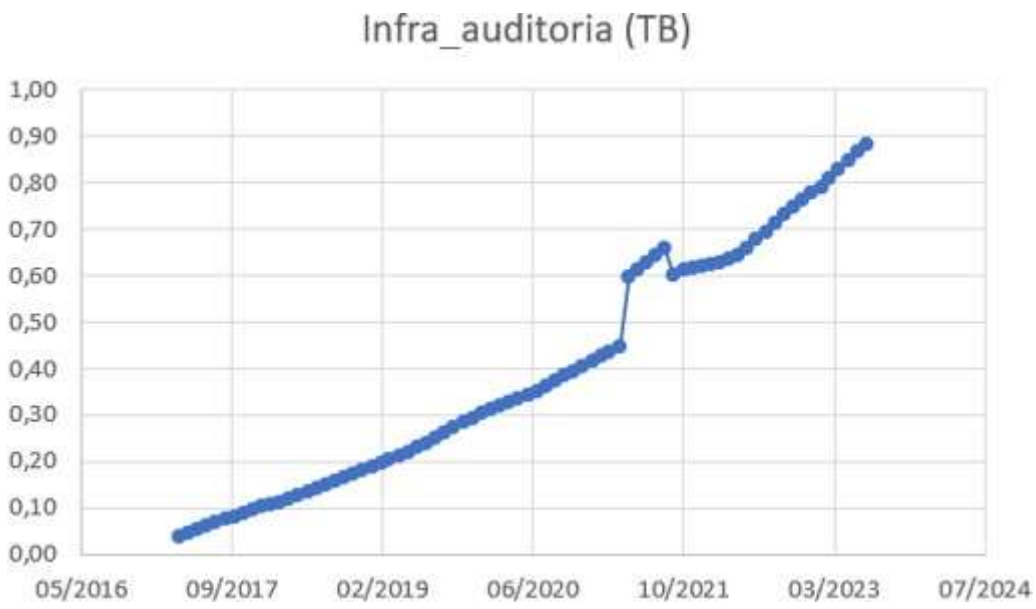

Figura 1.7: Evolução temporal do espaço alocado pela tabela infra_auditoria (em terabytes).

De acordo com o fornecedor do SQL Server (Microsoft), deveria haver ganhos consideráveis de desempenho ao dividir e classificar os dados em arquivos, uma vez que, ao consultar um registro em um banco específico, a busca seria limitada ao arquivo pertinente. No entanto, esses benefícios foram pouco evidentes nos testes realizados.

Nesse processo de segmentação, também foi decidido separar os registros com mais de 1 ano do banco do SEI. No início de cada mês, um processo automatizado recorta os dados do mês anterior e os transfere para um banco de dados separado chamado SEI_Auditoria. Esse banco contém apenas dados da tabela infra_auditoria, ocupando um espaço com recursos de hardware menos sofisticados. A estrutura é constituída por uma tabela única, também particionada por ano e indexada pela data.

A Figura 1.8 apresenta a evolução do espaço ocupado pelo SEI_Auditoria. A ausência de crescimento entre abril de 2021 e maio de 2022 pode ser atribuída ao saldo inicial de espaço alocado pelo analista ao banco. Foi criado um banco com 0,5 TB (em abril de 2021). Como o tempo, uma vez que esse saldo foi consumido (maio 2022), o banco passou a automaticamente aumentar gradual e uniformemente seu tamanho.

Figura 1.8: Evolução temporal do espaço alocado pelo banco adjacente SEI_Auditoria (em terabytes).

1.4.1 A coluna Requisição

O SEI opera como um sistema baseado na web, desenvolvido em Hypertext Preprocessor (PHP) com a utilização de objetos, classes e métodos.

Para registrar as operações nos registros de *log*, os desenvolvedores adotaram duas abordagens distintas.

A primeira, mais direta, está relacionada ao campo "requisicao".

Nesse caso, o campo consiste essencialmente em um instantâneo do objeto *HTML* em questão.

O objeto é convertido em uma *string* (por meio do método PHP denominado "printr") e então armazenado no banco de dados sem qualquer forma de tratamento adicional.

O trecho de código 1.1 foi extraído de *InfraAuditoria.php*. Esta seção do código de programação ilustra claramente essa abordagem. Notavelmente, o método realiza uma impressão literal dos métodos *GET* e *POST* do objeto *HTML*. A única adição feita é a inclusão dos separadores *GET* e *POST*.

```
1 objInfraAuditoriaDTO->setStrRequisicao('GET - '.print_r($arrGetClone,true)."\nPOST
      - ".print_r($arrPostClone,true));
```

Código 1.1: Trecho do Código PHP responsável pelo registro de instruções http.

A Figura 1.9 apresenta o formato no qual os dados são registrados no banco de dados.

```
GET - Array
(
    [acao] => procedimento_visualizar
    [acao_origem] => procedimento_visualizar
    [id_procedimento] => 9851047
    [infra_sistema] => 100000100
    [infra_unidade_atual] => 110000628
    [infra_hash] => c6e7adf48c89fed1618c7a13df49db495d
```

Figura 1.9: Amostra do formato da informação gravada no banco de dados.

Neste formato, temos a esquerda o nome do atributo e a sua direita seu respectivo valor. Neste caso, o atributo "acao"tem relação ao método que foi executado.

1.4.2 A coluna Operação

Há também uma segunda abordagem, a qual fora adotada para armazenar dados na coluna operação, conforme listado no trecho de código a seguir. (Listining 1.2).

```
1  public function formatarDados($varParam, $iteracao = 0){
2      $ret = '';
3      if ($varParam instanceof InfraDTO){
4        $prefixo = "\n".str_repeat(' ',$iteracao);
5        $ret = $prefixo.' '.$varParam->__toString();
6      }else if (is_array($varParam)){
7        $n = InfraArray::contar($varParam);
8        $prefixo = "\n".str_repeat(' ',$iteracao);
9        $ret .= $prefixo.'Array ('.$n.') {';
10       foreach($varParam as $chave => $valor){
11         $ret .= $prefixo.' ['.$chave.'] => '.$this->formatarDados($valor,$iteracao
               +1);
12       }
13       $ret .= "\n".'}';
14     }else if ($varParam === null){
15       $ret = '[null]';
16     }else if ($varParam === ''){
17       $ret = '[vazio]';
18     }else{
19       $ret = $varParam;
20     }
21     return $ret;
22   }
```

Código 1.2: Código da função PHP "formatarDados", responsável pelo registro em banco da operação.

Nesse caso, a informação é amazenada no campo "operacao".

Agora, os programadores empregam uma função recursiva que recebe um objeto. Essa função tem a capacidade de discernir se o objeto é do tipo "auditoria"ou se consiste em um array de informações.

Esse processo permite a preservação de detalhes mais refinados, particularmente em objetos que apresentam múltiplos atributos.

As figuras a seguir servem como amostras, tomadas ao acaso, de registros registrados na coluna "operacao", oferecendo uma visão do formato pelo qual os dados são inseridos no banco de dados. Trata-se de exemplos dos métodos Exibir (Figura 1.10), Assinar (Figura 1.11), Visualizar (Figura 1.12) e Consultar (Figura 1.13). Dado que esses registros são gerados através da conversão de objetos em texto, conforme explanado na seção anterior, eles mantêm precisamente o formato do objeto original.

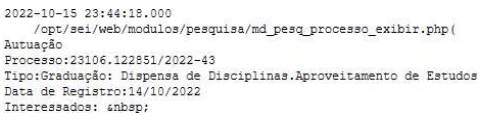

Figura 1.10: Amostra de ação Exibir

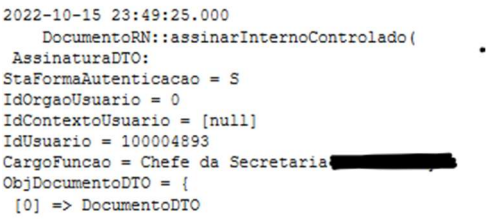

Figura 1.11: Amostra de ação Assinar

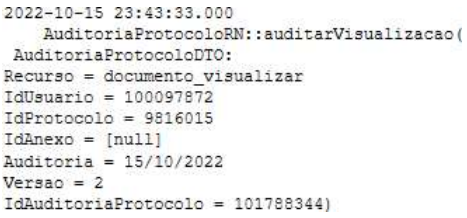

Figura 1.12: Amostra de ação Visualizar

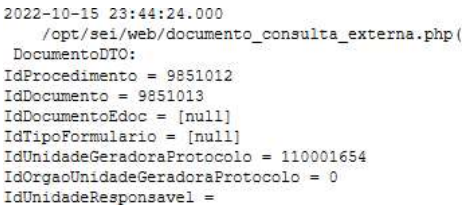

Figura 1.13: Amostra de ação Consultar

Nas referidas figuras, o primeiro atributo listado abaixo da data é o nome da classe, o qual corresponde à ação executada no sistema, ou seja, "exibir", "assinar", "visualizar"e "consultar", respectivamente. Os atributos subsequentes são específicos para cada tipo de objeto e não seguem a mesma estrutura uniforme.

Ao observar os dados representados nas figuras, fica claro que eles se referem a eventos distintos ("exibir", "assinar", "visualizar"e "consultar"). Embora todos esses eventos se-

jam armazenados no mesmo campo ("operacao"), cada um possui seus próprios atributos únicos, que podem variar em quantidade, nome e tipo de dado.

A primeira linha fornece a indicação do tipo de objeto, correspondendo a um total de 243 tipos únicos, análogos a classes e denominados como "OPERAÇÃO"no contexto deste estudo.

Cabe ressaltar que essas operações não se conformam a um padrão uniforme. Elas não compartilham os mesmos atributos, os quais podem variar conforme o tipo de objeto em questão.

Foram identificados 23.581 atributos.

A Lei Geral de Proteção de Dados (LGPD), em seu texto, estabelece critérios de privacidade que devem ser estritamente observados e obedecido. Neste caso, por se tratar de sistema arquivístico, uma grande quantidade de informações sigilosas, de acesso limitado, ou mesmo restritas, podem causar grande impacto caso tenham sua integridade comprometida.

Com a vênia do Encarregado de Proteção de Dados [8] os trabalhos são desenvolvidos sem que a referida proteção dos dados sejam colocados a prova.

Em suma, durante a fase de programação, não foi dada uma atenção específica ao modo de armazenamento dos dados. A abordagem adotada a época, limitou-se a registrar esses dados no banco, sem considerações mais detalhadas a respeito da sua estrutura. Aliado a este fato, o grande volume de dados dificulta o acesso a um registro específico a um bom termo, sem que se ponha em risco a privacidade dos mesmos, conforme proposto pela LGPD. A proposta é facilitar o acesso as informações de registros dessas interações do SEI, hoje concentrados em uma única coluna (operacao), de uma única tabela (infra_auditoria), organizando-os e usando de técnicas de otimização ("tunning") para assim viabilizar consultas pelo próprio usuário.

Capítulo 2

O Problema

Um caso que ilustra a presente pesquisa ocorreu quando a Secretaria de Tecnologia da Informação (STI) recebeu uma solicitação para esclarecer quem havia acessado um processo específico.

O cenário envolvia uma seleção de colaboradores, e o processo foi aberto de forma equivocada. No lugar de ser classificado como "sigiloso", restringindo o acesso a um grupo seleto, ele foi erroneamente classificado como "restrito".

Processos restritos normalmente contêm informações de natureza pessoal e funcional (vida privada, imagem, prontuário). Assim, o acesso é restrito apenas aos membros da mesma unidade.

Processos sigilosos contêm informações com o objetivo de proteger os interesses do órgão (seja orçamentária, financeira ou institucional). O acesso precisa ser explicitamente atribuído a um usuário.

No caso, havia a suspeita de que alguém dentro daquela unidade tivesse acessado o processo de seleção, potencialmente favorecendo um candidato ao cargo.

Consequentemente, a unidade abriu um chamado junto à STI, requisitando uma lista das pessoas que haviam acessado o processo em questão.

O analista encarregado tentou obter a informação usando a tela do SEI que permite realizar consultas ao log. Contudo, este resultado não pode ser obtido apenas especificando poucos parâmetros, ou buscando-os nos campos Requisição ou Operação. Isto fica evidente ao observarmos a referia tela (Figura 2.1). Ela traz em destaque o aviso, em letras vermelhas, "ATENÇÃO: Informar o maior número possível de critérios antes de realizar a pequisa".

Na tela apresentada na Figura 2.1, é possível realizar buscas detalhadas utilizando diferentes critérios, como usuário (sigla e nome), unidade (sigla e descrição), processo (recurso e período), conexão (IP, servidor), e fato (requisição e operação). Cada um destes campos refere-se a uma das colunas da tabela infra_auditoria.

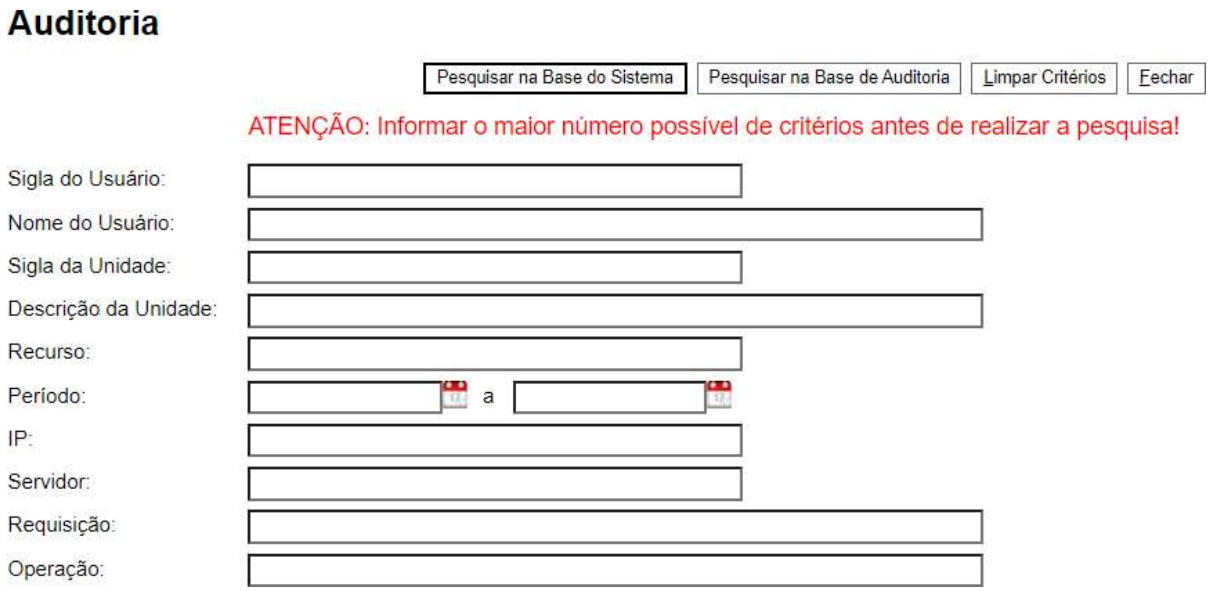

Figura 2.1: Imagem da tela de consulta disponível no SEI para auditoria.

O problema tornou-se mais sério quando a tentativa de recuperar os dados de "Requisicao"e "Operacao", os dois últimos campos, foi realizada. Isso ocorreu em nosso exemplo. Devido à ausência de atributos relacionais indicativos do tipo de operação (visualização, edição, etc.), tivemos que recorrer a esses dois campos.

Entretanto, não houve sucesso. A sessão do PHP expirou devido a um time-out.

Por conseguinte, o mesmo usuário nos enviou essa consulta para ser executada diretamente no console do SQL Server.

Na primeira tentativa, após a execução por mais de 24 horas, a execução foi manualmente interrompida. Não era possível identificar se ainda se encontrava em execução ou sequer se houvera falha da consulta. Então surgiu a ideia de limitar o período de tempo. Foram feitas três tentativas: um ano, seis meses e, finalmente, trinta dias. O sucesso foi alcançado após alguns minutos de espera (cerca de 20 minutos) nesta última tentativa.

Esse caso se destacou, não só devido à série de desafios enfrentados, mas também pelo longo tempo necessário para a execução. Era uma consulta complexa, que buscava informações específicas nos campos de texto da tabela "infra_auditoria".

Além disso, existem outras demandas, que constituem a maioria dos casos, que podem ser tratadas pelo analista responsável em atender o questionamento do usuário. Claro, isso envolve a restrição do número de campos e requer paciência. Sem a delimitação da data, as chances de sucesso são praticamente nulas. São questões mais simples, onde se busca quem editou ou quem leu determinado despacho, ou ainda que fez a exclusão.

2.1 Pergunta de Pesquisa

Dessa forma, pode-se chegar a seguinte questão de pesquisa:

Q1: Como proporcionar consultas analíticas e auditoria nas informações armazenadas no SEI com um tempo de resposta viável sem exigência de alguma variável?

2.2 Hipótese de Pesquisa

Acredita-se que, ao se extrair informações do SEI, limpá-las e organizá-las em um formato apropriado, será possível reduzir o tempo de resposta às consultas relacionadas à auditoria e analíticas.

2.3 Objetivo Geral

A partir do que foi exposto, este trabalho objetiva extrair, limpar e organizar as informações do SEI, em um formato apropriado, para agilizar e permitir consultas analíticas e auditoria.

2.4 Objetivos Específicos

Para alcançar o objetivo geral da presente pesquisa, serão percorridos os seguintes objetivos específicos:

- Analisar o formato e volume de informações de auditoria armazenadas no SEI;
- Extrair e organizar as informações de auditoria do SEI;
- Persistir as informações do SEI em um repositório a parte;
- Validar este repositório por meio de execução de consultas analíticas e de auditoria;

Capítulo 3

Procedimento Metodológico

Este capítulo analisa os aspectos da Metodologia Científica, também denominada Procedimento Metodológico. A seção 3.1 examina a classificação desta abordagem. Em seguida, a metodologia utilizada é discutida na seção 3.2, complementada pela seção 3.3, que aprofunda aspectos relevantes. Finalmente, os procedimentos de preparação são detalhados na seção 3.3.1.

Os detalhes completos do desenvolvimento metodológico estão apresentados no anexo I.4, que foi separado para proporcionar uma leitura mais fluida.

3.1 Classificação da Pesquisa

Este projeto pode ser classificado de acordo com o critério "Natureza", conforme descrito por Marconi [9], uma vez que lida com resultados aplicáveis à solução de problemas do mundo real, como enfatizado pelo autor.

Nesse contexto, a natureza da pesquisa é "predominantemente quantitativa", pois os resultados podem ser avaliados numericamente com base no desempenho do tempo de resposta.

Já, segundo o trabalho de Wazlawick [10], a pesquisa se enquadra na categoria "Original". Ela cria um conjunto de ferramentas de administração de dados, ainda que aplicadas a um caso específico, e emprega técnicas baseadas em boas práticas e conceitos estabelecidos, bem como conhecimentos adquiridos ao longo do tempo. O objetivo é alcançar resultados eficazes por meio dessas técnicas e, em consonância com a literatura, explicar o raciocínio por trás dos resultados obtidos.

Adicionalmente, também de acordo com Wazlawick [10], esta pesquisa se qualifica como "Pesquisa Explicativa", pois analisa o desempenho e o comportamento observado, procurando explicá-los para oferecer interpretações e alternativas. Ainda, segundo este

autor, também é considerada "Experimental", uma vez que envolve a aplicação de técnicas computacionais seguidas de observações para mensurar diferenças de desempenho.

3.2 Metodologia dos Trabalhos

A abordagem adotada para lidar com o problema proposto neste trabalho foi fundamentada principalmente na revisão da literatura especializada. Nessa revisão, destacam-se as contribuições de renomados autores, tais como [11] e [2], que são amplamente reconhecidos no campo de banco de dados. Além disso, [12] apresenta perspectivas adicionais relevantes.

Dentro desse contexto, o objetivo do trabalho é elaborar um plano para a implementação de uma nova estrutura destinada a armazenar dados em um repositório especializado conhecido como Data Warehouse (DW). O próximo passo envolve a extração de dados do banco de dados do SEI, a sua transformação e, finalmente, o carregamento desses dados no novo repositório, que foi otimizado para aprimorar o desempenho das consultas analíticas e de auditoria.

Adicionalmente, houve uma busca sistemática nas bases de dados acadêmicas *Web of Science* e *Scopus* com o propósito de identificar trabalhos relacionados que abordassem problemas semelhantes. A ideia era comparar esses trabalhos e suas respectivas soluções propostas.

3.3 Teoria do Enfoque Meta Analítico Consolidado (TEMAC)

A metodologia TEMAC, descrita por Mariano[13], adota três etapas para a pesquisa, a saber: a Pesquisa de Periódicos, que visa identificar literatura de impacto entre as publicações científicas; a Análise por Técnicas de Bibliometria; e, por fim, a Pesquisa Exploratória com abordagem quantitativa. Os detalhes desta etapa são apresentados no Anexo I (I.4).

A escolha da língua inglesa se deve à sua vasta quantidade de publicações. As consultas foram realizadas em janeiro de 2023. Os descritores selecionados de forma empírica foram *data warehouse*, *analytical research*, e *information management systems*. A busca foi restrita aos últimos cinco anos para garantir a atualidade das informações sobre o assunto.

A escolha do termo *Data Warehouse* se justifica pelo fato de que as técnicas utilizadas neste trabalho estão incluídas nas ferramentas de ETL. Além disso, em consonância com

as transformações aplicadas, buscou-se fornecer informações sobre o uso do sistema como um todo, o que motivou a inclusão do termo *analytical research* entre os argumentos.

O SEI é um sistema típico de administração de informações, como evidenciado pelo próprio nome. Portanto, foram buscados textos que também abordassem *information management systems*.

Adicionalmente, foram realizadas pesquisas incluindo o termo *system usage records* na tentativa de encontrar referências sobre a manipulação e a captação de informações relacionadas ao registro do uso do sistema como um todo. No entanto, esse termo não resultou em acréscimo de resultados relevantes.

3.3.1 Preparação da pesquisa

As expressões abaixo correspondem às estratégias de pesquisa utilizadas nas bases de dados Scopus e Web of Science, respectivamente. Elas foram detalhadas durante a fase de qualificação deste trabalho e serviram para prospectar obras que pudessem fornecer referências relevantes para contribuir com o desenvolvimento da pesquisa.

Scopus

A expressão utilizada no Scopus, importante fonte de artigos relacionados a tecnologia, foi "*(TITLE-ABS-KEY (data AND warehouse) OR TITLE-ABS-KEY (analytical AND research) OR TITLE-ABS-KEY (information AND management AND systems)) AND PUBYEAR > 2017 AND PUBYEAR > 2017* "que por sua vez retornou pouco mais de 170 mil registros.

Web Of Science (WoC)

A expressão aplicada foi "*((ALL=(data warehouse)) or ALL=(analytical research)) or ALL=(information management systems) and 2023 or 2022 or 2021 or 2020 or 2019 or 2018 (Anos da publicação)*". Foram encontrados 314.145 registros.

Limitados aos últimos 5 anos, também mostrou uma queda no último ano, contudo significativamente mais discreta que a do Scopus.

A seção a seguir apresenta uma análise dos trabalhos recuperados nas bases de dados.

Capítulo 4

Trabalhos Relacionados

Neste capítulo, será realizado um mapeamento das técnicas para tratamento de grande volume de dados, conforme literatura e trabalhos científicos listados na seção anterior. O objetivo é identificar aquelas que já foram exploradas em outros trabalhos, a fim de buscar otimizações para os resultados deste estudo.

4.1 Análise dos Trabalhos

Inicialmente, DWs foram concebidos com o propósito de organizar informações dispersas, principalmente em grandes bases de dados. No entanto, ao longo do tempo, sua função foi transformada de um objetivo em si para um meio essencial. Um grande número de sistemas de mineração de dados tornou-se intrinsecamente dependente dos DWs. Diversos estudos têm explorado esse campo, como destacam [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25].

Nesse contexto, tem havido esforços consideráveis para desenvolver um padrão de linguagem para OLAP [26]. Isso representa um desafio significativo, uma vez que está relacionado à programação, cuja eficiência está diretamente ligada à experiência do programador.

O estudo em [27] tem como objetivo avaliar a melhor forma de indexar repositórios de DW e é motivado pelo considerável aumento previsto de 200% nos acessos e compartilhamentos na internet em um curto período. Esse estudo ressalta a importância dos DWs e da DM na abordagem de questões para os usuários finais.

A construção e manutenção de DWs são atividades dispendiosas devido à complexidade envolvida, à necessidade de tecnologia de ponta e ao consumo contínuo de energia e recursos humanos especializados. Isso justifica a pesquisa na direção da mitigação do desperdício, como evidenciado no trabalho em [28], que analisa como otimizar e reduzir o consumo de energia na criação de DWs por meio da adoção de níveis de maturidade

desses DWs. A avaliação é realizada medindo a eficiência de cada etapa do ciclo de vida, desde a concepção até o usuário final.

A adaptação às tecnologias adequadas é abordada por [29]. A adoção precoce de novas tecnologias costuma ser dispendiosa, portanto, requer avaliações contínuas. Neste estudo, o autor adota o *benchmarking* de *warehouses* e informações sobre a estrutura dos dados (como o número de atributos) para fornecer informações para mecanismos de *machine learning*.

Uma ferramenta de código aberto para facilitar o gerenciamento de recursos de computação de alto desempenho é apresentada por [30]. Surpreendentemente, a proposta de usar um DW para controlar recursos de hardware consumidos por DWs com grandes volumes históricos é bastante interessante.

O trabalho de [31] demonstra como recursos simples, como a adoção de um carimbo de tempo (*timestamp*), podem melhorar o desempenho dos resultados de pesquisas em um DW composto por dados de fontes distintas.

No contexto dos Sistemas de Gerenciamento de SGBDR, as chamadas visões materializadas (*materialized views*) desempenham um papel importante. Elas são consultas pré-definidas que são atualizadas pelo SGBDR à medida que os registros de origem são atualizados. O estudo em [32] investiga como essas visões podem mitigar o uso excessivo de DWs. O autor as propõe como ferramentas úteis, tanto para consultas conhecidas quanto para consultas *ad-hoc*, sendo processadas tanto *off-line* para otimizar consultas conhecidas quanto *on-line* para otimizar consultas *ad-hoc*.

É comum encontrar trabalhos que coletam dados de alunos por meio de processos de Extração, Transformação e Carregamento (ETL), organizando-os e submetendo-os à mineração de dados em busca de inferências, como demonstrado em [18] e [19].

O trabalho chinês de [21] extrai conclusões a partir do repositório de textos da literatura mundial e comparativa, explorando dados de diversas origens.

Nos Estados Unidos, a aprovação da lei *Health Information Technology for Economic and Clinical Health (HITECH)*, que trata da automação de informações de prontuários médicos, levou à adoção generalizada de registros eletrônicos de saúde [16]. Isso resultou na criação de grandes sistemas de saúde com uma variedade diversa de propósitos, que têm potencial significativo para o futuro da medicina.

A partir desses registros, [33] utiliza dados coletados durante anestesias para auxiliar colegas na adoção de boas práticas nessa especialidade.

Em Uganda, mesmo sem legislação semelhante à HITECH, [17] relata que estão sendo compilados 20 anos de dados sobre HIV para desenvolver estratégias e diretrizes relacionadas a esse vírus, que tem um impacto semelhante ao que a COVID-19 tem tido recentemente.

Na Alemanha, [14] está promovendo a construção de ciclovias por meio da consolidação de repositórios heterogêneos usando ETL para formar um *datalake*. Esses dados podem variar desde informações de comitês de moradores até dados de entidades estatais.

Datalakes também são estudados por [34], que os define como repositórios de dados brutos de várias fontes, processados sob demanda.

Na China, [35] utiliza um DW de dados de saúde para minerar informações e fiscalizar seguros de saúde por meio de inferências.

Também na China, [23] se concentra na mineração de dados coletados em educação física, utilizando aprendizado profundo (*deep learning*) para desenvolver estratégias de ensino nas escolas.

Enquanto os autores de [36] preocupam-se com a saúde mental dos estudantes, eles analisam notas e desempenho escolar para fazer inferências em um grande volume de dados acadêmicos, buscando prever comportamentos de alunos que possam necessitar de apoio nas instituições de ensino superior para reduzir a evasão.

Na Arábia Saudita, [22] está trabalhando na criação de um repositório ubíquo para dados de saúde, semelhante ao estabelecido pela HITECH. O objetivo é tirar conclusões de repositórios de dados de saúde de diversas fontes.

O trabalho de [24] introduz a ideia de "Descobrir Conhecimento", que se refere à organização de informações por meio de ETL de maneira adequada para possibilitar inferências. Trata-se de abordagem alternativa para a mineração de dados. O autor lida com dados de diversos atores na área da saúde em Bangladesh, desde farmácias até grandes centros de saúde.

Os russos também estão focando na área da saúde. [25] consolida repositórios diversos com ênfase na segurança dos dados.

A análise de dados de mercado sempre foi uma tarefa desafiadora, que favorece aqueles que são mais habilidosos na manipulação de informações. [15] e muitos outros trabalhos propõem ferramentas de DW genéricas para o mercado financeiro, projetadas para fornecer uma base centralizada de armazenamento de informações associada a um *framework* de *Big Data Investment Application*. A partir desse repositório, são extraídas inferências que podem proporcionar vantagens competitivas.

A escassez de energia é um dos principais obstáculos para o crescimento econômico. Para abordar essa questão, a China desenvolveu um plano econômico para incentivar pesquisas que contribuam para o uso mais eficiente da energia. Isso provavelmente influenciou o número de artigos relacionados a esse tema na literatura. Dentre esses, destacamos dois.

O primeiro é o trabalho, em [37], a "política de desenvolvimento estratégico"é abordada, usando diversas informações, como previsões meteorológicas e dados de consumo,

para prever a melhor forma de gerenciar o setor de energia por meio de aprendizado de máquina.

O segundo apresenta a agregação de dados de leituras de equipamentos de mineração, abordada por [20]. Esses dados são usados para gerar inteligência e permitir o controle remoto de subestações. Os dados são classificados e padrões são identificados para ajustar os parâmetros de alarme.

Capítulo 5

Fundamentação Teórica

Este capítulo apresenta a Fundamentação Teórica, abordando os principais conceitos e tecnologias considerados essenciais para o desenvolvimento desta pesquisa. Na Seção 5.1, são discutidos os conceitos fundamentais relacionados aos Sistemas de Apoio à Decisão (SAD). A seguir, a Seção 5.2 explica conceitos envolvidos em um um DW para, em seguida, a Seção 5.3 descrever seu ciclo de vida. A Seção 5.4 oferece uma visão geral dos princípios dos bancos de dados NoSQL, destacando as diferenças em relação aos bancos de dados relacionais tradicionais. Por fim, na Seção 5.5.1, são apresentados conceitos relacionados à Lei Geral de Proteção de Dados (LGPD), dada sua relevância direta para as questões atinentes ao SEI.

A solução proposta neste trabalho é baseada em conceitos e definições extraídos da literatura, incluindo tanto obras canônicas e amplamente reconhecidas, como [2] e [11], quanto fontes mais especializadas, como [38]. Além disso, incorpora o que há de mais recente produzido por especialistas em ferramentas do mercado, a exemplo do SQL Server [12].

5.1 Sistemas de Apoio à Decisão (SAD)

Sistemas de Apoio à Decisão (SAD) são uma classe de sistemas de informação que auxiliam na tomada de decisão. Visa organizar a informação, através de ferramentas e modelos, viabilizando interpretações e análises. São particularmente úteis em cenários de decisão complexos, semi-estruturados ou não estruturados, onde o julgamento humano é essencial.

Esses sistemas abrangem diversas áreas, começando pela gestão de dados [39], frequentemente utilizando DW para armazenamento centralizado, o que facilita a recuperação e manipulação das informações necessárias. Eles também incluem a gestão de modelos [40], como modelos estatísticos e financeiros, que auxiliam na simulação de cenários e resultados, permitindo a exploração de alternativas e seus impactos. Além disso, contam

com interfaces de usuário [41] que tornam as interações mais acessíveis e intuitivas. Esses sistemas oferecem suporte a decisões [42], onde uma combinação de dados estruturados e julgamento humano ou intuição é essencial. Por fim, os SAD são flexíveis e adaptáveis a diversos processos de tomada de decisão e contextos, sejam eles operacionais, táticos ou estratégicos, podendo ser personalizados para atender às necessidades específicas da organização [43].

É útil situar o SAD no contexto dos Sistemas de Informação. Onde, de maneira resumida, são classificados em quatro níveis de decisão dentro de uma organização[1]: operacional, de conhecimento, gerencial e estratégico. A Figura 5.1 ilustra as principais aplicações para cada nível a saber:

- **Nível estratégico**: Sistemas de Apoio Executivo (SAE) [43].
- **Nível gerencial**: SAD e Sistemas de Informações Gerenciais (SIG) [41], com este trabalho focando especificamente nesse nível.
- **Nível do conhecimento**: Sistemas de Trabalhadores do Conhecimento (STC) e Sistemas de Automação de Escritórios [42].
- **Nível operacional**: Sistemas de Processamento de Transações (SPT) ou Sistemas Transacionais [39].

Figura 5.1: Classificação de Sistemas de Informação em Níveis de Decisão (adaptado de [1])

SAD e SIG possuem diferenças importantes. SIG é direcionado a problemas mais rotineiros, utilizando dados fornecidos pelos SPT e ferramentas tradicionais, como relatórios. SAD lida com problemas não rotineiros, portanto mais complexos e não usuais. O SAD pode consumir informações tanto internas, geradas pelos sistemas transacionais e pelos SIG, quanto externas, com o objetivo de complementar a consulta de interesse.

5.2 Arquitetura

A UnB optou por utilizar a arquitetura Online Analytical Processing (OLAP) [2] para o tratamento dos dados no SAD. Nessa arquitetura, destaca-se os conceitos de Data Warehouse (DW) e Business Intelligence (BI), sendo o DW um repositório multidimensional capaz de armazenar dados provenientes de diversas origens e formatos, e esta ferramenta que disponibiliza os dados para os usuários. O processo de carga é conhecido como Extraction, Transformation and Loading (ETL), que envolve a obtenção dos dados de suas fontes (extração), sua formatação e correlação (transformação), e, por fim, o seu armazenamento no DW (carga). Esses dados podem ser tanto internos à organização quanto externos.

A arquitetura OLAP é composta por uma fonte de dados; uma camada ETL, um DW e um Servidor de Relatórios Analíticos. O servidor de Relatórios Analíticos possibilita consultas sob demanda, que permitem uma navegação pelos dados organizados em dimensões no DW e a geração de painéis (*dashboards*) ou relatórios sem a necessária geração de código. O trabalho de [44], divide um SAD em dois sub sistemas: DW ("recebendo os dados") e BI ("entregando os dados"). A Figura 5.2 demonstra um DW comum.

Figura 5.2: Arquitetura de um DW ([2] adaptador por [3])

De maneira simplificada, segundo [45], dados podem ser entendidos como um conjunto qualquer de informação, um registro de acesso para exemplificar. A partir dos dados inferir algum tipo de informação, a exemplo, pode-se obter a que horas um sistema foi acessado. A partir daí pode-se evoluir o conhecimento até a competência. Com o conhecimento é possível saber o horário de maior movimento. E a partir competência fazer elucubrações, como por exemplo o que é possível se fazer para majorar o número de acesso.

Uma arquitetura OLAP pode ser implementada e classificada por diversas maneiras, a depender do repositório. Assim, o modelo Multidimentional Online Analytical Processing (MOLAP) os dados ficam armazenados em estrutura multidimensional. Enquanto que Relacional Online Analytical Processing (ROLAP) em estrutura relacional. A vantagem é utilizar tecnologia padronizada com escalabilidade e paralelismo de hardware. Enquanto que a desvantagem fica por conta de funções para análises dimensionais e o baixo desempenho da linguagem SQL.

Há ainda o conceito de Not Online Analytical Processing (NoLAP), que consiste em uma arquitetura OLAP baseados em Not SQL (NoSQL). Este, por sua vez é a abordado no trabalho de [46]. Aqui o conceito, embora não tenha sido chamado desta forma, utilizou estrutura NoSQL sobre arquitetura OLAP em ecossistema *BIG DATA* para tratamento e análise de dados abertos governamentais.

5.2.1 Data Warehouse (DW)

Warehouse, em sua tradução literal, refere-se aos grandes cilos de armazenamentos de grãos, capazes de armazenar grandes safras por períodos indefinidos. Data Warehouse é pois uma metáfora deste conceito, dedicando-se exclusivamente a dados.

O objetivo do DW é ([47]) "extrair, transformar e carregar dados de diferentes sistemas de origem em um repositório integrado". Isso se dá transpondo obstáculos, como origem diversas, fontes heterogêneas e ainda redundância.

A literatura oferece várias definições sobre DW. Segundo [11], um DW é uma coleção de dados orientada por assuntos, integrada, não volátil e variável ao longo do tempo, com o objetivo de apoiar os processos de tomada de decisão. Em [2], o DW é descrito como uma cópia de dados transacionais estruturada especificamente para consultas e análises. Já em [39], um DW é tipicamente um sistema de banco de dados dedicado, separado dos sistemas de Online Transaction Processing (OLTP) da organização. De acordo com Sen [48], o DW é construído para suportar a tomada de decisões empresariais, contendo dados históricos, sumarizados e consolidados provenientes de registros individuais de bancos de dados operacionais. Por fim, [49] define o DW como um banco de dados analítico, somente leitura, que serve como base para os SAD.

Por seu tamanho e complexidade, o DW pode ser classificado também com Data Mart (DM). De acordo com Kimball [2] um DW é união de DM orientados por assuntos, obedecendo acertos critérios: gravar os dados na maior detalhe possível (maior granularidade) organizados por dimensões e fatos, sempre tendo em vista a ubiquidade do significado. Normalmente um DM é dedicado a um único processo do negócio [47].

A solução aqui proposta não se trata de um DW, outro sim de um Data Mart (DM), que se distingue daquele pelo seu tamanho (menor) e finalidade, limitada a uma porção dos dados [50] e [2]. Contudo há sim emprego de métodos de ETL haja visto que, dados contidos em tabelas são interpretados, compilados e carregados em outra estrutura, otimizados de maneira a favorecer o desempenho. Vale ressaltar que, genericamente, é comum tratar DM como se DW o fosse, dado que aquele é uma generalização deste. São pois de grande valia para tomada de decisões [11].

Segundo [11], *Data Warehouse* é "uma coleção de dados, orientada a assuntos, integrada, variável no tempo e não volátil, para suporte ao gerenciamento dos processos de tomada de decisão"e desta forma:

- **Integrado**: Os dados são consolidados em armazém de dados (DW), partir de fontes distintas e consolidados em um ambiente íntegro e consistente. Fontes estas que podem ser internas (da organização), ou não e ainda de formatos diversos (planilhas, textos, listas).
- **Variável no tempo**: Possui algum tipo de *time stamp*, que faz a indicação do período em particular.
- **Não volátil**: Os dados são ali inseridos, porém nunca alterados. São "estáveis".

5.2.2 Extraction, Transformation and Loading (ETL)

Em um DW, são armazenados dados multidimensionais provenientes de diversas fontes, como SGBDR, planilhas eletrônicas e até arquivos de texto. Esses dados passam por um processo conhecido como *Extraction, Transformation and Loading (ETL)*, que consiste na extração, transformação e carga das informações, permitindo assim uma análise integrada e coesa.

A Figura 5.3 ilustra a arquitetura descrita por [2]. Nesse modelo, os dados passam por um primeiro processo de transformação, conhecido como *Back Room*, *Back stage* ou *Back End*. No *Back Room*, os dados são submetidos a um processo de ETL, visando otimizar o desempenho e facilitar o acesso às informações mais relevantes. O resultado desse processo é armazenado no DW, também chamado de EDW. O DW serve como base para o *Front Room*. O *Front Room* é uma ferramenta voltada para o usuário final, que

permite a criação e consulta de relatórios analíticos. Nesse ambiente, um novo processo de ETL pode ser aplicado para gerar um subconjunto de dados, otimizado para a análise específica.

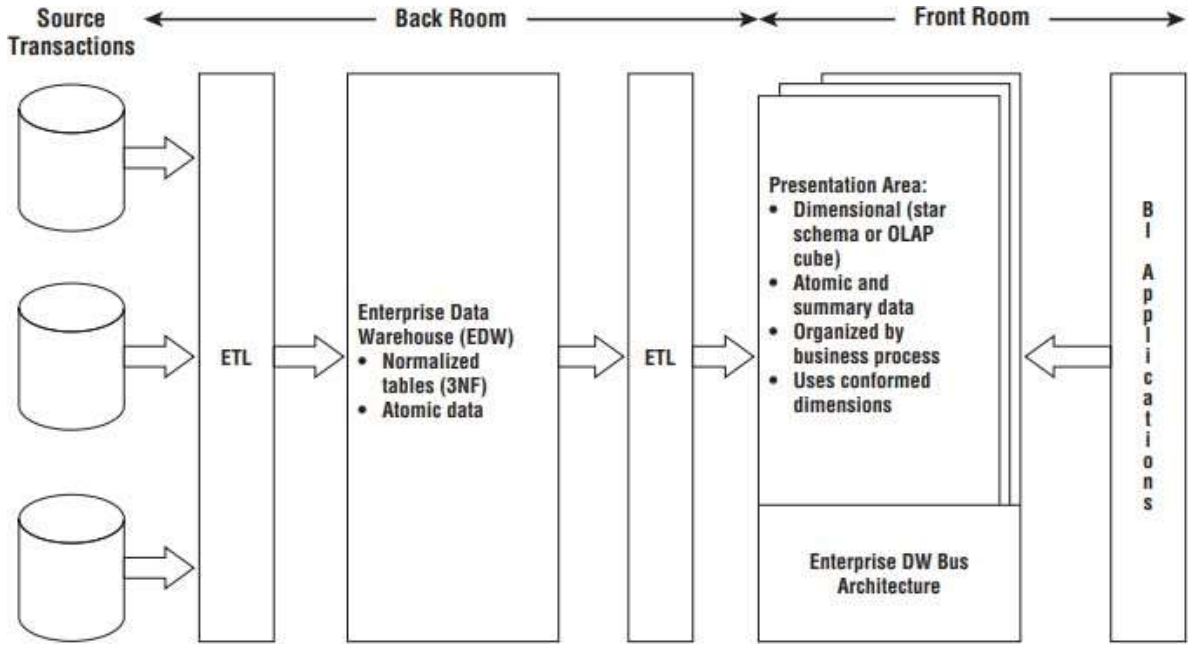

Figura 5.3: Arquitetura híbrida com estruturas 3FN e área de apresentação dimensionado [2]

5.2.3 Business Intelligence (BI)

BI é o consumidor deste produto. Recupera dados do DW fornecendo ao SSD orientado ao negócio através de ferramentas (como *dashboards*) ou ainda alimentando modelos preditivos. Na Figura 5.3 pode ser visto na última coluna, à direita.

O texto em [51] demonstra que BI pode ser usada efetivamente para integrar, e portanto entregar, grandes volumes de dados e ainda dar suporte a uma variedade de questões. Relatórios Analíticos lidam com consultas sob demanda, que permitem navegar através das dimensões do DW e geram relatórios e painéis customizados por livre demanda do usuário, até mesmo com técnicas de *drag and drop* através dos *dashboards*.

5.3 Ciclo de vida

O trabalho de [52] indica o caminho do desdobramento deste tipo de trabalho que pode se dar tem três etapas: identificação do problema, planejamento e elaboração do fato,

dimensão e esquema; construção do ETL; montagem do DW; por fim disponibilização por OLAP.

Em complemento, [2] (Figura 5.4) propõe um ciclo de vida do DW composto pelas seguintes atividades (também tratado em [3]):

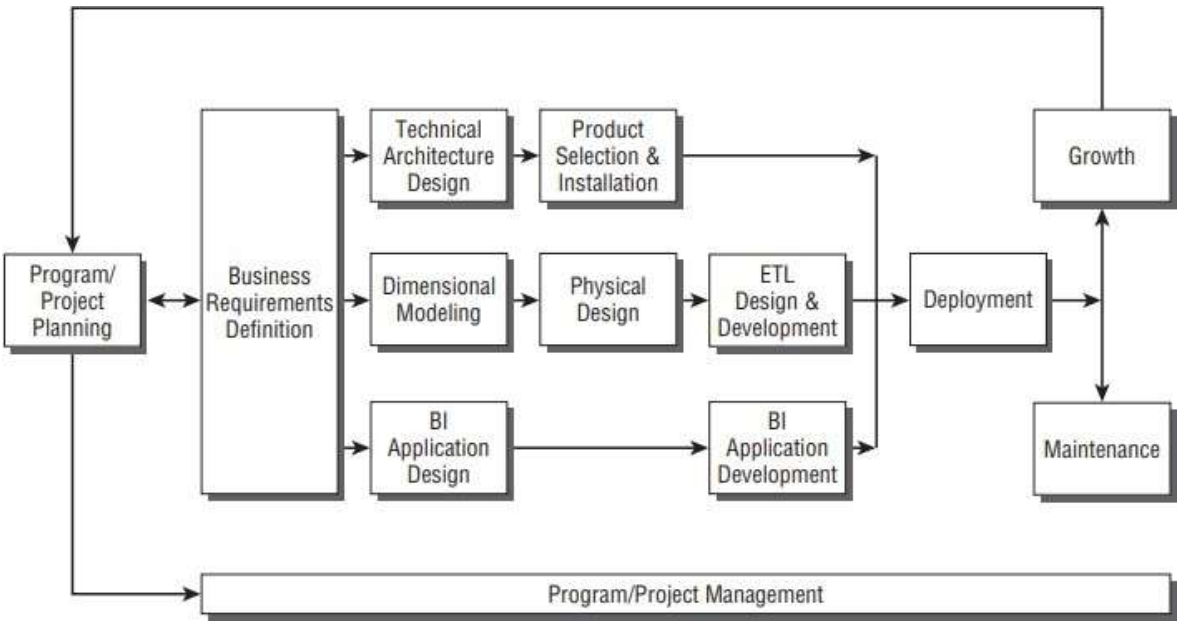

Figura 5.4: Diagrama do Ciclo de Vida de um DW [2]

- **Planejamento do Projeto**: Elaboração do plano, documento que contém o detalhes da execução de todas as tarefas.
- **Definição dos Requisitos**: Dividido em Levantamento das Necessidades do Negócio e Levantamento das Fontes de Dados.
  - **Levantamento das Necessidades do Negócio**: relação das necessidades que o sistema deverá atender. Identificação dos *steakholders* e requisitos funcionais e não funcionais.
  - **Levantamento das Fontes de Dados**: identificação das fonte de dados elegíveis e respectivos metadados. Ajuda não somente a construção do produto final como também o entendimento do negócio.
- **Projetar Arquitetura Técnica**: levantamento do volume de informação, tanto a respeito de bases de dados, quanto processamento e usuários simultâneos.

- **Modelagem Dimensional**: é a elaboração do modelo dimensional do DW. Identificando dimensões e fatos relativos aos temas da iteração. Cada fato deve ter dimensões, conteúdo e nível de detalhe definido.
- **Projetar Aplicação de Business Intelligence (BI)**: definição de consultas e o detalhamento dos indicadores identificados.
- **Projeto Físico**: são a elaboração dos modelos físico relacional e multidimensional. O primeiro derivado do modelo lógico, performance e controle; o segundo relativo a visões e consultas que serão oferecidas aos *steakholders*.
- **Selecionar e Instalar Produtos**: preparar o ambiente.
- **Implementar Rotinas ETL**: especificação e documentação dos processos ETL.
- **Implementar Aplicação de Business Intelligence (BI)**: implementação das consultas e fórmulas que compõem indicadores. Daqui surge o ambiente de desenvolvimento.
- **Implantação**: construção e avaliação de desempenho do ambiente de produção a partir do ambiente de desenvolvimento, que inclui documentação e treinamento de equipe de sustentação.
- **Nova Iteração**: próxima iteração levando e tratando novas funcionalidades bem como novas questões que tenham surgidos na iteração anterior.
- **Manutenção**: dar sustentação ao bom funcionamento. Inclui-se aqui eventuais funcionalidades (de porte pequeno ou moderado) que venham a surgir após o ciclo de iterações.
- **Gerenciamento do Projeto**: acompanhamento e controle da execução do projeto.

5.3.1 Modelagem Dimensional

Modelagem Dimensional é uma técnica de design de banco de dados utilizada para organizar dados de forma a facilitar a análise e a geração de relatórios [2]. Trata-se de abordagem focada para o negócio, a partir da qual os dados são organizados em torno de fatos e dimensões. "Fatos representam medidas ou métricas que se deseja analisar, enquanto dimensões fornecem o contexto para esses fatos". O objetivo é melhorar:

- **Desempenho**: A estrutura simplificada visa otimizar consultas complexas.
- **Intuitividade**: Mais fácil de entender e utilizar por usuários.

- **Flexibilidade**: Permite gerar relatórios e análises multidimensionais.

A modelagem dimensional adota uma abordagem diferenciada em relação aos bancos de dados transacionais, voltada especificamente para atender às necessidades de relatórios e análises de negócios. Nessa modelagem, existem dois tipos principais de tabelas: fato e dimensão.

No coração do modelo dimensional está a tabela fato, responsável por armazenar as métricas numéricas do negócio. Essas métricas são geradas pela interseção das dimensões, e a tabela fato contém as chaves que conectam essas dimensões. Ela deve capturar o nível mais detalhado de informações do processo de negócio disponível. Por isso, consultas à tabela fato podem acessar milhões de registros, permitindo a criação de relatórios complexos e análises detalhadas.

Por outro lado, a tabela dimensão é uma entidade independente que fornece o contexto necessário para a análise das métricas armazenadas na tabela fato. Ela geralmente possui uma chave primária artificial, chamada de *surrogate key*, e é menor em tamanho em comparação à tabela fato, armazenando informações descritivas sobre o negócio. A ligação entre as tabelas fato e dimensão é feita por meio da chave primária da tabela dimensão e da chave estrangeira da tabela fato, garantindo a integridade dos dados e a integração entre elas[52].

Modelos Dimensionais em Bancos de Dados Relacionais

- **Esquema Estrela**: Estrutura centralizada composta por uma tabela fato e várias tabelas de dimensão, facilitando consultas simples e diretas. Os dados armazenados no esquema em estrela são definidos como *desnormalizados* [2], o que significa que foram estruturados de acordo a necessidade.

  O modelo do esquema estrela ilustrado na Figura 5.5, possui 6 dimensões, contudo o número de dimensões deve variar conforme a complexidade dos dados, podendo chegar a dezenas. As dimensões descrevem as características, enquanto o fato quantifica a combinação dessas características.

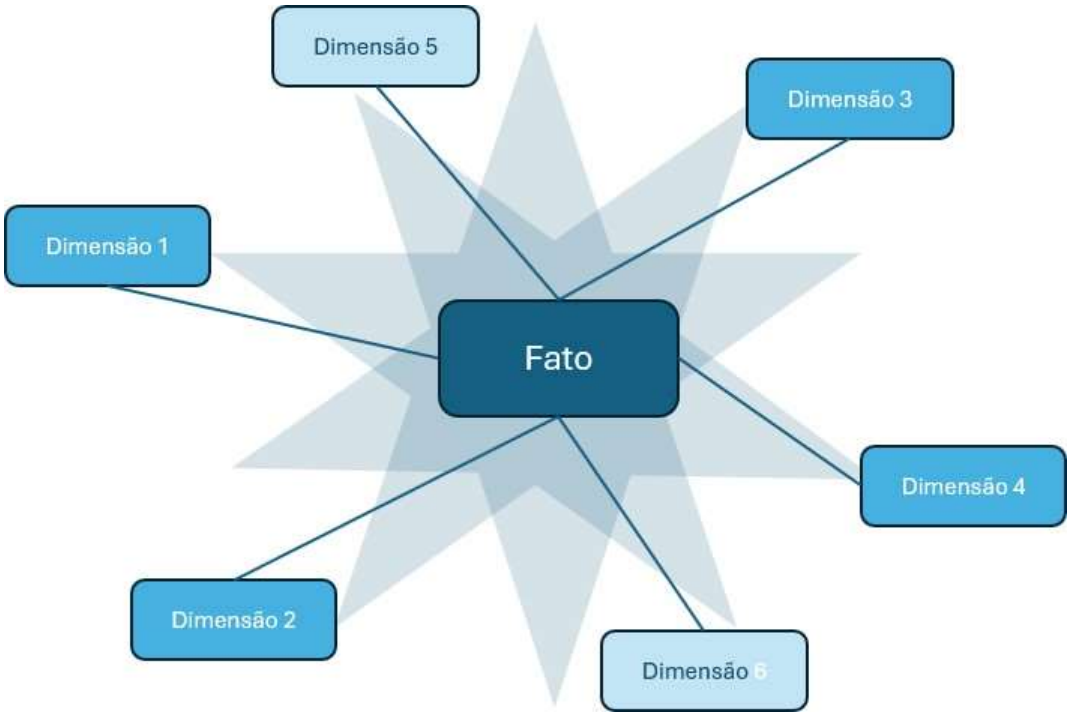

Figura 5.5: Representação do modelo Estrela

Esse modelo estrela, no nível conceitual, pode ser representado por meio do modelo entidade-relacionamento. A Figura 5.6 apresenta a representação parecida ao do modelo estrela da Figura 5.5 utilizando a simbologia do diagrama entidade-relacionamento da engenharia da informação.

O modelo entidade-relacionamento representado na Figura 5.6 é composto por cinco entidades. Ao centro, a representação do fato e as dimensões dimensões ligadas a ele. Neste tipo de reapresentação é também observável a identificação de cardinalidade, sendo de 1:n entre a dimensão e a fato.

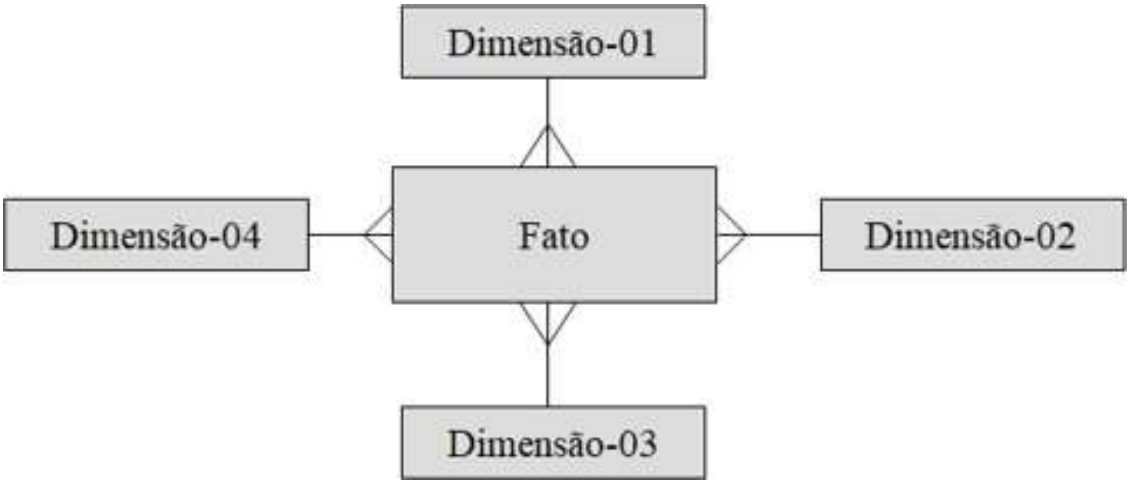

Figura 5.6: Representação do modelo estrela por meio de entidade-relacionamento

Ao mapearmos essas entidades para tabelas, a tabela de fatos será constituída por, no mínimo, quatro colunas, cada uma delas funcionando como chave estrangeira, referenciando a chave primária de uma dimensão. O conjunto dessas chaves estrangeiras compõe a chave composta da tabela de fatos. Dessa forma, a tabela de fatos é considerada dependente das dimensões, uma vez que seu conteúdo é contextualizado e quantificado por elas. As dimensões, por sua vez, são consideradas independentes, fornecendo os atributos descritivos para a análise dos dados. A Figura 5.7 apresenta o modelo lógico relacional resultante do mapeamento do modelo conceitual apresentado na figura Figura 5.6.

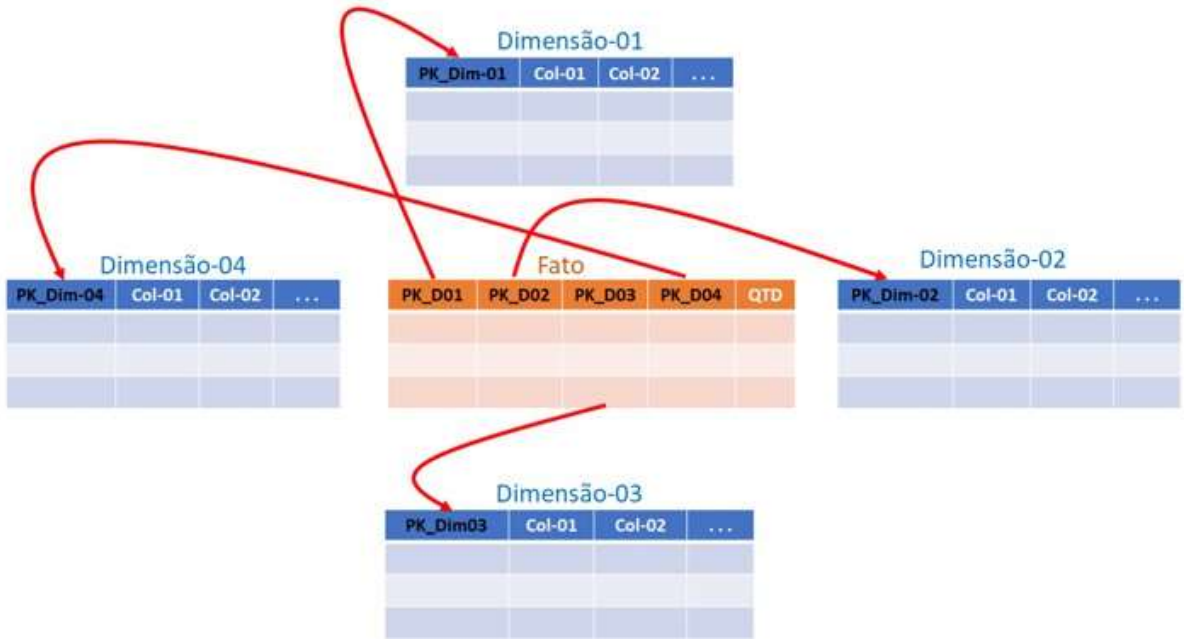

Figura 5.7: Representação do modelo estrela no nível lógico relacional

A Figura 5.7 ilustra um esquema estrela típico, composto por uma tabela de fatos denominada 'Fato' e quatro dimensões: Dimensão-01, Dimensão-02, Dimensão-03 e Dimensão-04. Cada dimensão possui uma chave primária (PK_Dim-01, PK_Dim-02, etc.) e atributos descritivos (Col-01, Col-02, ...).

A Fato, por sua vez, armazena as medidas (QTD) e as chaves estrangeiras que se referem às chaves primárias das dimensões. O conjunto dessas chaves estrangeiras forma a chave composta da tabela de fatos, estabelecendo um relacionamento um-para-muitos entre a tabela de fatos e cada dimensão.

- **Esquema Floco de Neve**: Variante do esquema estrela, onde as tabelas de dimensão são normalizadas, reduzindo redundâncias e aprimorando a organização dos dados. Figura 5.8.

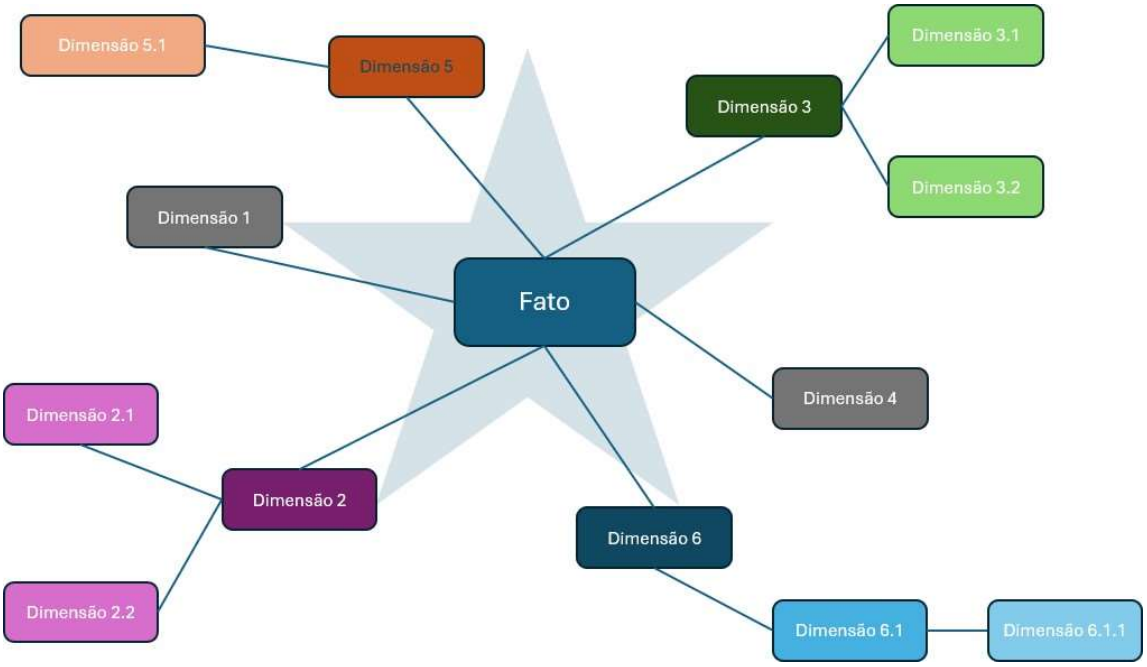

Figura 5.8: Representação do modelo Floco de Neve

- **Arquitetura ROLAP**: Relacional Online Analytical Processing (ROLAP). Implementa análise Online Analytical Processing (OLAP) utilizando bancos de dados relacionais.

Modelos Dimensionais em Bancos de Dados Multidimensionais

- **Cubos**: modelos dimensionais armazenados em bancos de dados multidimensionais. Trata-se de estrutura multidimensional que representa em cada dimensão um eixo, e os fatos são armazenados nas células do cubo. Figura 5.9.

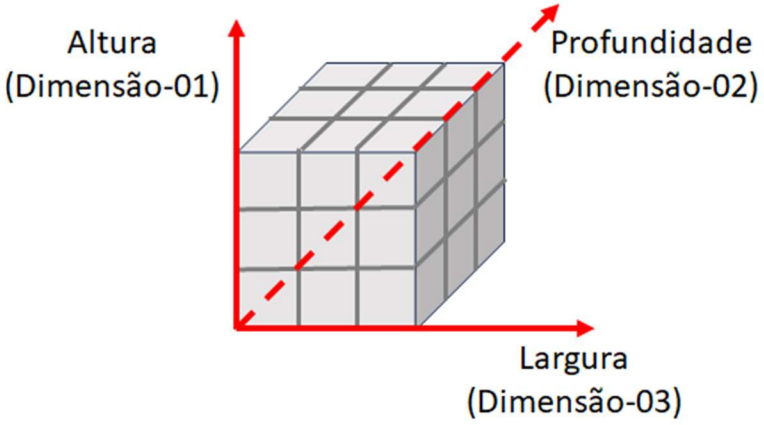

Figura 5.9: Representação do modelo Cubo

- **Hipercubos**: Extensão do conceito de cubo para mais de três dimensões.
- **Arquitetura MOLAP**: Multidimentional Online Analytical Processing (MOLAP) Baseada em bancos de dados multidimensionais, desempenho para consultas OLAP por meio de pré-processamento e otimização dos dados.

Arquitetura Híbrida

- **HOLAP**: Hybrid Online Analytical Processing (HOLAP). Combina as vantagens de ROLAP e MOLAP, oferecendo flexibilidade na organização relacional e desempenho otimizado.

Na Modelagem Dimensional, é preciso que se defina, de maneira precisa, a chamada granularidade da tabela dos fatos. Trata-se da determinação do nível de detalhe das informações. É preciso também detectar dimensões relevantes e seus respectivos atributos, assim forma-se o contexto para as medidas. As medidas são os valores numéricos que representam os fatos de negócio e são armazenados nas tabelas de fatos. As dimensões agrupam atributos textuais que descrevem os fatos, como data, produto, cliente e local.

Um ambiente DW pode conter ambos, estrela e cubo [2]. Contudo a modelagem estrela é suficiente para este DM.

5.4 Banco de Dados NoSQL

Os bancos de dados NoSQL adotam uma abordagem arquitetônica distinta em comparação com os bancos de dados convencionais [53], estes conhecidos como bancos de dados

relacionais. Uma das características marcantes é a flexibilidade de esquema, que permite o armazenamento de informações sem a necessidade de uma estrutura rigidamente definida.

Além disso, a arquitetura dos bancos NoSQL possibilita a adição de recursos de processamento aos servidores (*cluster*), operando de maneira ativa, caracterizando o que chamamos de escalabilidade horizontal. Isso se diferencia da escalabilidade vertical, que envolve a expansão dos recursos do servidor, como processador e memória, como é o caso do SQL Server, por exemplo.

Apesar de haver uma variedades de soluções NoSQL, ainda assim apresentam características em comum. Estas características podem ser descritas da seguinte maneira:

- Modelos de dados flexíveis, eventualmente nem possuem esquemas;
- Consistência eventual, que consiste no fato de que nem todos os nós do *cluster* possam estar sincronizados
- Desempenho otimizado tanto pela distribuição dos dados entre os nós, como também pela escalabilidade.
- Os dados replicados e particionados dentre os servidores.

Os bancos de dados NoSQL podem ser classificados em quatro categorias principais: documentos, chave-valor, família de colunas e grafos.

Aqueles baseados em documentos armazenam dados em formato JavaScript Object Notation (JSON), onde cada registro é um documento que é composto por pares de chave-valor organizados de forma hierárquica, proporcionando flexibilidade na estruturação dos dados.

Os bancos de dados do tipo chave-valor são os mais simples, consistindo apenas em pares de chaves e valores. Essa abordagem é eficiente para operações de leitura e gravação de dados.

Nas estruturas de famílias de colunas, os dados são armazenados em tabelas com colunas dinâmicas, permitindo uma modelagem mais flexível do que os bancos de dados relacionais. Isso é útil quando a estrutura dos dados não é predefinida.

Por fim, os bancos de dados do tipo grafos são ideais quando as relações entre as informações são cruciais, como em redes sociais, por exemplo.

Ainda em [53] as seguintes características são apontadas: não utilizam modelo relacional; tem boa execução em *cluster*; código aberto; não tem esquema.

5.4.1 MongoDB

Segundo a documentação [54], MongoDB é um Sistema de Gerenciamento de Banco de Dados não Relacional (DBMS), escrito em C e de código aberto, que utiliza documentos

flexíveis em vez de tabelas e linhas para processar e armazenar diferentes tipos de dados. Classificado como NoSQL, oferece modelo flexível de dados, o que permite a usuários armazenar e consultar diversos formatos, simplificando o gerenciamento e aumentando a escalabilidade. A estrutura de dados no MongoDB é composta por documentos, organizados em coleções. Esses documentos são formatados como BSON (Binário JSON).

Suas principais características são:

- Escalabilidade horizontal: pode ser distribuído em vários servidores.
- Esquema flexível: seus documentos não precisam de estrutura uniforme.
- Consultas ricas: diversos recursos de recuperação de registros.
- Alto desempenho: lida com grandes volumes.

Para implementar essa estrutura de maneira otimizada, seguindo as melhores práticas disponíveis no guia do [5], um *cluster* do MongoDB deve ser composto por três serviços:

- **SHARD (Fragmento)**: Cada nó do *cluster* deve ser responsável por uma parte dos dados, conhecida como *shard*, e é encarregado de lidar com consultas relacionadas a essa fração específica. Portanto, a tendência é de que quanto mais nós houver, melhor será a relação entre a quantidade de dados e servidores.
- **MONGOS**: Realiza o roteamento de consultas, servindo como uma interface entre os aplicativos clientes e o *cluster* fragmentado.
- **Servidor de Configuração**: Os servidores de configuração armazenam metadados e configurações definidas para o *cluster*.

A fragmentação dos dados no MongoDB ocorre através da eleição de um campo dos dados, a partir do qual é gerado um valor de *HASH* que é usado para distribuir os registros com base em faixas de valores.

Para ilustrar, se tivéssemos três nós e valores de *HASH* de 1 a 9, os valores de 1 a 3 seriam armazenados no primeiro nó, de 4 a 6 no segundo nó e de 7 a 9 no terceiro nó.

Cada nó é responsável por responder às consultas relacionadas aos registros que estão sob sua guarda.

5.5 Embasamentos Legais

5.5.1 Lei Geral de Proteção de Dados (LGPD)

A LGPD, Lei 13.709, instituiu uma nova fase no que tange tratamento de dados pessoais no país. Inspirada na General Data Protection Regulation (GDPR), da União Européia, ela

busca tentar assegurar a privacidade e a proteção de dados de indivíduos, estabelecendo diretrizes claras e rigorosas para o tratamento de informações pessoais por entidades públicas e privadas.

A este respeito ela consolida o conceito do que são dados pessoais e avança também daquilo que chama como "dados pessoais sensíveis". O primeiro sendo qualquer informação relacionada a uma pessoa natural identificada ou identificável, tanto o ambiente físico quanto o digital. O segundo inclui ainda informações relativas à origem racial ou étnica, convicções religiosas, opiniões políticas, filiação sindical, saúde, vida sexual e dados genéticos ou biométricos. Tendo este segundo uma proteção especial para mitigar questões discriminatórias.

Esta questão é baseada em respeito à privacidade, a autodeterminação informativa, a inviolabilidade da intimidade, a honra e a imagem, e o desenvolvimento econômico e tecnológico sustentável.

Este assunto é trazido aqui dado ao seu enfoque que enfatiza a promoção da transparência e a segurança jurídica, pois permite que cidadãos saibam o que terceiros sabem a seu respeito, o que coletaram, processaram e armazenaram obrigando a estes últimos possuir mecanismos de controle e auditoria sobre essas práticas.

A LGPD também estabelece princípios que o tratamento de dados devem contemplar quais sejam finalidade, necessidade, livre acesso, qualidade dos dados, segurança, prevenção, não discriminação e prestação de contas.

Uma vez que este trabalho lida com dados pessoais, sobretudo sobre a guarda do estado, fica vinculada a necessidade de se tratar a respeito de proteção de dados. A obra de [55] aponta a privacidade como desafio capital no tratamento e coleta de dados. Seu foco é abordar o desafio de coletar dados (heterogêneos) de diversas fontes, centralizando-as em um repositório central. Os autores adotam uma camada de *wrapper* a qual anonimizava os dados na coleta, mitigando o risco de vazamentos. A questão de privacidade deste trabalho possui características diferentes. Embora seja necessário coadunar com a LGPD, os dados aqui não saem da área de proteção da universidade, ou sequer sofrem transporte entre bancos.

O trabalho de [8] apresenta a figura do *Data Protection Officer (DPO)*, análogo ao *Encarregado* de nossa LGPD, na *General Data Protection Regulation (GDPR)* da União Europeia. O autor faz o comparativo entre os dois e detalha responsabilidades do Encarregado em nossa pátria. Em ambas as legislações, europeia e nacional, a privacidade é condição primaz de trabalho. Os dados tratados aqui são, na maioria dos casos de ordem pública, no entanto há registros de natureza restrita, e diversos outros sigilosos. Entretanto, serão atingidos somente metadados, ou seja, não será dada luz a conteúdo de documentos propriamente dito. E ainda, os usuários para os quais é destinado, são

aqueles que já tem acesso a todo repositório do SEI. Não cabendo, pois, julgamento e/ou atribuição de permissões.

5.5.2 Lei de Acesso à Informação (LAI)

Um segundo regimento de destaque é a Lei de Acesso à Informação (LAI) (Lei 12.527) 5.5.2 . Esta trata de prazos para o atendimento das solicitações de informação são bem definidos, garantindo rapidez e eficiência no processo de transparência.

Esta lei foi um avanço no que tange transparência e acesso à informações públicas no país. Constitui-se importante ferramenta para o cidadão a obter informações pertinentes bem como deu novo prisma a fiscalização e combate a corrução. Ela permite e fomenta iniciativas de dados abertos e ferramentas para facilitar o acesso a mesmas, tais como o Portal da Transparência.

Estabelece que o órgão público tem até 20 dias corridos (a partir do dia da solicitação) para responder a pedidos de informação, sendo prorrogável por mais 10 dias, se justificado. Se o caso de pedido for negado, o solicitante pode interpor recurso, cabendo ao órgão 5 dias para responder. Se ainda assim negado, pode o requisitante recorrer à CGU. A exceção a estes prazos pode ser justificada pela complexidade, seja de coleta ou tratamento.

Negativas de acesso só são admissíveis se há imposição de sigilo. E ainda assim, neste caso, o solicitante pode pedir a revisão deste sigilo. Normalmente este tipo de sigilo são impostos a informações que envolvem riscos à segurança do Estado, à integridade física de pessoas ou ao sigilo comercial.

Este regimento deve ser observado por todos os órgãos da administração direta dos poderes Executivo, Legislativo e Judiciário, incluindo tribunais de contas, Ministério Público e autarquias, empresas públicas e entidades que recebem recursos públicos. Inclusive quaisquer entidade privada que execute serviços públicos ou que se beneficiam de verbas públicas.

Capítulo 6

Estudo do Caso

Dando continuidade à análise do problema, a seção 6.1 descreve o papel da STI. Em seguida, a seção 6.2 explora a arquitetura do SEI, enquanto a seção 6.3 aborda o entendimento detalhado do problema. Por fim, a proposta de solução é discutida na seção 6.4 e como foi a carga dos dados em 6.5.

6.1 Secretaria de Tecnologia da Informação (STI)

A Secretaria de Tecnologia da Informação desempenha um papel fundamental na Universidade de Brasília, atuando como a provedora de serviços telemáticos para toda a instituição. Em colaboração com a Arquivo Central, a qual é encarregada da guarda arquivística, realizaram conjuntamente em 2015 a implementação deste sistema na Universidade, seguindo o exemplo de outras entidades do âmbito federal [56].

> "O SEI, desenvolvido pelo Tribunal Regional Federal - 4ª Região (TRF4), é uma plataforma que engloba um conjunto de módulos e funcionalidades que promovem a eficiência administrativa. Trata-se também de um sistema de gestão de processos e documentos eletrônicos, com interface amigável e práticas inovadoras de trabalho, tendo como principais características a libertação do paradigma do papel como suporte físico para documentos institucionais e o compartilhamento do conhecimento com atualização e comunicação de novos eventos em tempo real" [57].

Desde então, tem sido dedicado um contínuo esforço para garantir o adequado funcionamento do sistema. Isso envolve tanto a rápida resposta a demandas imprevistas dos usuários quanto a garantia da sólida performance do sistema, assegurando sua robustez e confiabilidade.

6.2 Arquitetura do Sistemas Eletrônico de Informação (SEI)

Como um repositório de documentos públicos, os quais podem ter uma vida útil indeterminada, conforme previsto na lei [58], o SEI assume a responsabilidade primordial de guardar e preservar permanentemente seu conteúdo e informações arquivísticas. Em um estudo realizado [59], é ressaltada a importância não apenas de armazenar tais informações, mas também de mantê-las acessíveis, a fim de atender às demandas, como aquelas definidas pela Lei de Acesso à Informação (LAI) [7], que estabelece prazos legais para o fornecimento de informações a cidadãos mediante solicitações.

Com esse propósito, o sistema foi dotado de recursos que possibilitam o registro de eventos e interações dos usuários com os documentos e a plataforma.

Além disso, um detalhe significativo a ser mencionado é que os documentos em si não são armazenados diretamente no banco de dados. Os arquivos anexados e os conteúdos textuais dos documentos em formato HTML são mantidos em um repositório separado, na forma de arquivos físicos. Para possibilitar consultas eficientes por parte dos usuários, esses documentos são indexados por meio do banco de dados textual *SolR* [60]. É relevante destacar que esses aspectos específicos não serão abordados no escopo deste trabalho.

6.3 Entendimento do Problema

6.3.1 Estruturas NoSQL

Considerando as características delineadas anteriormente, uma abordagem alternativa que foi considerada foi a transferência dos dados para uma ferramenta externa ao banco de dados do SEI, estrutura NoSQL, como MonboDB, Cassandra ou Spark. A questão central teria sido avaliar se essa abordagem se revela vantajosa para a situação em pauta.

É importante destacar que, independentemente da ferramenta escolhida para análise e teste, a etapa mais significativa do trabalho envolve o processo de Extraction, Transformation and Loading (ETL). Esse esforço é inevitável e comum a todas as opções avaliadas. O Capítulo 7.1.3 apresenta o resultado de uma tentativa de realizar esse processo sem o devido tratamento dos dados.

Nesse contexto, o escopo deste projeto foi delimitado para abranger especificamente o processo de ETL, deixando as avaliações de desempenho a serem conduzidas no próprio Sistema Gerenciador de Banco de Dados (SGBD) utilizado, neste caso, o SQL Server. Consequentemente, como neste caso, as consultas realizadas dentro da estrutura proposta

atingiram resultados satisfatórios em termos de desempenho, não foram realizados testes em outros ambientes.

Além disso, é importante salientar que manter a solução dentro da mesma estrutura possui um relevante requisito funcional, visto que isso não apenas simplifica a implantação, mas também facilita a replicação da solução em outras unidades que utilizem o SEI e estejam interessadas em adotá-la, haja visto que não é preciso adquirir e configurar novos equipamentos.

A expectativa subjacente é que esse processo de ETL seja capaz de resolver as questões de normalização dos dados, proporcionando, assim, um desempenho adequado no sistema.

6.3.2 Outras Iniciativas

No ano de 2020, uma tentativa foi empreendida com o objetivo de abordar a questão de desempenho através da implementação de práticas que pudessem aprimorar o rendimento das consultas SQL Server. A técnica selecionada para essa finalidade foi o armazenamento em modo *row-based* [12]. Essa técnica envolve a fragmentação física do repositório da tabela em arquivos, com base em um critério específico. No contexto dessa ação, a escolha recaiu sobre a data do evento, resultando na criação de arquivos individuais para cada ano. A expectativa subjacente era que essa abordagem melhorasse o desempenho, já que, ao executar uma consulta para o ano de 2018, o banco de dados lidaria exclusivamente com os registros desse ano. Contudo, essa iniciativa demonstrou ser ineficaz, apresentando um impacto limitado no tempo de resposta.

As manobras empreendidas são responsáveis pelas anomalias evidenciadas nos gráficos 1.2 e 1.3, os quais estão apresentados na Seção 1, estas, por sua vez, causaram uma alocação de espaço excedente que foram em seguida ajustados.

Em uma segunda tentativa, optou-se por separar os registros da tabela infra_auditoria com mais de 12 meses de idade. Essa separação é automatizada por meio de um agendamento que realiza a transferência dos dados a cada mês.

Assim, a mesma consulta apresentada na Seção 2, realizada pela interface do usuário, foi restringida ao último ano. Caso seja necessário acessar dados de anos anteriores, é preciso recorrer à STI, a qual realiza essa tarefa diretamente no console do banco de dados. Mesmo assim, ainda persistem consultas que demandam vários minutos para serem executadas.

A suposição é que essa problemática esteja relacionada ao volume de dados. De um lado, a quantidade substancial de registros (superior a 300 milhões) e, de outro, o considerável número de informações contidas em cada registro, amplificado pelo tamanho, especialmente em campos do tipo *varchar(max)*.

Os impactos dessa segunda abordagem também se manifestam nas anomalias de evolução de crescimento de disco observadas nas Figuras 1.7 e 1.8 da Seção 1.

6.3.3 Colunas *operacao* e *requisicao*

Em relação à coluna *operacao*, três aspectos de relevância merecem destaque.

No primeiro aspecto, na primeira linha do texto gravado da coluna operação, assinala-se o início do registro, onde é atribuído o tipo de ocorrência em questão. Inicialmente, foram identificados 149 tipos distintos, tais como *AcessoRN listarConectado*, *UsuarioRN gerarSenhaControlado* e *DocumentoRN gerarPdfConectado*. O procedimento inicial consistiu em percorrer toda a tabela infra_auditoria, registrando a incidência de cada um desses atributos, a fim de identificar os possíveis tipos de acesso. Esse processo define os "Tipos de Operação".

No segundo aspecto, na linha subsequente, é indicado o nome do objeto ao qual a informação está relacionada. Nas Figuras 1.10, 1.11, 1.12 e 1.13, nelas, o objeto de origem é nomeado como *AcessoDTO*. Vale recordar que a função *formatarDados* (conforme listado na Seção 1.2) recebe um objeto como argumento, percorre-o para coletar todos os atributos e seus valores, e então retorna um texto que é armazenado na coluna *operacao*.

Por fim, a terceira informação disponível é a associação entre atributo e valor. À esquerda, encontra-se o nome do atributo do objeto, e à direita, o seu respectivo valor. Em casos onde se trata de um "array", o nome do atributo é sucedido por "IN ", e os respectivos valores para cada item do "array"são enumerados logo a seguir.

No que concerne à coluna *requisicao*, conforme ilustrado na Figura 1.9 (Seção 1.1), há uma distinção sutil na formatação, sendo identificadas apenas duas formas.

A primeira delas denota o início e o término da ocorrência. O início é indicado por "GET - Array"e/ou "POST - Array", dependendo do objeto HTTP de onde provém a informação. Em ambos os casos, a linha subsequente contém um sinal de parêntese aberto - "(". O término é indicado por um parêntese fechado - ")".

Em seguida, de forma análoga à coluna *operacao*, a coluna *requisicao* apresenta uma combinação de chaves e valores. A diferença reside na abordagem em relação a "arrays". Nesse contexto, apenas um valor singular é fornecido para cada atributo. Portanto, quando o evento se trata de listagem de documentos, os registros na coluna *requisicao* não fornecem informações pertinentes. Nesse caso, a *operacao* é a coluna relevante para esclarecimentos. A respeito disso, o atributo chave é *acao*. Uma análise preliminar identificou 283 tipos distintos de ações. Por conseguinte, todas as entradas das colunas serão percorridas com a finalidade de enumerar essa classificação de ações.

6.4 Proposta

A primeira etapa envolveu percorrer todos os registros, examinando cada um deles em detalhes para identificar, nesse primeiro escaneamento, os atributos pertencentes aos objetos PHP. Esses atributos, cujos nomes são essenciais para uma classificação subsequente, são de grande utilidade para a fase seguinte do processo.

Uma segunda passagem tem como foco a leitura dos valores. Dessa vez, o objetivo principal foi separar os dados, organizando-os cuidadosamente de acordo com as categorias definidas pelos atributos identificados na etapa anterior.

Na sequência, esses dados foram reestruturados e enriquecidos por meio da integração com informações de outras tabelas que possuam relevância para o projeto. Essa integração resulta na construção de um repositório normalizado, utilizando uma estrutura em formato estrela, a qual desempenha o papel de um DW, destinado a atender às consultas analíticas com eficácia.

6.4.1 Modelo Conceitual

Para a realização deste projeto, foi concebida uma estrutura de dados dimensional com a expectativa de alcançar melhorias substanciais no tempo de resposta das consultas. A extração, transformação e carregamento (ETL) dos dados foram efetuados a partir da tabela *infra_auditoria*.

O processo de carregamento foi dividido em duas etapas. A primeira fase consistiu na identificação das operações e seus respetivos atributos, conforme já delineado. Em seguida, na segunda iteração, os dados foram adquiridos e processados, sendo categorizados e armazenados de acordo com a estrutura proposta. A imagem 6.1 ilustra o processo, no qual toma-se uma base monolítica que, através de técnicas de ETL é gerado um Modelo Estrela, a partir do qual consulta e auditorias podem ser realizadas, inclusive através de painéis.

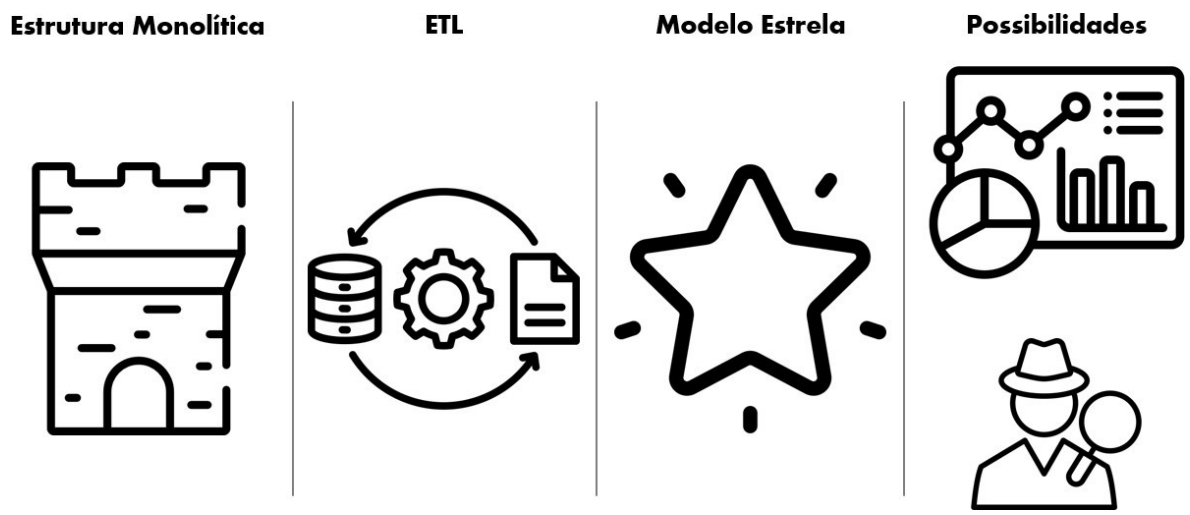

Figura 6.1: Processo do modelo Conceitual Preliminar - Composição de imagens de [4]

O modelo conceitual preliminar é retratado na Figura 6.2. Nessa representação, a tabela fato fato_Auditoria é central, cercada pelas dimensões dim_Usuarios, dim_Operacao, dim_Tempo e dim_Unidades, as quais constituem os elementos fundamentais desse ambiente dimensional.

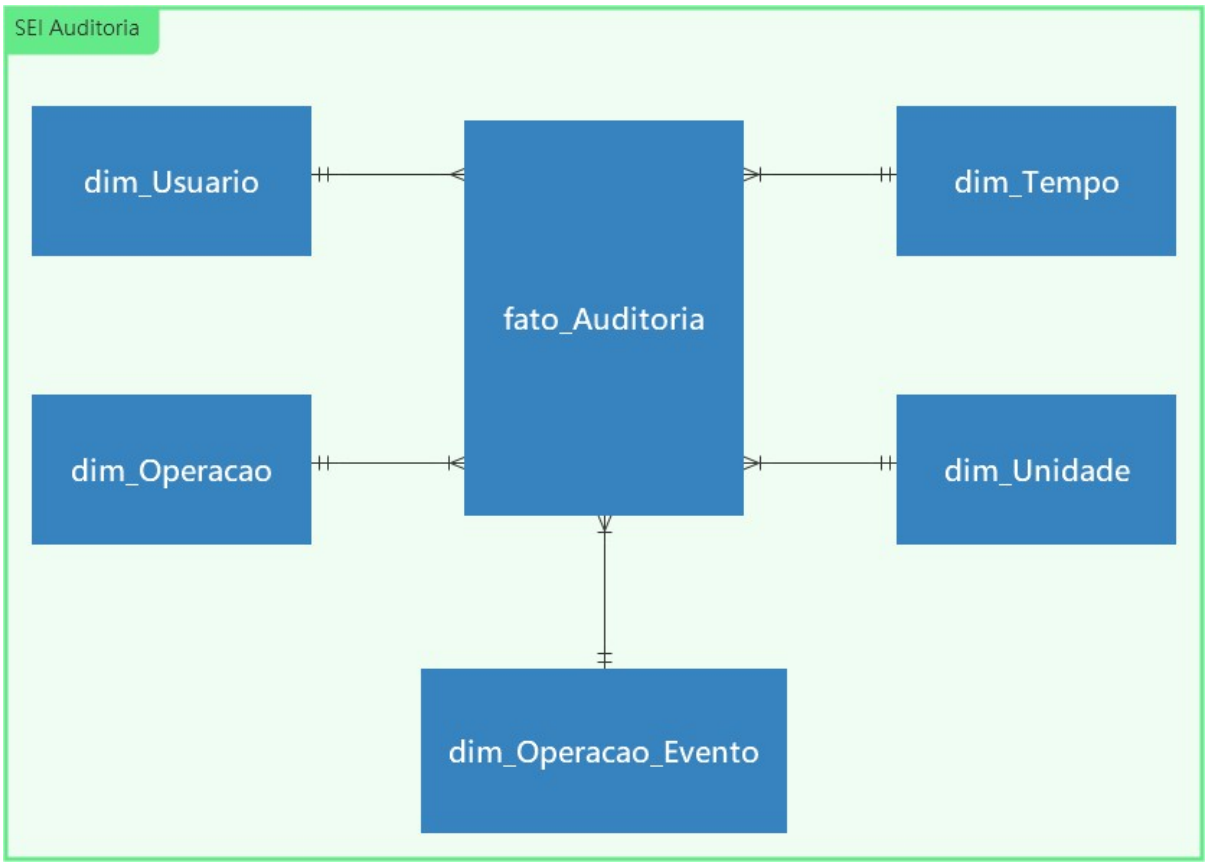

Figura 6.2: Modelo Conceitual Preliminar Proposto.

A tabela *fato_Auditoria* serve como repositório para as informações extraídas da tabela *infra_auditoria*. Paralelamente, a *dim_Operacao* contem os distintos tipos de operações identificados a partir dos códigos executados, como anteriormente mencionado.

As dimensões dim_Usuario e dim_Unidade desempenharão a função de concentrar as identificações dos usuários e unidades que interagem no SEI. Elas irão buscar informações pertinentes diretamente do banco de dados do sistema ou de outros repositórios que possam vir a ser identificados.

Cada categoria de operação, por sua vez, está associada a um conjunto variável e diversificado de atributos. Portanto, a dimensão dim_Operacao_Evento é encarregada de registrar, para cada um desses atributos, os respectivos valores correlacionados na tabela fato_Auditoria.

6.4.2 Bastidores

Considerando o processamento dos dados realizado em etapas, foi desenvolvida uma estrutura intermediária para os dados (o modelo físico é apresentado na figura 6.3). Essa

abordagem segue a estratégia conhecida como *staging area* [11].

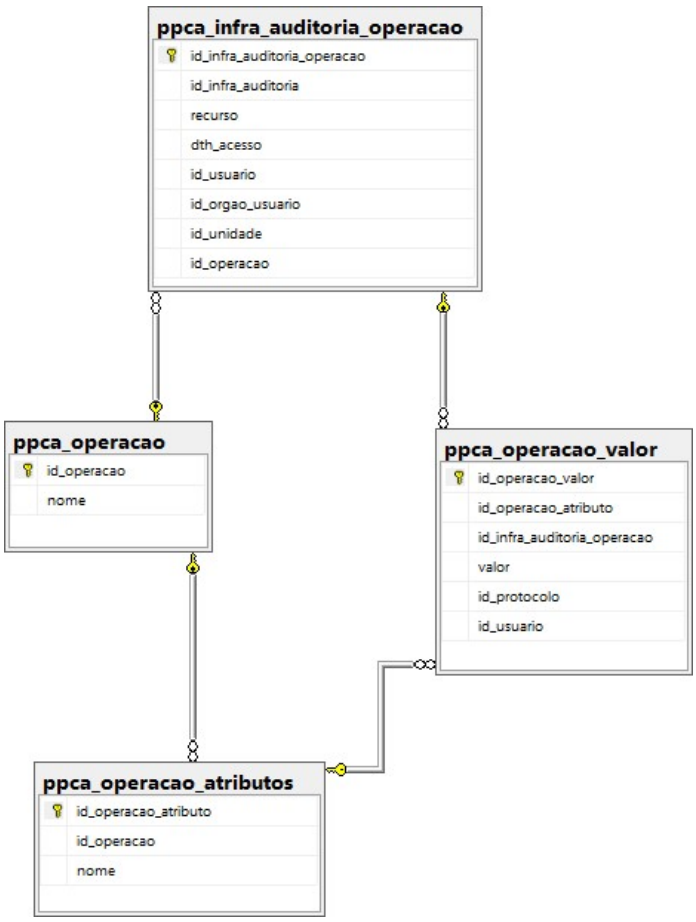

Figura 6.3: Modelo Físico - Bastidores

Um código em linguagem SQL percorre as ocorrências dessa coluna, identificando os tipos de objetos e seus atributos correspondentes. Posteriormente, esses dados são segregados e organizados em tabelas e colunas apropriadas.

A seguir a descrição das entidades representadas na figura 6.3. O prefixo encontrado nelas (ppca) serve meramente para distinguir os objetos tratados neste trabalho daqueles originais do repositório.

Tabela ppca_infra_auditoria_operacao

Esta tabela mantém uma relação de correlação 1:1 (um para um) com a tabela *infra_auditoria* (conforme figura 1.4, Seção 1.1). Em outras palavras, para cada registro na tabela *infra_auditoria*, há um correspondente nessa nova tabela. No entanto, a nova tabela foi projetada para excluir informações de pouca relevância para o escopo deste

trabalho, tais como endereços IP, siglas e descrições. O intuito é concentrar-se nos principais eventos (interações) relacionados aos documentos. Essas informações podem ser reintroduzidas posteriormente, quando formos estabelecer o repositório definitivo (DW).

Tabela ppca_operacao

Nesta tabela, abordamos a identificação das classes encontradas no código PHP, cujos objetos foram convertidos em texto e posteriormente armazenados no banco de dados. Durante o desenvolvimento, foram identificadas 140 operações distintas.

Tabela ppca_operacao_atributos

Continuando a análise desses objetos mencionados anteriormente, os seus atributos são registrados nesta tabela após terem sido identificados. No ambiente de desenvolvimento, um total de 61.726.073 atributos foram registrados.

Tabela ppca_operacao_valor

Por fim, concentramos-nos na informação em si. Cada valor encontrado no texto é identificado, separado e, em seguida, armazenado. Esse processo preserva a organização, o registro e a classificação dos valores de maneira estruturada.

6.5 Carga de dados

Após a elaboração e teste dos /emphscripts em PL/SQL, iniciou-se o processo de importação, cuja apenas a execução levou, ininterruptamente, aproximadamente 15 dias (14 dias, 19 horas e 55 minutos). Durante essa etapa, foram processadas 289.572.382 registros operações da tabela infra_auditoria, resultando na geração de 1.516.540.638 registros de operação armazenados na nova estrutura.

Esses registros estão classificados em, pelo menos, uma dos 243 tipos de operação, que possuem somados 23.581 atributos.

Capítulo 7

Comparativos

Seria possível resolver essa questão de outra forma? Existiria uma solução ou ferramenta já disponível capaz de mitigar a lentidão observada? Bancos de dados NoSQL são conhecidos por oferecer soluções simples para problemas complexos, como o presente caso. Para investigar essa possibilidade, realizamos testes utilizando uma das principais ferramentas desse segmento. Este capítulo apresenta o relato detalhado desses experimentos.

7.1 Comparativo

A questão surgiu sobre se um banco de dados NoSQL poderia oferecer um tempo de resposta melhor do que o SQL Server, respondendo a consultas sem a necessidade de tratamento de dados além da simples importação.

Assim, foi realizado um processo de seleção para encontrar o banco NoSQL mais adequado para essa questão. Entre os quatro tipos mencionados na Seção 5.5.2, aquele que melhor se adequou foi o banco de dados baseado em documentos. Isso se deve ao fato de que esse tipo de banco não requer uma definição prévia de dados.

A coluna de dados chamada "operacao", descrita na Seção 6.5, é do tipo "varchar(max)"em sua estrutura original, ou seja, é uma coluna de texto que não possui um padrão de dados estabelecido.

O MongoDB foi escolhido como o banco de dados ideal para este caso, devido à sua notoriedade significativa entre os bancos NoSQL e à disponibilidade de uma documentação abrangente e acessível.

7.1.1 Ambiente de testes

A STI providenciou o ambiente necessário para conduzir os testes, em resposta a uma solicitação justificada, devido ao interesse na análise.

Foram disponibilizadas três máquinas virtuais, cada uma executando o sistema operacional Debian 11, com dez processadores virtuais, trinta e dois gigabytes de memória RAM e dois terabytes de espaço em disco. Essas máquinas foram interconectadas em rede, permitindo assim a configuração de um cluster MongoDB entre elas.

A versão do MongoDB instalada foi a *Community Server* 7.0.

No entanto, devido às características dos dados e das consultas realizadas, o mecanismo de *Shard*, que é um dos principais fatores para o ganho de desempenho no MongoDB, não pôde ser efetivamente utilizado. Isso ocorreu porque esse mecanismo requer que o campo a ser consultado seja classificável, para que possa ser organizado e dividido entre os nós. No entanto, o campo onde as consultas foram realizadas é do tipo texto, e as consultas envolveram a busca por trechos de texto nesse campo. Portanto, não foi possível segmentar um trecho de texto em um índice que abrangesse todo o campo, em vez de trechos específicos. Como resultado, as consultas foram realizadas de forma paralela, mas não distribuída.

7.1.2 MongoDB

Em cada um dos nós, o serviço de Shard foi instalado, conforme ilustrado na Figura 7.1, seguindo uma arquitetura semelhante àquela proposta na documentação disponível em [5].

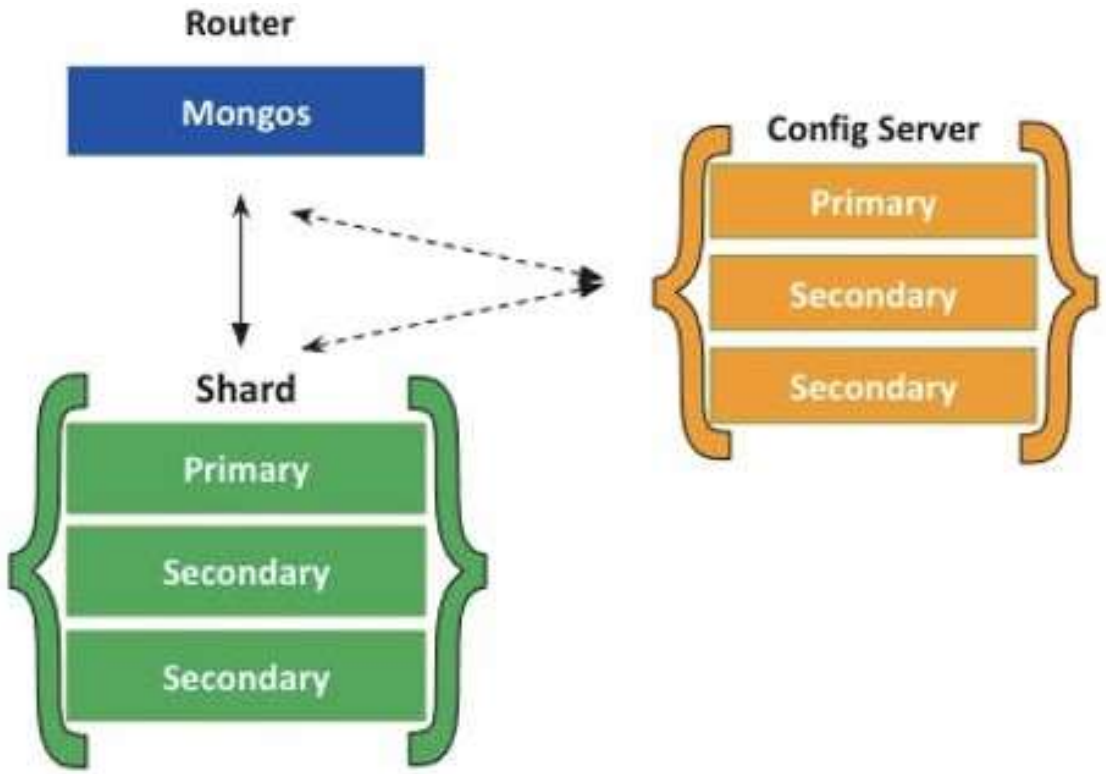

Figura 7.1: Arquitetura MongoDB[5]

Dessa forma, em todos os nós, os serviços de *Shard* e *Config Server* foram instalados, com um deles desempenhando também a função de *Router*.

Em seguida, os dados foram exportados do banco SQL Server no formato *Comma Separated Values (CSV)*, separados por ano devido ao tamanho considerável. Esses arquivos foram então transferidos para um dos nós do cluster e importados. No total, foram importados 289.572.382 registros. Não foi necessário criar a estrutura previamente, pois o banco a construiu com base na definição encontrada nos arquivos. O processo de importação e exportação levou aproximadamente 24 horas.

7.1.3 Resultados

A dificuldade de realizar consultas nos registros do SEI se torna mais evidente quando é necessário recuperar dois trechos de texto (*substring*) distintos no campo "operacao"da tabela "infra_auditoria".

Para abordar essa dificuldade, selecionamos dois trechos de texto presentes nos dados e elaboramos uma consulta de baixa complexidade. Em linguagem natural, essa consulta equivaleria a: "**Na tabela infra_auditoria, encontre os registros em que o**

campo operação contenha o trecho de texto 'Visualiza' e o trecho de texto 'Id_usuario=100009063'".

No banco de dados de produção do SEI (SQL Server), essa consulta tem uma duração aproximada de 45 minutos e retorna um total de 784 registros.

A expressão de consulta utilizada no MongoDB pode ser visualizada na Figura 7.2.

```
db.infra_auditoria.find(
    {$and: [
    {operacao:{$regex:'Visualiza'}},
    {operacao:{$regex:'IdUsuario = 100009063'}}]})
```

Figura 7.2: Expressão de consulta utilizada no MongoDB

Após aproximadamente noventa minutos (1 hora e 30 minutos e 98 segundos), a consulta foi concluída sem retornar registros.

Em uma segunda tentativa, foi realizada a indexação textual da coluna "operacao". No entanto, após quatorze dias de execução, a indexação precisou ser interrompida e não foi possível realizar os testes.

Também é importante registrar que um grande número de registros no campo utilizado para os testes possui um formato semelhante a um campo JSON. É bastante plausível que os resultados pudessem ser significativamente melhores se houvesse um tratamento desses dados (ETL), assim com o proposto neste trabalho.

Capítulo 8

Resultados

Conforme descrito na Seção 6.4 e como resultado das atividades abordadas em 6.4.2, chegou-se à estrutura final do banco de dados, detalhada na Seção 8.1. Após a realização das cargas de dados (ETL), foram realizados testes de desempenho com tempos de execução das consultas, cujos resultados são apresentados em 8.2. Diante do bom desempenho, e ainda como prova de conceito, foram desenvolvidos *dashboards*, que estão apresentados na Seção 8.3. Por fim, abre-se um novo campo de possibilidades, algumas das quais são discutidas em 8.4.

8.1 Estrutura Definitiva

O modelo físico final, implementado e carregado está ilustrado na Figura 8.1

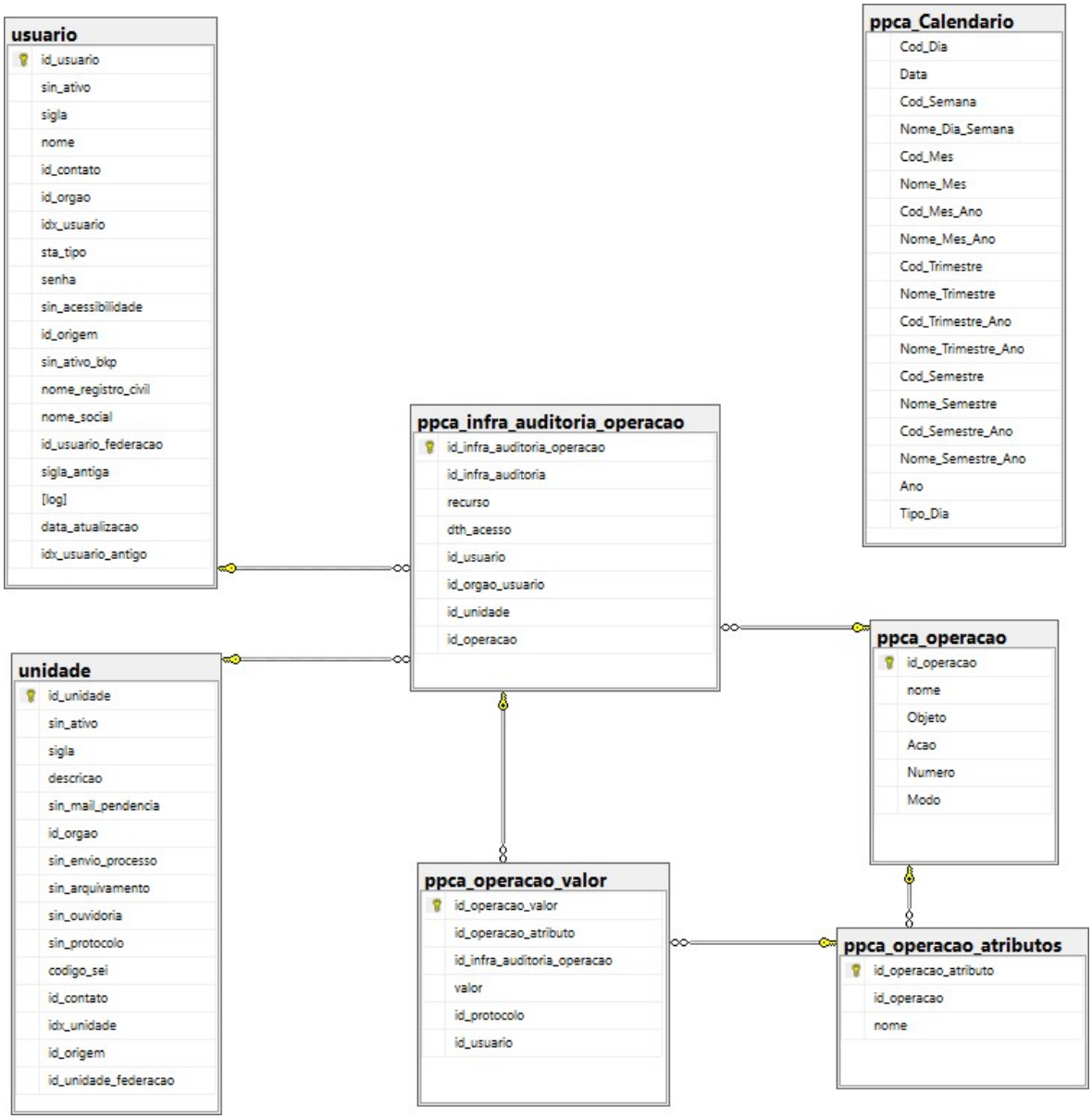

Figura 8.1: Modelo Físico Final

A tabela *ppca_infra_auditoria_operacao* desempenha um papel central, sendo assim a tabela fato desta solução. Ela possui uma relação de cardinalidade 1:1 com a tabela original *infra_auditoria*.

A tabela *ppca_operacao* é o resultado das operações identificadas no processamento da tabela *infra_auditoria*. A partir dessas operações, também foram identificados os respectivos atributos (armazenados na tabela *ppca_operacao_atributos*). Por exemplo, uma operação como uma leitura possui diversos atributos, como data e usuário. Os valores correspondentes desses atributos são armazenados na tabela *ppca_operacao_valor*.

Além disso, a relação entre as tabelas *ppca_operacao_valor* e *ppca_infra_auditoria_operacao* foi adicionada para facilitar consultas por valores que não demandem outros atributos. Um exemplo disso seria a consulta de todos os registros ocorridos em uma determinada data.

As tabelas *unidade* e *usuário* são fornecidas pelo próprio SEI, contendo os respectivos dados de unidades e usuários.

A tabela *ppca_Calendario* não possui um relacionamento explícito com o modelo, mas é essencial para navegação em *dashboards*, pois permite ao usuário final realizar pesquisas por datas, com opções de intervalos por dia, semana, etc. As caixas de pesquisa são preenchidas com os dados dessa tabela, de onde o usuário define o intervalo de tempo a ser filtrado, relacionado à coluna *dth_acesso* da tabela *ppca_infra_auditoria_operacao*.

8.1.1 Quantitativos

A Tabela 8.1 a seguir apresenta a quantidade de registros após o processamento e a carga de todos os dados.

A tabela *ppca_infra_auditoria_operacao*, que possui uma relação de 1:1 com a tabela *infra_auditoria*, contém aproximadamente 270 mil registros. As operações identificadas (armazenadas na tabela *ppca_operacoes*) totalizam 228 registros, enquanto seus respectivos atributos, na tabela *ppca_operacao_atributos*, somam pouco mais de 23 mil registros. O destaque vai para a tabela *ppca_operacao_valor*, que contém mais de 1,5 bilhões de registros.

Tabela	Quantidade
ppca_infra_auditoria_operacao	269.882.453
ppca_operacao	228
ppca_operacao_atributos	23.571
ppca_operacao_valor	1.516.540.625
Usuario	182.485
Unidade	1.741

Tabela 8.1: Quantitativo de registros após processamento dos dados.

8.2 Testes de Desempenho

Foi selecionado aleatoriamente um registro da tabela, cujo conteúdo na coluna "operacao"está exemplificado na Fig. 8.2. A partir deste registro, foi elaborado uma consulta

que filtra o equivalente, em linguagem natural, a "Quais documentos foram visualizados pelo usuário 100009063".

```
operacao
-----------------------------------------
AuditoriaProtocoloRN::auditarVisualizacao(
 AuditoriaProtocoloDTO:
Recurso = procedimento_visualizar
IdUsuario = 100009063
IdProtocolo = 309192
IdAnexo = [null]
Auditoria = 23/11/2016
Versao = [null]
IdAuditoriaProtocolo = 6204163)
```

Figura 8.2: Registro selecionado aleatoriamente

A sintaxe SQL utilizada para realizar esta consulta na tabela original (infra_auditoria) pode ser visualizada na Figura 8.3.

```
select *
from infra_auditoria
where operacao like '%auditarVisualizacao%IdUsuario = 100009063%'
```

Figura 8.3: Expressão inicial da consulta em SQL

Durante a execução das consultas, foi observada uma variação significativa nos tempos de resposta. Essa diferença pode ser explicada por fatores relacionados ao acesso às informações pelo SGBD, influenciados por aspectos de hardware. Na primeira execução, por exemplo, os registros são lidos diretamente do disco e armazenados na memória RAM pelo SGBD; nas execuções subsequentes, essa leitura ocorre diretamente da RAM, resultando em tempos consideravelmente mais rápidos. Além disso, o tempo de resposta pode ser afetado pela concorrência por recursos de hardware devido a outras consultas ou atividades em execução no banco.

Para mitigar essa variabilidade, a consulta foi repetida mil vezes, uma após a outra. Essa repetição ajuda a reduzir a influência de eventos esporádicos no mesmo servidor que poderiam distorcer os resultados. Os tempos de execução foram registrados e analisados com Python e o pacote de análise de dados Pandas, gerando as Figuras 8.4 e 8.6.

O tempo médio da execução destas consultas, sob as condições originais do banco de dados, foi de 16 minutos e 12 segundos, equivalente a uma média de 972,5 segundos, e resultou na obtenção de 784 registros, conforme evidenciado na Figura 8.4.

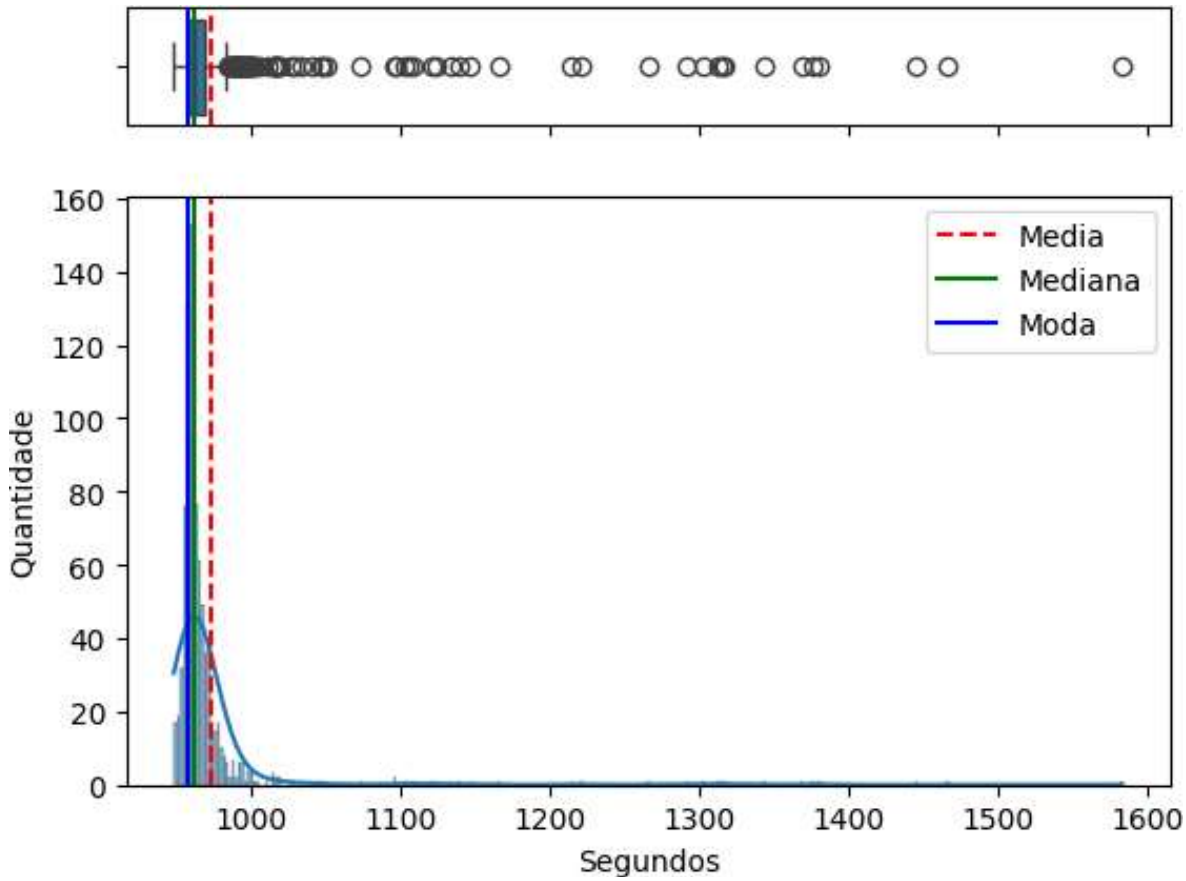

Figura 8.4: Distribuição da tomada de tempo de duração das execuções da estrutura original.

Para melhorar a visualização da distribuição dos resultados, a Figura 8.5 apresenta uma representação gráfica da distribuição após a remoção dos 50 maiores valores de tempo (5%). Com essa filtragem, observa-se uma distribuição "normal". A média representa o ponto de equilíbrio dos dados, a mediana divide a distribuição ao meio, e a moda indica o valor mais frequente. Como esses três valores não coincidem, verifica-se uma assimetria, e como a calda está a direita, é denominada positiva.

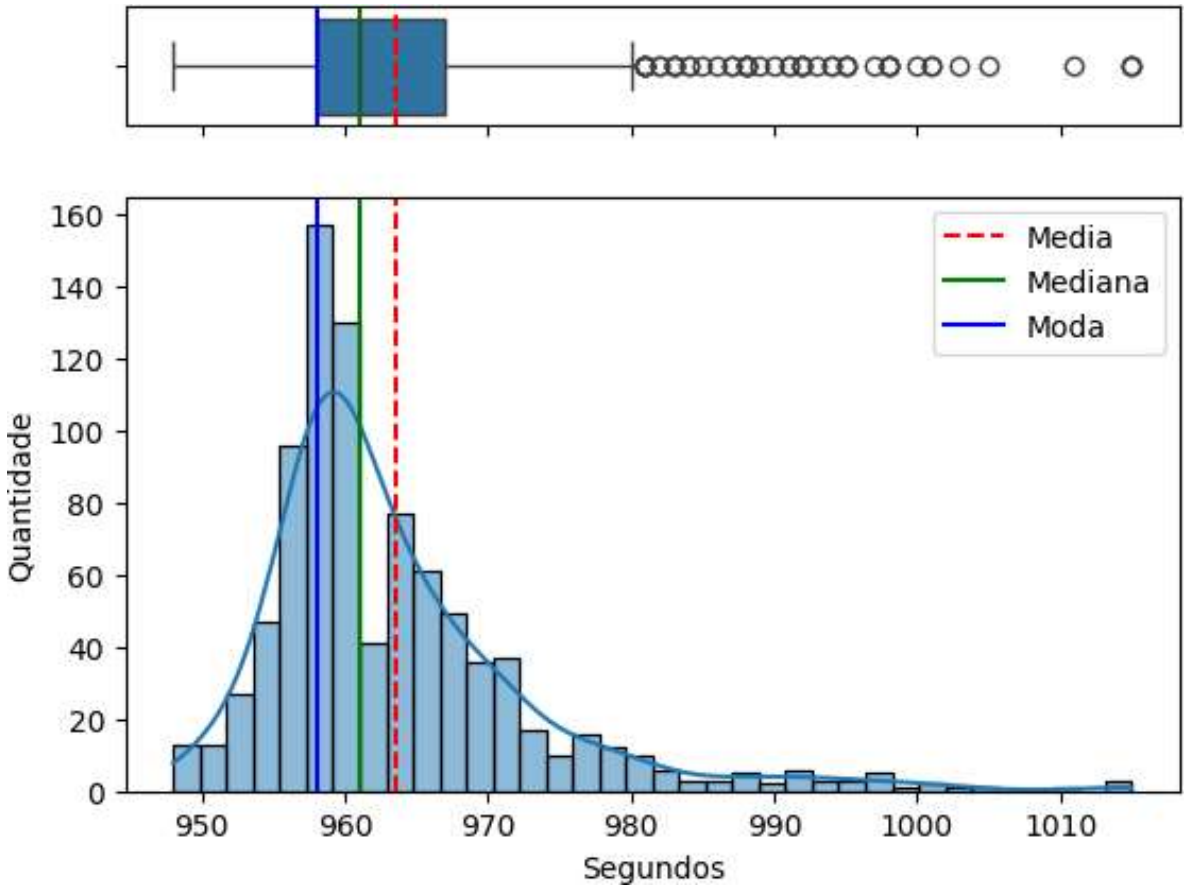

Figura 8.5: Distribuição da tomada de tempo excluído os 50 maiores valores.

Utilizando os dados tratados, observamos uma melhoria significativa no desempenho. Ao realizar a consulta diretamente no banco de dados usando a estrutura criada e carregada do *data warehouse*, obtivemos os mesmos 784 registros em um tempo médio de apenas 1 minuto e 21 segundos (ou seja, 81,7 segundos em média), como mostrado na Figura 8.6. Nesta análise gráfica, nota-se uma convergência entre a média e a moda dos tempos de execução o que indica uma convergência nos tempos de execução.

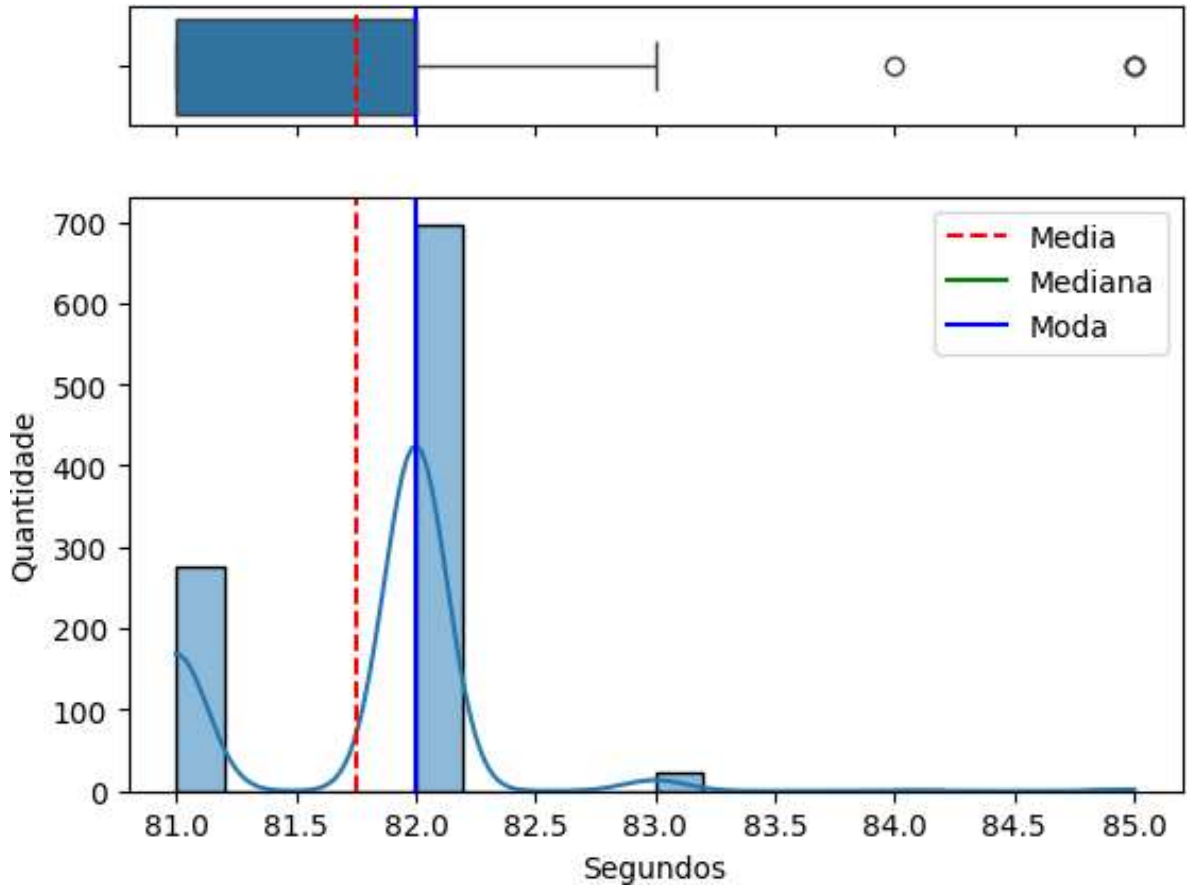

Figura 8.6: Distribuição da tomada de tempo de duração das execuções da nova estrutura

A sintaxe desta consulta é apresentada na Figura 8.7.

```
select * from ppca_infra_auditoria_operacao
where id_infra_auditoria_operacao in (
    select id_infra_auditoria_operacao
    from ppca_operacao_valor
    where valor like '%100009063%'
    and id_operacao_atributo in (
        select id_operacao_atributo
        from ppca_operacao_atributos
        where Nome like '%idusuario%'
        and id_operacao in (
            select id_operacao
            from ppca_operacao
            where acao like '%auditar%' and modo like '%Visualizacao%')))
```

Figura 8.7: Consulta em SQL utilizada na nova estrutura

Inicialmente, a tabela infra_auditoria contém 289.572.382 registros, abrangendo o período de novembro de 2015 a agosto de 2022, ocupando um espaço de 820 gigabytes.

Este espaço engloba não apenas a coluna "operação", mas também outras colunas da tabela. A estrutura criada e alimentada pelo processo de Extração, Transformação e Carga (ETL), que compreende o mesmo período, mas apenas com os dados da coluna "operação", ocupa 166 gigabytes, contendo um total de 1.786.446.877 registros.

É importante destacar que a consulta a essa nova estrutura foi realizada por meio de uma única operação, envolvendo junções entre diversas tabelas, o que impactou o tempo de resposta. No entanto, uma alternativa viável é disponibilizar esses mesmos dados aos usuários por meio de um painel de consulta, como o Power BI, onde a filtragem dos dados ocorre de forma sequencial, proporcionando aos usuários uma experiência de recuperação instantânea de informações.

8.3 Dashboard

Com o objetivo de fornecer aos administradores do SEI um mecanismo que conceda autonomia nas consultas de auditoria, foi desenvolvida uma interface de consulta utilizando o Power BI. Esta interface capacita os usuários a interagirem através de filtros e visualizarem os acessos registrados.

Além disso, a interface possibilita a apresentação de informações relacionadas à utilização do SEI, obtidas tanto da tabela infra_auditoria quanto das tabelas de registros de interação de documentos, conforme descrito na seção 1.1.

8.3.1 Processos

Gráficos de barras foram utilizados para representar a quantidade de documentos gerados e processos ao longo do tempo (por ano) e por unidade. Os dados quantitativos de documentos e processos (Figura 8.8) permitem análises sobre tendências temporais quando confrontados com o calendário acadêmico, por exemplo. Essa abordagem possibilita a identificação de picos de atividade e períodos de baixa demanda, auxiliando os administradores a planejar recursos e suporte durante os períodos de maior demanda.

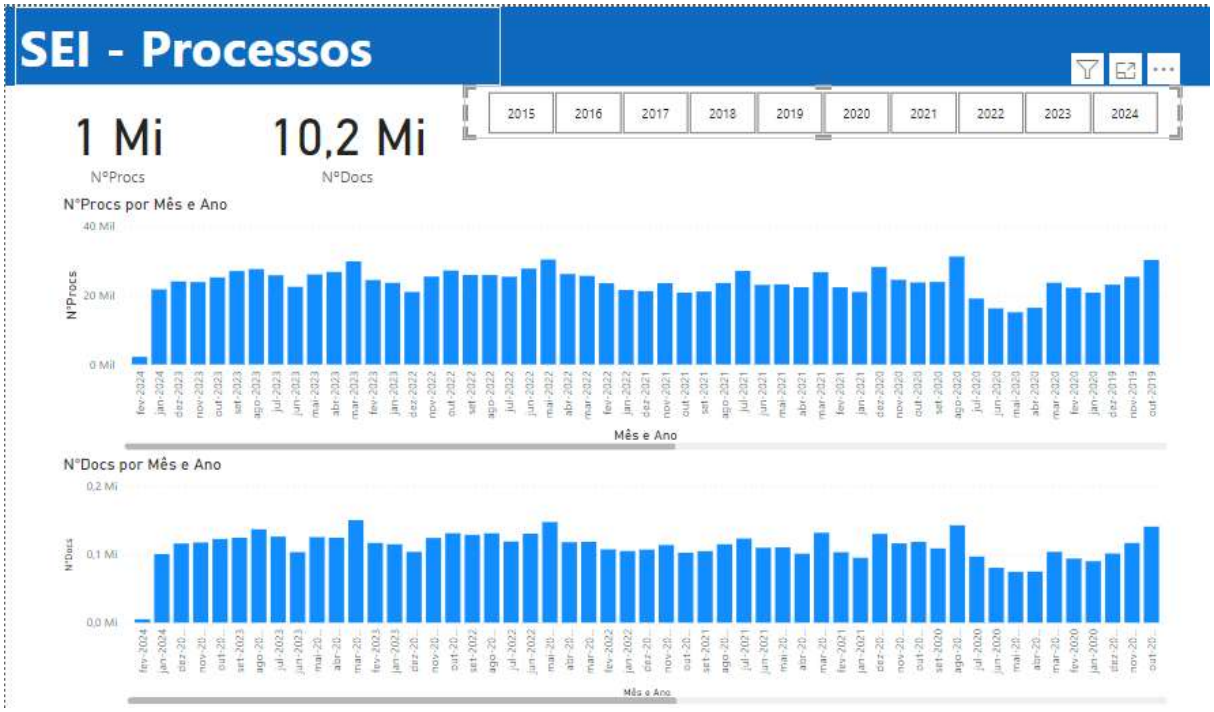

Figura 8.8: Dashboard com Quantitativo de Processos do SEI

A análise por unidade (Figura 8.9) demonstra a distribuição de documentos entre as diferentes unidades. Essa análise pode revelar disparidades e padrões interessantes que informam estratégias de alocação de recursos e treinamento.

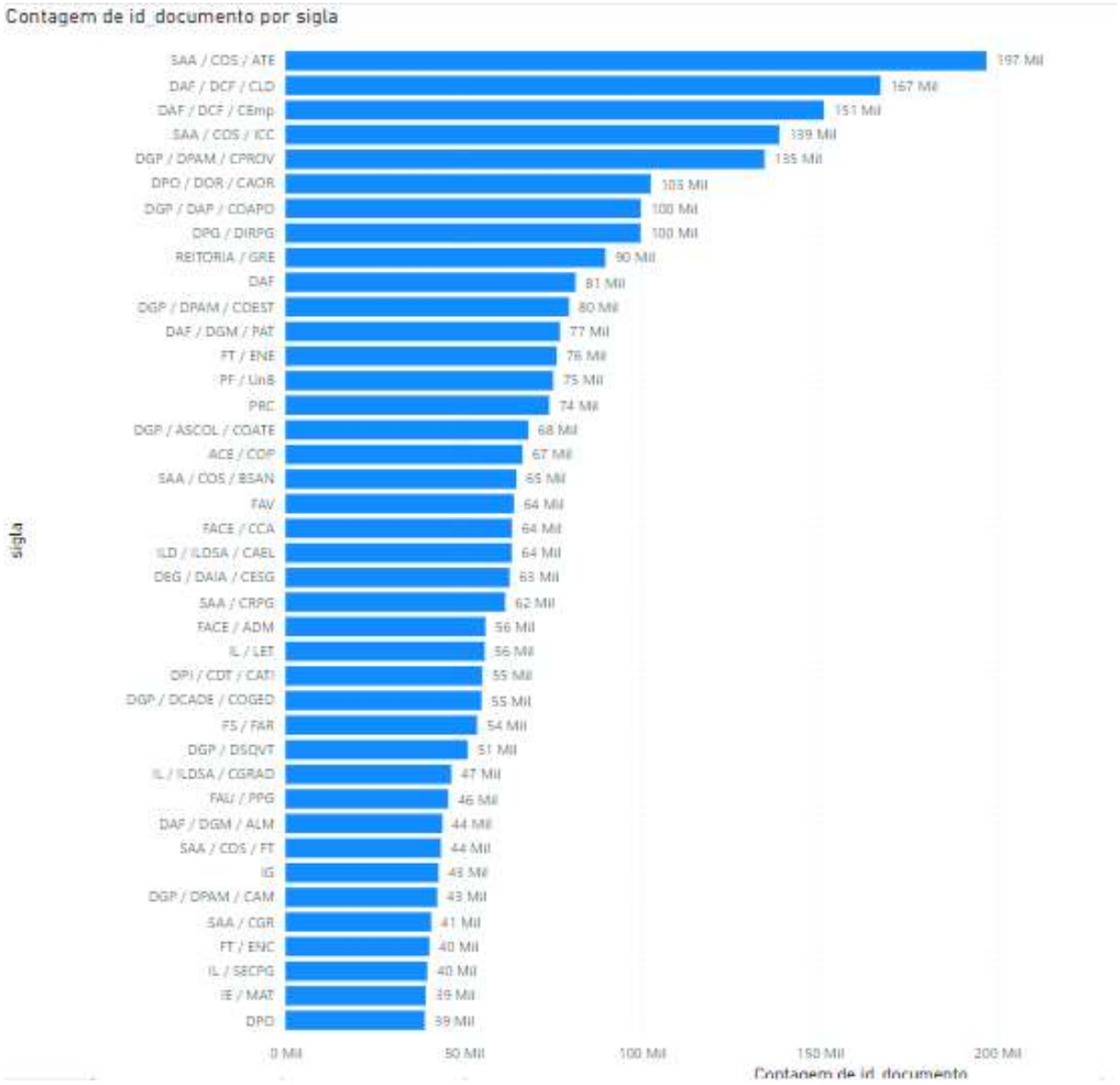

Figura 8.9: Dashboard com a Distribuição do quantitativo de processos por unidade

Essas distribuições também permitem analisar o contexto de eventos importantes que impactaram na utilização do SEI, como períodos de matrícula e a pandemia de COVID-19. Por exemplo, é possível identificar se determinadas medidas durante o período de matrícula influenciaram a geração de processos em uma unidade específica não diretamente correlacionada.

Os gráficos das Figuras 8.8 e 8.9 são interativos, o que significa que permitem a filtragem por período de tempo e unidade por meio de navegação com mouse.

8.3.2 Atividades

Este painel (Figura 8.10 elaborado oferece acesso aos registros das atividades arquivísticas do sistema. Essas atividades são registradas em uma tabela chamada "atividades"e podem também ser acessadas em páginas dedicadas dentro do sistema.

Neste painel, os registros podem ser obtidos de maneira mais prática e filtrados por usuário do sistema e processo.

A Figura 8.10 ilustra a distribuição das atividades por unidades. Após selecionar um nome específico, os processos são filtrados e um deles pode ser escolhido. Em seguida, todas as atividades relacionadas a esse processo são apresentadas.

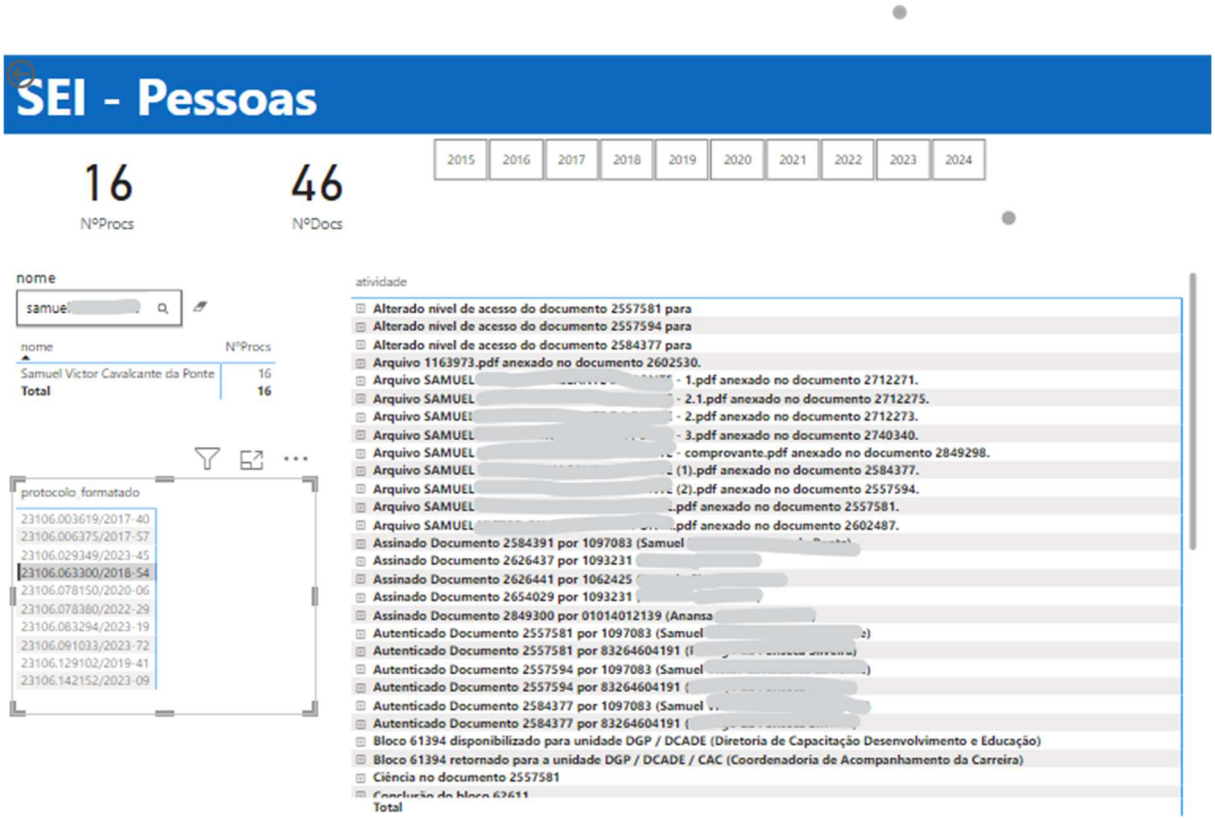

Figura 8.10: Dashboard com consulta de processos por Pessoas

8.4 Trabalhos Futuros

Foi feita o processamento e carga do passivo. A próximas cargas podem ser feitas ad hoc ou podem ser programadas conforme a necessidade, uma vez que não se mostraram demandar processamento ou tempo que seja fator de relevância.

Este projeto tem como objetivo aprimorar o acesso aos registros de auditoria, viabilizando a recuperação das informações de acessos futuros sem a necessidade de especificar um período ou outras restrições.

Adicionalmente, essa solução poderá ser prontamente replicada em outras unidades. Apesar de demandar um processamento inicial considerável, sua implementação será direta e poderá ser aplicada em qualquer instituição que utilize o SEI.

Outra vantagem consiste na capacidade de conduzir análises estatísticas e inferenciais no futuro. Os dados tratados poderão ser processados com efetividade, facilitando a identificação de padrões de comportamento, tendências e outros achados relevantes.

Além disso, essa solução viabiliza aos desenvolvedores do SEI, possam oferecer aos usuários um histórico mais completo do documento. Por exemplo, é possível agora adicionar um ícone à visualização do documento que exibirá todas interações e acessos já realizados.

Capítulo 9

Conclusão

O objetivo deste trabalho foi viabilizar consultas eficientes aos logs do SEI por meio de técnicas de ETL, limpando e organizando os dados para garantir desempenho e eficiência nas respostas às consultas dos usuários.

O resultado foi o desenvolvimento de uma estrutura adjacente otimizada, com estratégias de particionamento dos dados para melhor aproveitamento dos recursos de processamento do servidor. Além disso, foi criado um conjunto de códigos em PL/SQL que realiza a leitura na estrutura original, aplica o tratamento necessário e carrega os dados no novo modelo. Isso resultou em, conforme testes executados em consultas complexas, uma redução de 91% no tempo de resposta, passando de uma média de 16 minutos e 12 segundos para apenas 1 minuto e 21 segundos.

A implicação prática desse trabalho é que os usuários agora contam com uma estrutura de consulta de alto desempenho para buscas de auditoria, de maneira independente e eficiente. Além disso, qualquer unidade organizacional interessada pode adotar essa solução com facilidade, bastando executar os scripts de criação da estrutura e de carga dos dados.

Entre as dificuldades encontradas, destaca-se a impossibilidade de a UnB integrar essa consulta diretamente ao aplicativo do SEI, uma vez que a ela é apenas usuária do sistema, cujo desenvolvimento é centralizado pelo TRF4. Dado que o órgão responsável deve possuir outras demandas prioritárias, a implementação de funcionalidades adicionais de consulta de logs pode não ser uma prioridade imediata, dificultando a adoção direta dessa solução no curto prazo. Questão contornada com a elaboração de *dashboards*.

Outro ponto relevante é o grande volume de dados envolvido, que demanda um uso significativo de espaço, tempo e capacidade de processamento. A estrutura adjacente criada aumenta ainda mais esse consumo, o que pode representar um desafio para outras instituições que utilizam o SEI e desejem implementar a mesma solução.

Em síntese, todos os objetivos foram atingidos de maneira satisfatória, proporcionando à UnB um significativo ganho de produtividade nas atividades relacionadas à análise de logs do SEI.

Referências

[1] Laudon, Kenneth C e Jane P Laudon: *Sistemas de informação gerenciais: administrando a empresa digital*. Bookman Editora, 2022. xi, 25

[2] Kimball, Ralph e Margy Ross: *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011. xi, 18, 24, 26, 27, 28, 29, 30, 31, 32, 37

[3] Fonseca Silveira, Rodrigo da, Maristela Holanda, Marcio de Carvalho Victorino e Marcelo Ladeira: *Educational data mining: Analysis of drop out of engineering majors at the unb-brazil*. Em *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, páginas 259–262. IEEE, 2019. xi, 26, 30

[4] *Flaticon*. https://www.flaticon.com. Accessed: 2024-08-01. xii, 47

[5] https://www.mongodb.com/docs/. Accessed: 2023-10-10. xii, 39, 52, 53

[6] Brasil: *Lei geral de proteção de dados pessoais (lgpd). (redação dada pela lei n° 13.853, de 2019)*. Diário Oficial da República Federativa do Brasil, 2018, ISSN 00. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm, acesso em 2022-11-15. 2

[7] Brasil: *Lei de acesso à informação - lei n° 12.527, de 18 de novembro de 2011*. Diário Oficial da República Federativa do Brasil, 2011, ISSN 00. https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm, acesso em 2022-11-15. 2, 43

[8] Oliveira, Denis Lima de: *Agentes de tratamento de dados pessoais e encarregado: guia prático sobre suas atribuições, responsabilidades e boas práticas*. Tese de Doutoramento, unknow, 2021. 13, 40

[9] Marconi, Marina de Andrade e Eva Maria Lakatos: *Fundamentos de metodologia científica*. atlas, 2003. 17, 84

[10] Wazlawick, Raul Sidnei: *Metodologia de pesquisa para ciência da computação*, volume 2. Elsevier, 2009. 17, 85

[11] Inmon, William H: *Building the data warehouse*. John wiley & sons, 2005. 18, 24, 27, 28, 49

[12] Korotkevitch, Dmitri: *SQL Server Advanced Troubleshooting and Performance Tuning*. O'Reilly Media, Inc, 2022. 18, 24, 44

[13] Mariano, Ari Melo e Maíra Santos Rocha: *Revisão da literatura: apresentação de uma abordagem integradora*. Em *AEDEM International Conference*, volume 18, páginas 427–442, 2017. 18, 74

[14] Schering, Johannes, Jorge Marx Gómez, Lena Büsselmann, Federico Alfaro e Jan Stüven: *Potentials of bicycle infrastructure data lakes to support cycling quality assessment*. INFORMATIK 2022, 2022. 20, 22

[15] Casturi, Rao e Rajshekhar Sunderraman: *Cost effective, rule based, big data analytical aggregation engine for investment portfolios*. Wireless Networks, páginas 1–7, 2018. 20, 22

[16] Visweswaran, Shyam, Brian McLay, Nickie Cappella, Michele Morris, John T Milnes, Steven E Reis, Jonathan C Silverstein e Michael J Becich: *An atomic approach to the design and implementation of a research data warehouse*. Journal of the American Medical Informatics Association, 29(4):601–608, 2022. 20, 21

[17] Ndyanabo, Anthony, Kevin Footer, Tanvir Ahmed, Alex Glogowski, Christopher Whalen, Joseph Ssekasanvu, Lloyd Ssentongo, Tom Lutalo, Fred Nalugoda, Grace K Ha *et al.*: *Establishing a centralized data mart from the rakai community cohort study to improve hiv research in rakai, uganda*. JAMIA Open, 5(2):ooac032, 2022. 20, 21

[18] Zhao, He *et al.*: *Research on construction of educational management model based on data mining technology*. Journal of Applied Science and Engineering, 26(5):613–621, 2022. 20, 21

[19] Chernyshenko, Serge e Vsevolod Chernyshenko: *University digital document management and optimal strategy of education data warehouses' placement*. Em *2022 2nd International Conference on Technology Enhanced Learning in Higher Education (TELE)*, páginas 237–243. IEEE, 2022. 20, 21

[20] Zhang, Songtao e Guijun Yuan: *Substation operation information maintenance based on intelligent data mining*. Wireless Communications and Mobile Computing, 2022, 2022. 20, 23

[21] Xu, Wen: *Reflections on the discipline construction environment of world literature and comparative literature in the era of big data analysis.* Journal of Environmental & Public Health, 2022. 20, 21

[22] Almalawi, Abdulmohsen, Asif Irshad Khan, Fawaz Alsolami, Yoosef B Abushark, Ahmed S Alfakeeh e Walelign Dinku Mekuriyaw: *Analysis of the exploration of security and privacy for healthcare management using artificial intelligence: Saudi hospitals.* Computational Intelligence & Neuroscience, 2022. 20, 22

[23] Zhao, Bo, Yanjin Liu *et al.*: *Application of data warehouse technology based on neural network in physical education quality management*. Mathematical Problems in Engineering, 2022, 2022. 20, 22

[24] Sakib, Nazmus, Shah Jalal Jamil e Saddam Hossain Mukta: *A novel approach on machine learning based data warehousing for intelligent healthcare services*. Em *2022 IEEE Region 10 Symposium (TENSYMP)*, páginas 1–5. IEEE, 2022. 20, 22

[25] Zalozhnev, Alexey Yu, Vasily N Ginz e Anatoly Eu Loktionov: *Intelligent system and human-computer interaction for personal data cyber security in medicaid enterprises*. Em *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, páginas 1–4. IEEE, 2022. 20, 22

[26] Kovacic, Ilko, Christoph G Schuetz, Bernd Neumayr e Michael Schrefl: *Olap patterns: A pattern-based approach to multidimensional data analysis*. Data & Knowledge Engineering, 138:101948, 2022. 20

[27] Luo, Jia, Junping Xu, Obaid Aldosari, Sara A Althubiti e Wejdan Deebani: *Design and implementation of an efficient electronic bank management information system based data warehouse and data mining processing*. Information Processing & Management, 59(6):103086, 2022. 20

[28] Hidayanto, AN, HS Indriany, A Prastya, SF Mardiansyah *et al.*: *Data warehouse capability maturity model assessment for efficient monitoring process: a case study in national narcotics board*. Em *IOP Conference Series: Earth and Environmental Science*, volume 969, página 012055. IOP Publishing, 2022. 20

[29] Tufano, Alessandro, Riccardo Accorsi e Riccardo Manzini: *A machine learning approach for predictive warehouse design*. The International Journal of Advanced Manufacturing Technology, 119(3):2369–2392, 2022. 21

[30] Dean, Gregary, Joshua Moraes, Joseph White, Robert Deleon, Matthew Jones e Thomas Furlani: *Performance optimization of the open xdmod datawarehouse*. Em *Practice and Experience in Advanced Research Computing*, páginas 1–7. unknow, 2022. 21

[31] Sun, Yue yue: *Research and implementation of an efficient incremental synchronization method based on timestamp*. Em *2022 3rd International Conference on Computing, Networks and Internet of Things (CNIOT)*, páginas 158–162. IEEE, 2022. 21

[32] Mouna, Mustapha Chaba, Ladjel Bellatreche e Narhimene Boustia: *Prores: Proactive re-selection of materialized views*. Computer Science and Information Systems, 1(00):3–3, 2022. 21

[33] Lamer, Antoine, Mouhamed Djahoum Moussa, Romaric Marcilly, Régis Logier, Benoit Vallet e Benoît Tavernier: *Development and usage of an anesthesia data warehouse: lessons learnt from a 10-year project*. Journal of Clinical Monitoring and Computing, páginas 1–12, 2022. 21

[34] Guyot, Alexis, Annabelle Gillet, Eric Leclercq e Nadine Cullot: *A formal framework for data lakes based on category theory*. Em *Proceedings of the 26th International Database Engineered Applications Symposium*, páginas 75–83, 2022. 22

[35] Chen, Cuixia: *Risk analysis and countermeasures of social medical insurance based on random matrix theory and data mining*. Mathematical Problems in Engineering, 2022, 2022. 22

[36] Li, Kang, Junpeng Huang e Jie Lin: *The architecture of college psychological teaching management system based on data mining technology*. Security and Communication Networks, 2022, 2022. 22

[37] Jing, Zhixin, Rui Fan, Wanying Liu, Yan Shi e Fengjiu Yang: *Evaluation of online tool data management for warehouse management for power big data*. Em *2022 11th International Conference of Information and Communication Technology (ICTech))*, páginas 307–311. IEEE, 2022. 22

[38] Cordeiro, Kelli de Faria: *Modelagem Multidimensional de Dados*. John Wiley & Sons, 2011. 24

[39] Watson, Hugh J e Paul Gray: *Decision support in the data warehouse*. Prentice Hall Professional Technical Reference, 1997. 24, 25, 27

[40] Power, Daniel J: *Decision support systems: concepts and resources for managers*, volume 13. Quorum Books Westport, 2002. 24

[41] Efraim, Turban, E Aronson Jay, Ting Peng Liang e RV McCarthy: *Decision support systems and intelligent systems*. Yogyakarta: Andi, 2005. 25

[42] Keen, Peter GW e Michael S Scott Morton: *Decision support systems: an organizational perspective*. (No Title), 1978. 25

[43] Morton, Michael S Scott: *Decision support systems: Current practice and continuing challenges*. Sloan Management Review (pre-1986), 21(3):77, 1980. 25

[44] Watson, Hugh J e Barbara H Wixom: *The current state of business intelligence*. Computer, 40(9):96–99, 2007. 26

[45] Setzer, Valdemar W: *Dado, informação, conhecimento e competência*. DataGramaZero Revista de Ciência da Informação, n. 0, 28, 1999. 27

[46] Victorino, Marcio de Carvalho, Marcelo Shiessl, Edgard Costa Oliveira, Edson Ishikawa, Maristela Terto de Holanda e Marcal de Lima Hokama: *Uma proposta de ecossistema de big data para a análise de dados abertos governamentais concetados*. Informação & sociedade, 27(1):225–242, 2017. 27

[47] Antunes, António Lorvão, Elsa Cardoso e José Barateiro: *Incorporation of ontologies in data warehouse/business intelligence systems-a systematic literature review*. International Journal of Information Management Data Insights, 2(2):100131, 2022. 27, 28

[48] Sen, Arun e Varghese S Jacob: *Industrial-strength data warehousing*. Communications of the ACM, 41(9):28–31, 1998. 27

[49] Poe, Vidette, Stephen Brobst e Patricia Klauer: *Building a data warehouse for decision support*. Prentice-Hall, Inc., 1997. 27

[50] Tangsripairoj, Songsri e Premmanat Natseevatana: *A business intelligence system for radio communication licensing: A case study of the national broadcasting and telecommunications commission of thailand*. Em *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, páginas 1–6. IEEE, 2018. 28

[51] Radenković, Miloš, Jelena Lukić, Marijana Despotović-Zrakić, Aleksandra Labus e Zorica Bogdanović: *Harnessing business intelligence in smart grids: A case of the electricity market*. Computers in industry, 96:40–53, 2018. 29

[52] Chandra, Edward, Abba Suganda Girsang, Ryo Hadinata e Sani Muhamad Isa: *Analysis students' graduation eligibility using data warehouse*. Em *2018 International Conference on Information Management and Technology (ICIMTech)*, páginas 61–64. IEEE, 2018. 29, 32

[53] Sadalage, Pramod J e Martin Fowler: *NoSQL essencial: um guia conciso para o mundo emergente da persistência poliglota*. Novatec Editora, 2019. 37, 38

[54] *O que é mongodb*. https://www.ibm.com/br-pt/topics/mongodb. Accessed: 2024-08-01. 38

[55] Mia, Md Raihan, Abu Sayed Md Latiful Hoque, Shahidul Islam Khan e Sheikh Iqbal Ahamed: *A privacy-preserving national clinical data warehouse: Architecture and analysis*. Smart Health, 23:100238, 2022. 40

[56] *Guia prático do sei na unb*. https://www.portalsei.unb.br/images/documentos_sei/Guia_v3_0_Atualizado_10_7_17.pdf. Accessed: 2022-11-07. 42

[57] TRF4: *Manual do SEI*, 2022. https://softwarepublico.gov.br/social/sei/manuais. 42

[58] Brasil: *Lei n° 8.159, de 8 de janeiro de 1991*. Diário Oficial da República Federativa do Brasil, 1991, ISSN 00. https://www.planalto.gov.br/ccivil_03/leis/l8159.htm, acesso em 2022-11-15. 43

[59] Carvalho, Priscila Freitas de e Regina de Barros Cianconi: *A gestão de informações arquivísticas sob a vigência da lei de acesso à informação em ambiente universitário*. Em *XVI Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação*, 2015. 43

[60] *Solr - indexado textual*. https://solr.apache.org/. Accessed: 2022-11-16. 43

[61] *Vosviewer is a software tool for constructing and visualizing bibliometric networks*. https://www.vosviewer.com/. Accessed: 2022-11-07. 74

[62] *Tag crowd: word cloud for any text*. https://tagcrowd.com/. Accessed: 2022-11-07. 82

Anexo I

Teoria do Enfoque Meta Analítico Consolidado (TEMAC)

A metodologia TEMAC, descrita por [13], propõe três etapas principais para a pesquisa: (i) pesquisa de periódicos, na qual se busca identificar a literatura de maior impacto entre publicações científicas; (ii) análise por técnicas de bibliometria; e (iii) pesquisa exploratória com uma abordagem quantitativa.

Os dados obtidos a partir dos portais de periódicos foram posteriormente submetidos ao *software* VOSViewer [61] para compilação. Os resultados estão apresentados a seguir.

As pesquisas foram realizadas nos dois maiores portais de periódicos da área de tecnologia, o Scopus e o *Web Of Science*. A língua inglesa foi escolhida por conter a maior quantidade de publicações.

Entre os descritores, foram selecionados empiricamente os termos "*data warehouse*", "*analytical research*"e "*information management systems*". A busca foi limitada aos últimos cinco anos, visando a utilização de referências atualizadas sobre o tema.

A escolha do primeiro termo, "*Data Warehouse*", deve-se ao fato de que as técnicas aplicadas neste estudo estão entre as ferramentas de ETL.

Além das transformações realizadas, buscou-se prover informações sobre a utilização do sistema como um todo, o que motivou a inclusão do termo "*analytical research*".

O SEI caracteriza-se como um sistema típico de administração de informações, algo evidente em sua própria nomenclatura. Por isso, também foram incluídos textos que abordam "*information management systems*".

Adicionalmente, foram realizadas pesquisas incluindo o termo "*system usage records*"para buscar referências sobre a captação e manipulação de informações relacionadas ao registro de utilização do sistema. Contudo, esse termo não resultou em adição de novos resultados.

Observa-se uma queda na produção de trabalhos no último ano, conforme mostrado na Figura I.1, possivelmente em decorrência da recente pandemia.

I.1 Preparação da pesquisa

I.1.1 Scopus

A expressão utilizada no Scopus, importante fonte de artigos relacionados a tecnologia, foi "*(TITLE-ABS-KEY (data AND warehouse) OR TITLE-ABS-KEY (analytical AND research) OR TITLE-ABS-KEY (information AND management AND systems)) AND PUBYEAR > 2017 AND PUBYEAR > 2017* "que por sua vez retornou pouco mais de 170 mil registros.

I.1.2 Web Of Science (WoC)

A expressão aplicada foi "*((ALL=(data warehouse)) or ALL=(analytical research)) or ALL=(information management systems) and 2023 or 2022 or 2021 or 2020 or 2019 or 2018 (Anos da publicação)*". Foram encontrados 314.145 registros.

Limitados aos últimos 5 anos, também mostrou uma queda no último ano, contudo significativamente mais discreta que a do Scopus.

I.2 Apresentação e Inter-relação dos dados

As assertivas a seguir foram obtidas através da análise quantitativa dos resultados, em cada uma das plataformas.

I.2.1 Scopus

O *ranking* dos principais autores pode ser observado na figura I.2, com predominância daqueles elaborados na China, figur I.3. O Brasil figura na 13ª posição, com 4.695 publicações. Alguns países foram omitidos na figura I.3 para que o Brasil fosse representado na imagem.

I.2.2 Web Of Science (WoC)

A figura I.4 mostra a distribuição do quantitativo de publicações nos últimos cinco anos. Seus principais autores estão enumerados na figura I.5 e a distribuição dos principais países na figura I.6. O Brasil está na 15ª posição desta última relação.

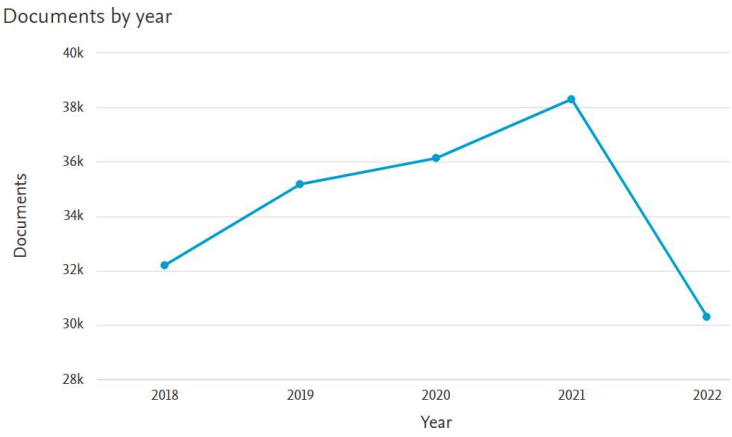

Figura I.1: SCOPUS - Distribuição por ano

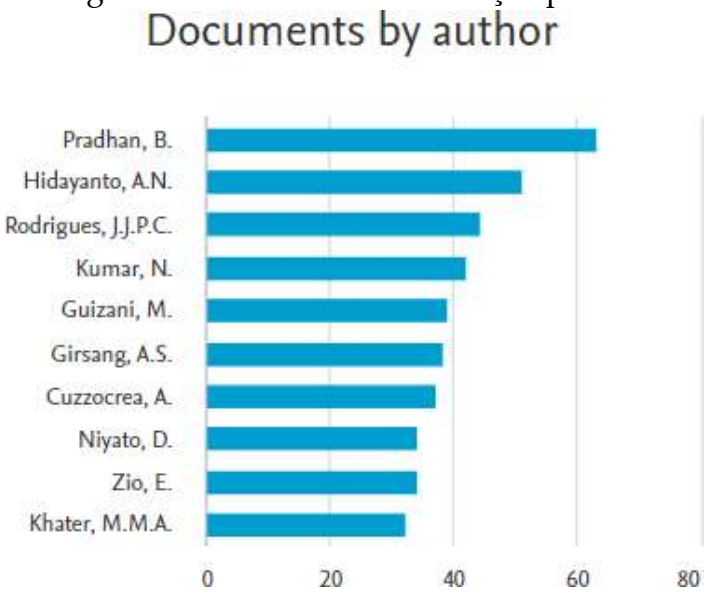

Figura I.2: SCOPUS - Principais autores

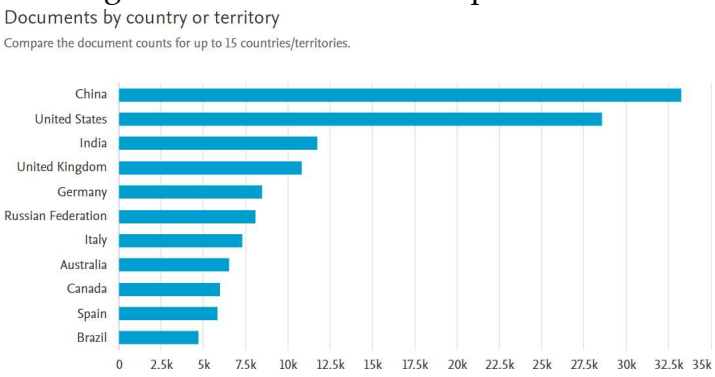

Figura I.3: SCOPUS - Ranking de países

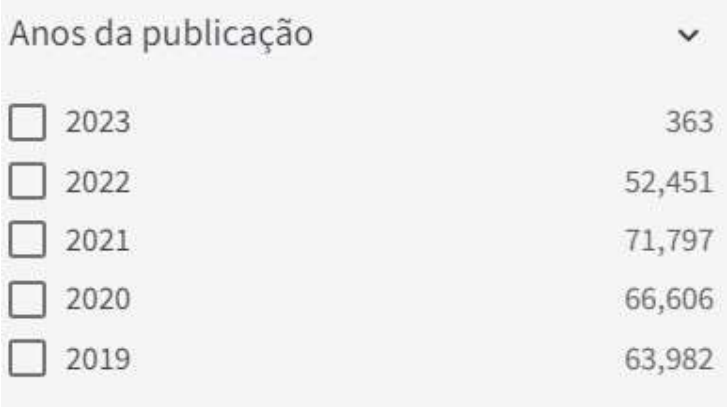

Figura I.4: WoC - Distribuição por ano

Zhai, Tianyou	109
Pradhan, Biswajeet	109
Baleanu, Dumitru	94
Asiri, Abdullah M.	85
Hou, Changjun	83
Amal, Rose	81
Wang, Jianhua	80
Yuan, Ruo	79
Yang, Haiping	77
Huo, Danqun	77
Cheng, Laifei	76
Kumam, Poom	72
Dadios, Elmer P.	69
Dwivedi, Yogesh Kumar	69
Li, Huiqiao	65

Figura I.5: WoC - Principais autores

PEOPLES R CHINA	77,285
USA	56,713
INDIA	21,280
GERMANY	20,416
ENGLAND	19,953
AUSTRALIA	14,876
CANADA	13,758
RUSSIA	13,544
ITALY	13,253
SPAIN	12,317
FRANCE	11,936

I.2.3 Análise Consolidada

A partir do levantamento desta etapa, houve uma análise ostensiva com o intuito de focar nos resultados que de fato tivessem pertinência com este trabalho. A seção a seguir, faz a apresentação desta fase.

Os primeiros resultados de cada uma das fontes foram analisados e aqueles que não se apresentaram pertinentes descartados. Assim, após os descartes, restaram na plataforma Scopus 19 documentos e, em WoC, 30 trabalhos.

Destaca-se que as análises feitas com ajuda do VOSviewer, da participação de coautoria das obras, apresentaram os mesmos resultados gráficos como podem ser vistos nas figuras I.7 e I.8, embora houvesse alteração dos autores das obras. Todos os autores desta seleção possuíam apenas uma obra na relação.

A tabela I.1 traz a relação dos países com o maior número de publicações, limitando-se àqueles com pelo menos 3 obras foram:

País	**Obras**
Estados Unidos	7
Espanha	5
China	3
Rússia	3
Portugal	3

Tabela I.1: Obra por países

O Brasil tem 2 obras, apenas na plataforma "*Web Of Science*".

I.3 Detalhamento do modelo integrador e validação por evidências

Esta terceira etapa do TEMAC tem por objetivo apresentar os fatos bibliométricos relevantes. Para tanto, foram selecionados os critérios de cocitação, acoplamento bibliográfico e a frequência de palavras-chave.

I.3.1 Scopus

A análise de cocitação, feita no VOSviewer (figura I.9), não apontou artigo, ou grupo de artigos, que se destacasse dentre as citações. Ou seja, não há na produção referência significativa.

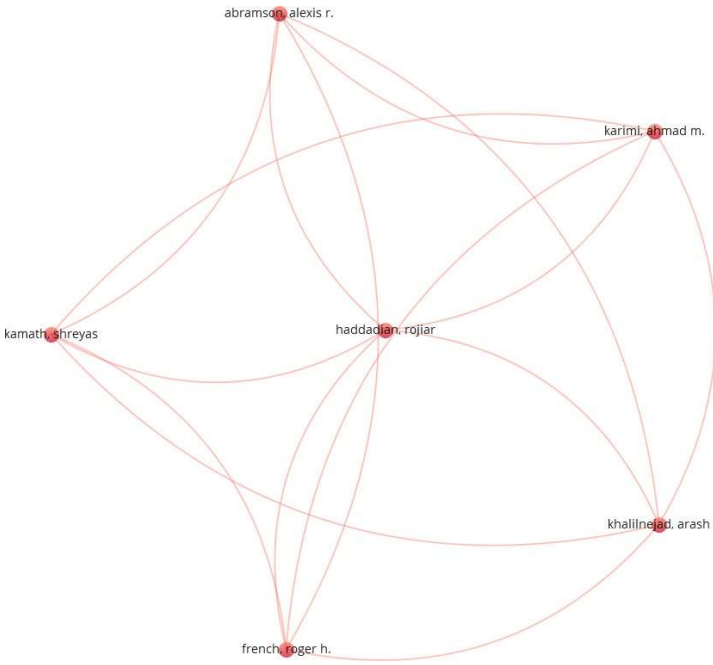

Figura I.7: WoC - Coautoria

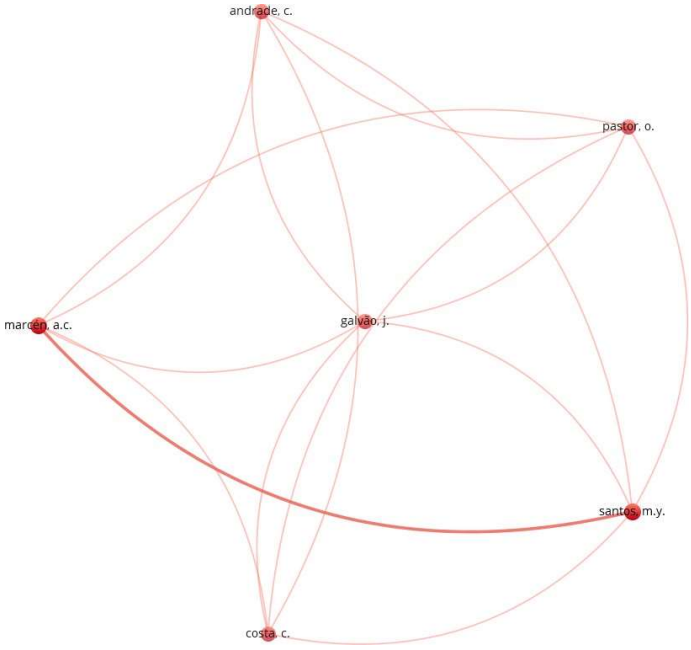

Figura I.8: Scopus - Coautoria

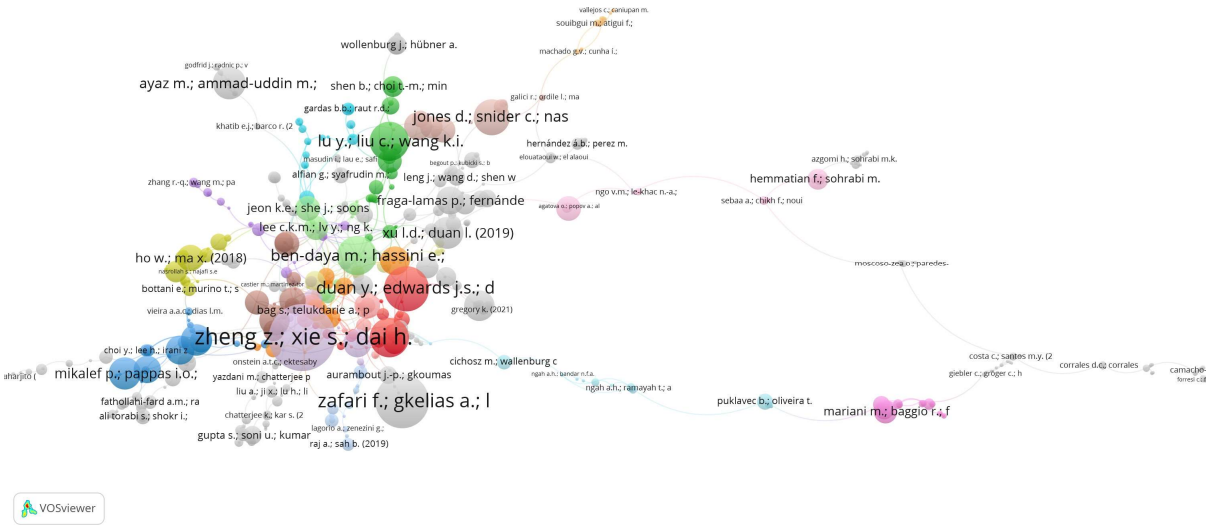

Figura I.9: Scopus - Correlações de citações

De maneira análoga, o acoplamento de termos, que busca identificar artigos que tenham algum tipo de correlação aos assuntos tratados, mostra, na figura I.10, os principais "*clusters*"encontrados. Também espaços, denotam pouca interdependência.

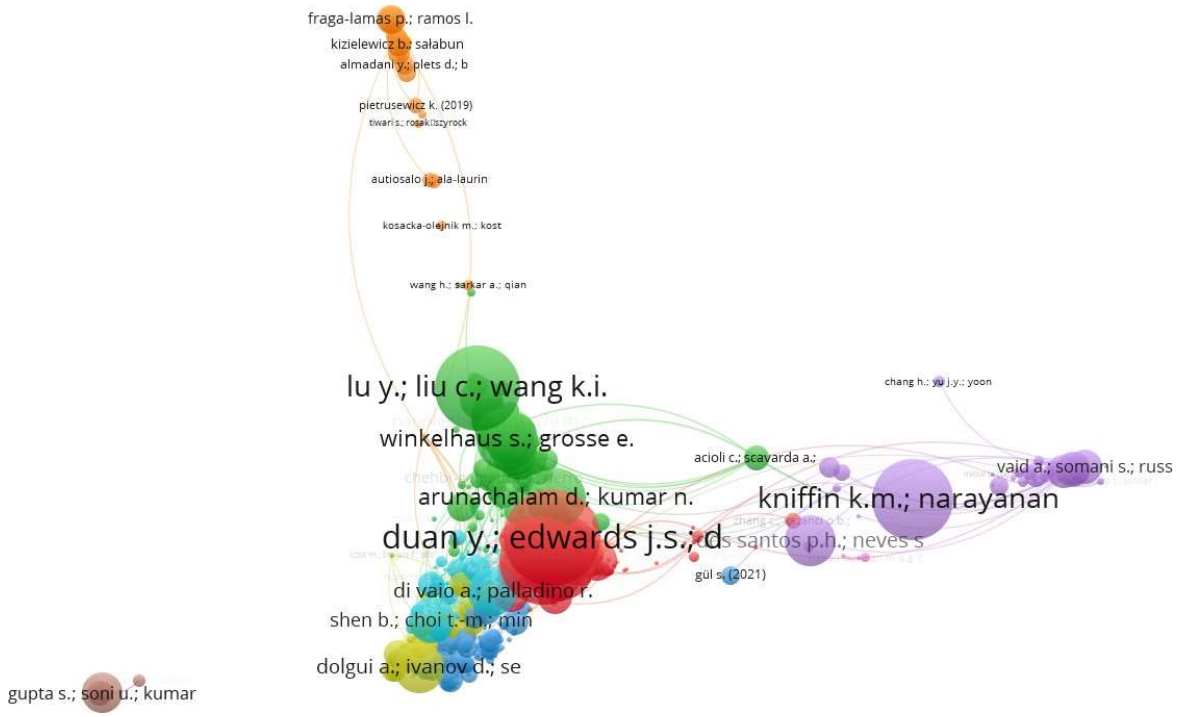

Figura I.10: Scopus - Acoplamento de Termos

I.3.2 Web Of Science (WoC)

A análise do VOSviewer de cocitações de obras (figura I.11) aponta que, embora existam obras que se apresentem com maior destaque que outras, estas ainda não possuem relevância que influenciem muito além de seu ciclo de pesquisa.

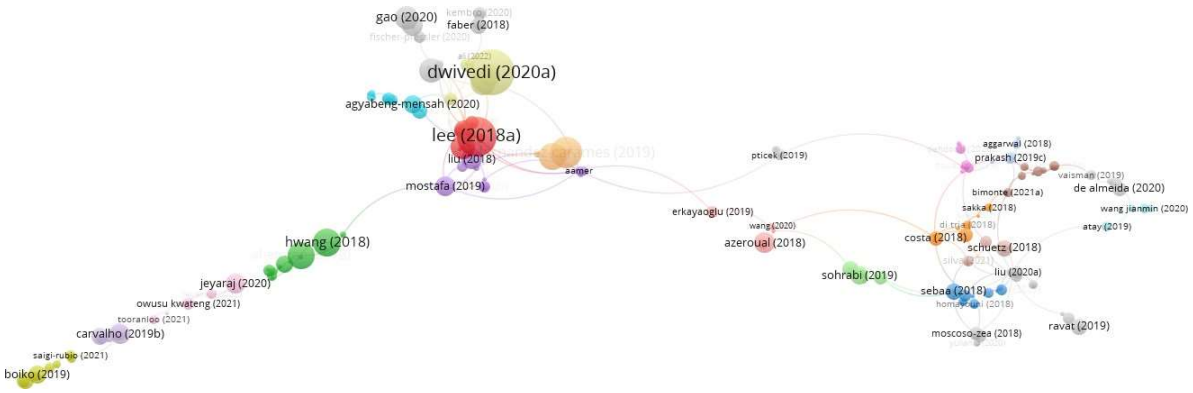

Figura I.11: WoC - Citações de Obras

O acoplamento de termos do WoC (figura I.12) aponta resultado mais interessante que o do Scopus (figura I.10). Pode-se observar que neste caso há nuvens, bastante distintas, as quais indicam que há "*clusters*"de assuntos significativos e estes se relacionam com outros "*clusters*".

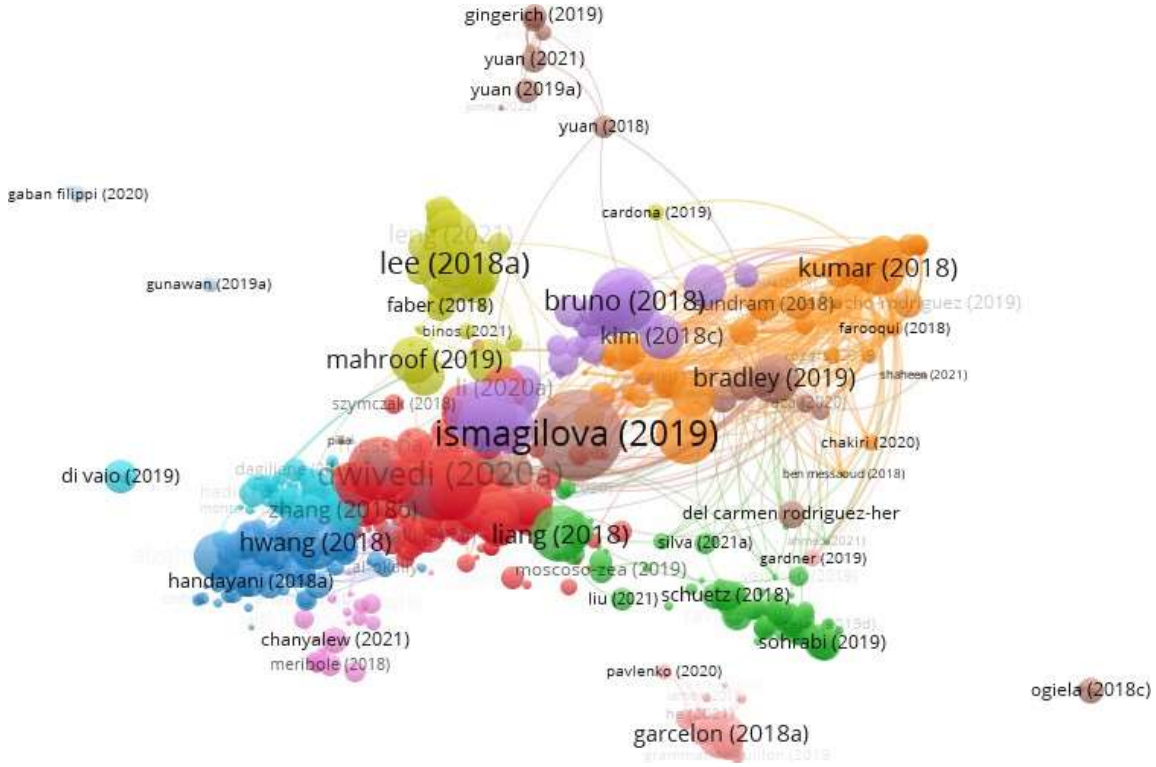

Figura I.12: WoC - Acoplamento de Termos

I.3.3 Frequência de Palavras Chave

Para a contagem da frequência de palavras-chave, foram extraídos dados de ambos, Scopus e WoC, que consolidadas, foram analisadas como se um conjunto fosse. Esta etapa considerou duas formas de contagem. Na primeira, foram contadas palavras independentemente de sua locução. Como exemplo, no termo "*data warehouse*"a palavras "*data*"e "*warehouse*"são contadas uma vez cada uma.

A nuvem de palavras da figura I.13 foi elaborada a partir destes termos. Em seguida, esta extração foi submetida ao portal [62]. Esta nuvem apresenta as 50 palavras de maior ocorrência, majorando os atributos da fonte (cor, tamanho e transparência) de acordo com sua relevância.

Figura I.13: Nuvem de palavras consolidada

A tabela I.2 lista as 10 palavras que mais ocorrem e sua respectiva frequência.

Nº	**Palavra**	**Ocorrências**
01	data	136
02	information	47
03	management	40
04	systems	31
05	analysis	30
06	warehouse	25
07	system	22
08	processing	20
09	analytics	18
10	decision	18

Tabela I.2: 10 palavras mais repetidas

Na segunda forma (figura I.14) foram consideradas as locuções. Portanto, o termo "*data warehouse*", por exemplo, é como se uma palavra o fosse. A tabela I.3 traz os 10 primeiros termos de maior ocorrência.

Figura I.14: Nuvem de locuções consolidada

Nº	Termo	Ocorrências
01	data warehouse	23
02	olap	12
03	data mining	8
04	big data	7
05	business intelligence	7
06	analytics	5
07	data analysis	5
08	data mart	4
09	data visualization	4
10	management	4

Tabela I.3: 10 primeiras locuções repetidas

I.4 Classificação da Pesquisa

Este projeto pode ser classificado pelo critério "Natureza", segundo [9], pois trata de resultados que podem ser aplicados na solução de problemas que ocorrem na realidade, segundo o autor.

E desta forma, sua Natureza é quantitativa, pois seus resultados podem ser aferidos numericamente, de acordo com o desempenho do tempo de resposta.

E ainda, também deve ser classificada como "Original", de acordo com Wazlawick [10], uma vez que cria um conjunto de ferramentas de administração de dados, embora aplicados a um caso em específico, e também faz uso de técnicas, de acordo com melhores regras de boas práticas, a partir de conceitos já estabelecidos, bem como conhecimentos tácitos adquirido ao longo dos anos.

Busca assim, um bom resultado a partir destas técnicas e explicar, de acordo com a literatura, o porquê do resultado obtido. É informação relevante, uma vez que são entendidos processos e ainda com implicação prática.

Enquadra-se, também segundo Wazlawick [10] como uma "Pesquisa Explicativa", pois analisa desempenho e comportamento observados, tentando explicá-los para interpretá-los oferecendo alternativa aos mesmos.

É "Experimental", dado que são empregadas técnicas de computação seguidas de observações para que se possa mensurar diferenças de desempenho [10].