UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

VINICIUS LUIZ DA SILVA FRANÇA

Integração de dados para a construção de KPIs em uma empresa de varejo: um estudo comparativo entre uma abordagem distribuída e uma centralizada

RECIFE

2023

UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

SISTEMAS DE INFORMAÇÃO

VINICIUS LUIZ DA SILVA FRANÇA

Integração de dados para a construção de KPIs em uma empresa de varejo: um estudo comparativo entre uma abordagem distribuída e uma centralizada

TCC apresentado ao Curso de Sistemas de Informação da Universidade Federal de Pernambuco, Centro de Informática, como requisito para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador(a): Prof. Robson do Nascimento Fidalgo

RECIFE

2023

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

França, Vinícius Luiz da Silva.
Integração de dados para a construção de KPIs em uma empresa de varejo: um estudo comparativo entre uma abordagem distribuída e uma centralizada / Vinícius Luiz da Silva França. - Recife, 2023.
45 p. : il., tab.

Orientador(a): Robson do Nascimento Fidalgo
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Sistemas de Informação - Bacharelado, 2023.

1. KPI. 2. Integração de dados. 3. Varejo. I. Fidalgo, Robson do Nascimento. (Orientação). II. Título.

000 CDD (22.ed.)

Integração de dados para a construção de KPIs em uma empresa de varejo: um estudo comparativo entre uma abordagem distribuída e uma centralizada

TCC apresentado ao Curso de Sistemas de Informação da Universidade Federal de Pernambuco, Centro de Informática, como requisito para a obtenção do título de Bacharel em Sistemas de Informação.

Aprovado em: ___/___/_____.

BANCA EXAMINADORA

_____
Profº. Dr. Robson do Nascimento Fidalgo (Orientador)
Universidade Federal de Pernambuco

_____
Profº. Dr. Luciano De Andrade Barbosa (Examinador Interno)
Universidade Federal de Pernambuco

Dedico a minha mãe, família e amigos.

AGRADECIMENTOS

Venho agradecer primeiramente a Deus por me guiar na minha jornada, à minha mãe e meu irmão, que acompanharam de perto minha dedicação para alcançar meus objetivos e, com isso, me deram forças para continuar lutando dia após dia. Venho agradecer também aos meus amigos da faculdade, em especial a Alexsandro Henrique, Luis Gabriel, Daniel Moraes, Gustavo Prazeres, João Pedro e Flávio Henrique, que compartilharam comigo esse período de aprendizado, desafios e conquistas. Agradeço a todos os meus colegas de trabalho, pelos quais sou grato por me fazer evoluir não apenas como profissional, mas também como pessoa. Por fim, venho agradecer aos meus antigos colegas de trabalho que comemoraram comigo minha entrada na Universidade Federal de Pernambuco.

RESUMO

Os Indicadores-Chave de Desempenho (KPIs) representam uma ferramenta essencial na gestão de empresas, fornecendo uma visão clara e concisa do desempenho organizacional. No entanto, a escolha da abordagem de integração de dados para a geração desses KPIs apresenta desafios e vantagens. O uso de um sistema de banco de dados distribuído pode oferecer escalabilidade, mas também pode acarretar problemas de sincronização e consistência de dados. Este estudo tem como objetivo principal realizar uma análise comparativa entre duas abordagens de integração de dados na geração de KPIs. Na primeira abordagem, utilizando Database Links e Procedures de banco de dados Oracle, os cálculos dos KPIs são realizados em bancos de dados distribuídos. Por outro lado, na segunda abordagem, que faz uso da ferramenta de replicação de dados Qlik Replicate, os cálculos dos KPIs são centralizados em um único banco de dados. A meta central é otimizar a capacidade da organização de medir e melhorar seu desempenho por meio da geração de KPIs, sem prejudicar sua operação, assegurando a eficiência na disponibilização de dados. Para alcançar esse objetivo, o estudo propõe levantar métricas que avaliam o processamento de dados nas diferentes abordagens de integração e analisar o processo de implantação da integração de dados no ambiente centralizado de banco de dados, explorando desafios técnicos e organizacionais associados. Os resultados deste estudo destacam que a utilização de KPIs baseados em dados integrados em um banco de dados centralizado oferece uma solução mais eficiente para as necessidades de análise em empresas de varejo. A centralização dos dados proporciona vantagens significativas, incluindo maior segurança e privacidade dos dados. No entanto, desafios técnicos e organizacionais são identificados em ambos os cenários de teste, sendo mais proeminentes no primeiro cenário devido ao impacto direto no atendimento ao cliente e na operação das lojas. Portanto, a escolha entre essas abordagens dependerá das prioridades da empresa em relação ao equilíbrio entre eficiência, estabilidade e impacto organizacional, ressaltando a importância da conscientização e do alinhamento das estratégias de integração.

Palavras-chave: KPI, Integração de dados, Varejo.

ABSTRACT

Key Performance Indicators (KPIs) represent an essential tool in business management, providing a clear and concise view of organizational performance. However, choosing the data integration approach for generating these KPIs presents challenges and advantages. Using a distributed database system can offer scalability, but it can also lead to data synchronization and consistency issues. This study's main objective is to carry out a comparative analysis between two data integration approaches in generating KPIs. In the first approach, using Database Links and Oracle Database Procedures, KPI calculations are performed in distributed databases. On the other hand, in the second approach, which makes use of the Qlik Replicate data replication tool, KPI calculations are centralized in a single database. The central goal is to optimize the organization's ability to measure and improve its performance through the generation of KPIs, without harming its operation, ensuring efficiency in data availability. To achieve this objective, the study proposes to raise metrics that evaluate data processing in different integration approaches and analyze the process of implementing data integration in the centralized database environment, exploring associated technical and organizational challenges. The results of this study highlight that the use of data-based KPIs integrated into a centralized database offers a more efficient solution to the analysis needs of retail companies. Centralizing data provides significant benefits, including greater data security and privacy. However, technical and organizational challenges are identified in both test scenarios, being more prominent in the first scenario due to the direct impact on customer service and store operations. Therefore, the choice between these approaches will depend on the company's priorities in relation to the balance between efficiency, stability and organizational impact, highlighting the importance of awareness and alignment of integration strategies.

Keywords: KPI, Data Integration, Retail.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIAÇÕES

CDC	Change data capture
SQL	Structured Query Language
DDL	Data Definition Language
DCL	Data Control Language
DML	Data Manipulation Language
DBMS	Database Management System
KPI	Key Performance Indicator
DBLINK	Database Link
BD	Banco de Dados

SUMÁRIO

1 INTRODUÇÃO

1.1 Contexto

Os Indicadores-Chave de Desempenho (KPI) são números projetados para transmitir de forma sucinta o máximo de informações possível. Bons indicadores-chave de desempenho são bem definidos, bem apresentados, criam expectativas e impulsionam ações [1]. Os KPIs atuam como uma base estruturante para fornecer direcionamento e simplificação das decisões dos gestores, pois eles são ferramentas projetadas para simplificar a relação das pessoas com os dados da web e orientar a ação. Ao fornecer informações relevantes e em um formato fácil de entender, os KPIs permitem que os gestores avaliem rapidamente o desempenho e tomem decisões apropriadas [1].

A utilização de um sistema de banco de dados distribuído para análise de dados apresenta várias vantagens, como escalabilidade, tolerância a falhas, segurança e desempenho. No entanto, essa abordagem também apresenta desvantagens, como a necessidade de sincronização, a possibilidade de inconsistência de dados [15].

O método incremental de extração de dados (Change Data Capture - CDC) é utilizado para determinar e detectar mudanças nos dados que ocorreram durante uma transação no sistema operacional. Ele pode ser usado para suportar o sistema ETL, reduzindo a quantidade de dados processados. O processo ETL pode ser executado de forma mais eficiente porque processa apenas os dados que foram alterados [3]. O método de extração completa (Full Load) é um método lógico de extração de dados que extrai todos os dados do sistema de origem, sem a necessidade de rastrear as alterações feitas no sistema de origem em relação à extração anterior. Com isso, o Full Extraction pode ser ineficiente e consumir muitos recursos, especialmente quando há grandes volumes de dados [4].

1.2 Motivação

No cenário atual do mercado varejista, a obtenção e análise de informações estratégicas desempenham um papel fundamental na busca pela competitividade e sucesso empresarial. A crescente complexidade das operações de varejo, juntamente com a demanda por decisões embasadas em dados, amplifica a necessidade de uma abordagem robusta na integração de informações provenientes de múltiplas fontes. A construção de KPIs (Indicadores-Chave de Desempenho) a partir de um ambiente de banco de dados centralizado surge como um diferencial, permitindo a otimização operacional, a tomada de decisões informadas e a adaptação ágil às dinâmicas de mercado, consolidando as empresas como líderes em um ambiente dinâmico e competitivo.

Neste contexto, a motivação para este estudo é impulsionada pela necessidade de comparar e avaliar duas abordagens de integração de dados no contexto da geração de KPIs. Para isso, utilizamos cenários de teste distintos: o primeiro cenário empregou Database Links e Procedures do banco de dados Oracle para a integração de dados em uma abordagem de banco de dados distribuído, enquanto o segundo cenário utilizou a ferramenta de replicação de dados Qlik Replicate para realizar a integração de dados em uma abordagem de banco de dados centralizado.

A mudança em direção a um ambiente de banco de dados centralizado surge como resposta aos desafios enfrentados pelas operações varejistas. A geração descentralizada de KPIs tem resultado em alto tempo de execução de processos, dificuldades no controle dos procedimentos e obstáculos na correção de erros. Esses obstáculos impactam diretamente a habilidade da empresa de adaptar-se rapidamente às mudanças e tomar decisões estratégicas embasadas. A transição para um banco de dados centralizado oferece a perspectiva de superar esses obstáculos, fornecendo uma abordagem centralizada e unificada que permitirá à empresa agir de maneira mais ágil, eficiente e proativa.

A consolidação de dados de várias fontes em um repositório único permite que as empresas tenham uma visão mais completa e integrada de seus dados, o que pode levar a insights mais profundos e informados [5].

1.3 Objetivo geral

O objetivo geral deste estudo é analisar e comparar duas abordagens de integração de dados (Database Links e Procedures de banco de dados Oracle vs. Qlik Replicate) no contexto da geração de KPIs. Nesta direção, esse estudo busca analisar a capacidade da organização em medir e melhorar seu desempenho sem impactar negativamente em sua operação, garantindo a eficiência na disponibilização de dados.

1.3.1 Objetivos secundários

- Levantar métricas que avaliam o processamento de dados nas diferentes abordagens de integração.
- Analisar o processo de implantação da integração de dados no ambiente centralizado de banco de dados, abordando desafios técnicos e organizacionais.

1.4 Perguntas de pesquisa

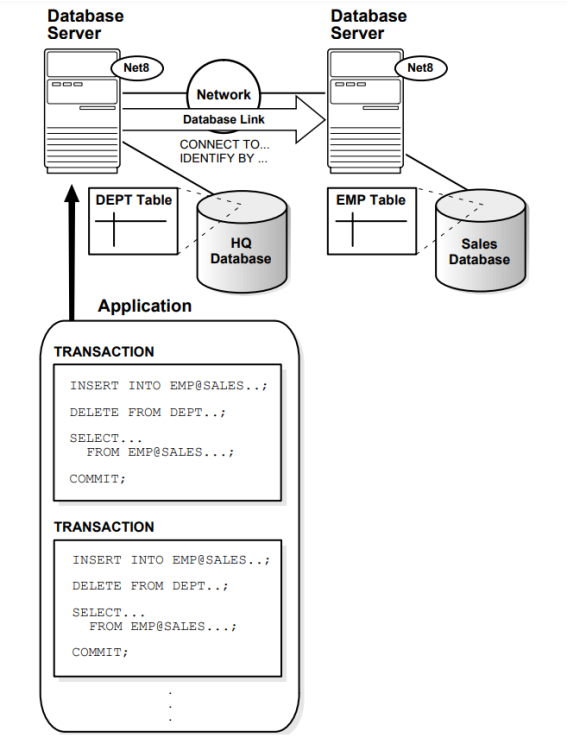
- Quais métricas de processamento de dados se destacaram nas duas abordagens de integração e como essas métricas afetaram a eficiência na disponibilização de dados?
- Quais são os desafios técnicos e organizacionais enfrentados durante a implementação da integração de dados nas duas abordagens e como esses desafios afetam a operação da organização?

2 REFERENCIAL TEÓRICO

2.1 Banco de dados distribuído

Um banco de dados distribuído é um sistema de gerenciamento de banco de dados que armazena dados em vários computadores em uma rede. Esses computadores são conectados por meio de uma rede de comunicação e trabalham juntos para fornecer aos usuários acesso aos dados armazenados. Cada computador no sistema é capaz de processar consultas e transações de banco de dados, permitindo que o sistema seja escalável e tolerante a falhas. O objetivo de um banco de dados distribuído é fornecer aos usuários acesso rápido e confiável aos dados, independentemente de sua localização física. Além disso, um banco de dados distribuído pode ser configurado para fornecer segurança, privacidade e controle de acesso aos dados. Os bancos de dados distribuídos são importantes porque permitem que as organizações gerenciem grandes quantidades de dados de forma eficiente [14]. Por exemplo, na Figura 1, o computador que gerencia o banco de dados HQ atua como um servidor de banco de dados quando uma instrução é emitida contra seus dados locais e atua como um cliente quando emite uma instrução com dados remotos presentes no banco de dados SALES.

Figura 1 – Ilustração de um sistema de banco de dados distribuído.



Fonte: ORACLE. Oracle® Database Database Administrator's Guide, 2021. Figura extraída da página 524.

Embora os bancos de dados distribuídos ofereçam muitas vantagens, eles também apresentam desafios significativos que precisam ser considerados [14]. Alguns dos principais desafios incluem:

- **Complexidade:** os bancos de dados distribuídos são mais complexos do que os bancos de dados centralizados, pois envolvem vários servidores e redes de comunicação. Isso pode tornar a configuração, manutenção e solução de problemas mais desafiadores.
- **Consistência:** manter a consistência dos dados em um ambiente distribuído pode ser um desafio, pois as atualizações em um servidor podem não ser imediatamente refletidas em outros servidores. Isso pode levar a problemas de integridade de dados e conflitos.
- **Segurança:** garantir a segurança dos dados em um ambiente distribuído pode ser um desafio, pois os dados são armazenados em vários servidores e

podem ser acessados por usuários em diferentes locais. Isso pode aumentar o risco de violações de segurança e acesso não autorizado.

- **Desempenho:** garantir um desempenho consistente em um ambiente distribuído pode ser um desafio, pois os dados precisam ser transferidos entre servidores e redes de comunicação. Isso pode levar a problemas de latência e gargalos de rede.
- **Custo:** os bancos de dados distribuídos podem ser mais caros do que os bancos de dados centralizados, pois envolvem a configuração e manutenção de vários servidores e redes de comunicação.

2.2 Gerenciamento de Metadados

Metadados são informações que descrevem os dados em um sistema de informação. Eles fornecem contexto e significado para os dados, permitindo que os usuários entendam melhor o conteúdo e a estrutura dos dados. Eles podem ser gerenciados por meio de ferramentas de metadados, como catálogos de dados, que permitem que os usuários pesquisem e naveguem pelos dados de maneira mais fácil e eficiente [8].

Existem diferentes tipos de metadados que podem ser gerenciados em um banco de dados. Alguns dos principais tipos incluem [8]:

1. Metadados técnicos: esses metadados descrevem as características técnicas dos dados, como o formato, a estrutura, o tipo de dados e outras informações relevantes. Eles são importantes para garantir que os dados sejam processados e armazenados corretamente.

2. Metadados de negócios: esses metadados descrevem os dados em termos de seu significado e contexto de negócios. Eles incluem informações sobre os processos de negócios, as regras de negócios e outras informações relevantes que ajudam os usuários a entender o significado dos dados.

3. Metadados operacionais: esses metadados descrevem as operações que são realizadas nos dados, como a limpeza, transformação e agregação. Eles são importantes para garantir que os dados sejam processados corretamente e que os resultados sejam precisos.

As informações de metadados podem afetar significativamente a interpretação e análise de KPIs organizacionais. Os metadados de negócios são particularmente importantes para a interpretação de KPIs, pois fornecem informações sobre o significado e o contexto dos dados. Sem essas informações, os usuários podem interpretar erroneamente os KPIs e tomar decisões equivocadas [8].

Além disso, a gestão de metadados é fundamental para garantir a qualidade dos dados e para facilitar a análise de dados [8].

2.3 KPI (Indicador-chave de desempenho)

KPI significa "Key Performance Indicator" ou Indicador-chave de Desempenho, em português. Eles são métricas utilizadas para avaliar o desempenho empresarial e medir o progresso em relação a objetivos específicos. Os KPIs são importantes porque ajudam as empresas a entenderem como estão se saindo em relação às suas metas e a identificar áreas que precisam de melhorias [1].

A escolha de KPIs relevantes e alinhados aos objetivos estratégicos da empresa é fundamental para garantir que os KPIs sejam eficazes na avaliação do desempenho empresarial. Os KPIs devem ser bem definidos, bem apresentados, criar expectativas e impulsionar ações. Além disso, é importante que os KPIs sejam apresentados em uma linguagem relevante para os negócios, usando taxas, proporções, porcentagens e médias em vez de números brutos. Os KPIs também devem fornecer contexto temporal e destacar mudanças em vez de apresentar tabelas de dados. Quando os KPIs são escolhidos com cuidado e alinhados aos objetivos estratégicos da empresa, eles podem ajudar a impulsionar ações críticas de negócios e melhorar o desempenho empresarial [1].

2.3.1 KPIs no mercado varejista

No mercado varejista, os KPIs desempenham um papel fundamental, especialmente para grandes varejistas online. Eles são usados para medir o desempenho em áreas críticas, como conversão e receita, e para avaliar o impacto de pequenas mudanças nas operações de varejo. Alguns exemplos de KPIs recomendados para varejistas online incluem taxas de conversão de pedidos e

compradores, valor médio do pedido, receita média por visita, custo médio por conversão e percentual de clientes satisfeitos e insatisfeitos [1].

Os KPIs são selecionados a partir de um conjunto de métricas relevantes e importantes para a empresa. As métricas são usadas para medir o desempenho em áreas específicas do negócio, enquanto os KPIs são usados para avaliar o desempenho geral da empresa em relação a seus objetivos estratégicos. As métricas são, portanto, um componente fundamental na construção de KPIs eficazes e relevantes [1].

Além disso, os KPIs são usados para avaliar o desempenho do varejista em áreas como atendimento ao cliente, fidelidade do cliente e eficácia da campanha de marketing. Alguns exemplos de KPIs recomendados para estrategistas de varejo de nível médio incluem tempo médio de resposta a consultas por e-mail, taxa de conversão de novos visitantes e percentual de receita de novos clientes [1].

2.4 Integração de Dados

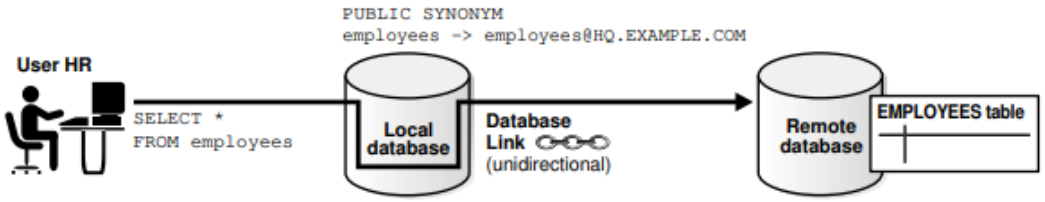
A Integração de Dados é um processo que envolve a combinação de dados de diferentes fontes, formatos e estruturas em um único sistema, a fim de fornecer uma visão unificada e coerente dos dados. Esse processo é essencial para empresas e organizações que precisam lidar com grandes volumes de dados de diferentes fontes, como bancos de dados, arquivos, planilhas, entre outros. A integração de dados permite que as empresas obtenham insights mais precisos e valiosos a partir de seus dados, além de melhorar a eficiência e a tomada de decisões. Existem várias técnicas e ferramentas disponíveis para a integração de dados, incluindo ETL (Extração, Transformação e Carga), ELT (Extração, Carga e Transformação), integração de dados em tempo real e integração de dados federada [6]. A extração pode ser incremental ou completa. No modelo de extração incremental de dados é um método de extração de dados em que apenas as alterações feitas nos sistemas de origem são extraídas em relação à extração anterior. O CDC (Change Data Capture) é um mecanismo que utiliza a extração incremental. Esse modelo é uma alternativa ao modelo de extração completa, que extrai todos os dados do sistema de origem, independentemente de terem sido alterados ou não. A extração incremental é uma técnica importante para garantir a

eficiência e a precisão do processo de ETL (Extract, Transform, Load) em um ambiente de data warehousing [9]. Por sua vez, o modelo de extração completa de dados é um método de extração de dados em que todos os dados brutos são extraídos de uma fonte de dados e transformados em um modelo de dados predefinido [7]. Ademais, este modelo é útil em situações em que os dados precisam ser limpos, transformados ou enriquecidos antes de serem integrados em um único repositório. Além disso, essa técnica permite que os dados sejam armazenados em um formato padronizado, o que facilita a análise e a tomada de decisões [6].

2.5 Database Link

Um Database Link (DBLINK) é um ponteiro que define um caminho de comunicação unidirecional de um servidor de Banco de Dados Oracle para outro servidor de banco de dados. A Figura 2 ilustra uma instância em que o usuário 'hr' está realizando o acesso à tabela 'funcionários' dentro de um banco de dados remoto, identificado globalmente como 'hq.example.com'. Notavelmente, a utilização de um sinônimo associado a 'funcionários' desempenha o papel de mascarar tanto a identidade quanto a localização do objeto pertencente ao esquema remoto [10].

Figura 2 – Ilustração do funcionamento de um Database Link.



Fonte: ORACLE. Oracle® Database Database Administrator's Guide, 2021. Figura extraída da página 524.

O uso de Database Links em um banco de dados Oracle oferece a vantagem de permitir que os usuários acessem objetos em um banco de dados remoto sem a necessidade de copiar ou duplicar os dados. Isso pode ser útil em situações em que os dados precisam ser compartilhados entre diferentes bancos de dados ou quando é necessário acessar dados em um banco de dados remoto para realizar uma tarefa específica. Além disso, o uso do Database Links permite que os usuários acessem

objetos em um banco de dados remoto sem precisar ser um usuário no banco de dados remoto [10].

2.6 Redo log

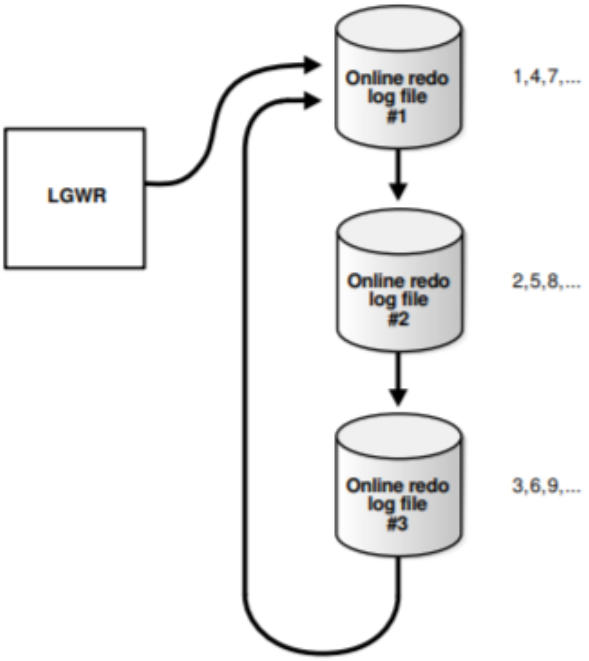
O Redo Log é uma estrutura crucial para operações de recuperação em bancos de dados Oracle. Ele consiste em dois ou mais arquivos pré-alocados que armazenam todas as alterações feitas no banco de dados à medida que ocorrem. Cada instância do Oracle Database tem um Redo Log associado para proteger o banco de dados em caso de falha da instância [10].

Os arquivos do Redo Log são preenchidos com registros de Redo, que são compostos por um grupo de vetores de mudança, cada um dos quais é uma descrição de uma mudança feita em um único bloco no banco de dados. Como demonstrado na Figura 3, esses registros de Redo são armazenados em buffer de forma circular no Redo Log Buffer da SGA e são gravados em um dos arquivos do Redo Log pelo processo de fundo do banco de dados Log Writer (LGWR) [10].

O Oracle Database usa apenas um arquivo de "redo log" por vez para armazenar registros de redo gravados no buffer de redo log. O arquivo de redo log no qual o LGWR está gravando ativamente é chamado de arquivo de redo log atual [10].

Uma troca de log é o ponto em que o banco de dados para de gravar em um arquivo de redo log e começa a gravar em outro. Normalmente, uma troca de log ocorre quando o arquivo de redo log atual está completamente preenchido e a gravação deve continuar para o próximo arquivo de redo log [10].

Figura 3 – Reutilização de arquivos de redo log por LGWR.



Fonte: ORACLE. Oracle® Database Database Administrator's Guide, 2021. Figura extraída da página 586

Os registros de Redo gravados no Redo Log podem ser usados para reconstruir todas as alterações feitas no banco de dados, incluindo os segmentos de undo. Portanto, o Redo Log também protege os dados de rollback. Quando o banco de dados é recuperado usando dados de Redo, ele lê os vetores de mudança nos registros de Redo e aplica as alterações aos blocos relevantes [10].

2.7 Qlik Replicate

Qlik Replicate é uma solução para implementar replicação de dados entre vários endpoints. Com o Qlik Replicate, é possível carregar dados de forma rápida em armazéns de dados operacionais, criar cópias de endpoints de produção e distribuir dados entre endpoints. O Qlik é projetado para escalar e dar suporte a cenários de replicação de dados empresariais em grande escala, com uma arquitetura escalável de vários servidores, várias tarefas e várias threads. O Qlik Replicate também oferece recursos de captura de dados de alteração (CDC) em tempo real, garantindo a integridade transacional no endpoint de destino. Ademais, é possível replicar dados entre fontes de dados heterogêneas e homogêneas,

incluindo bancos de dados relacionais, como Oracle, Microsoft SQL Server e IBM DB2. Ele também fornece aos usuários visibilidade instantânea sobre tarefas atuais e históricas, status, desempenho e informações de uso de recursos por meio de um console baseado na web [11].

A arquitetura do sistema Qlik Replicate é composta por vários componentes que trabalham juntos para fornecer replicação de dados, incluindo terminais, agentes, um console, um servidor de replicação, transformação e CDC.

Endpoints são repositórios de dados que podem ser usados como origem e/ou destino em uma tarefa do Qlik Replicate. Conforme explica, "O termo 'endpoint' é usado genericamente ao longo deste texto para se referir a um repositório de dados que pode ser usado como origem e/ou destino em uma tarefa do Qlik Replicate". O Agente são processos instalados em cada endpoint e são responsáveis pela leitura e gravação de dados nos endpoints. O console é uma interface baseada na Web que permite aos usuários configurar e monitorar tarefas de replicação. De acordo com, "Replicate permite configurar e monitorar tarefas por meio de um console baseado na web". O servidor de replicação gerencia tarefas de replicação e coordena a comunicação entre agentes. Conforme explica, “O servidor de replicação gerencia as tarefas de replicação e coordena a comunicação entre os agentes”.

3 METODOLOGIA

3.1 Comparação de Cenários de Teste

A metodologia adotada nesta avaliação envolve uma abordagem comparativa entre dois cenários de teste distintos para analisar a replicação e processamento de dados no contexto da geração de KPIs. A relevância dos KPIs nesta avaliação está relacionada à necessidade de disponibilizar os dados de maneira mais eficiente, sem impactar negativamente na operação da organização. O primeiro cenário utiliza Database Links e Procedures de banco de dados Oracle para a integração, enquanto o segundo cenário emprega a ferramenta de replicação de dados Qlik Replicate. A análise detalhada destes cenários permite uma avaliação das duas abordagens de integração de dados, visando otimizar a capacidade da organização de medir e melhorar seu desempenho sem comprometer sua operação.

3.2 Domínio do problema

Dentro do contexto da empresa de varejo em questão, foi delineado áreas-chave para monitoramento e análise, estas áreas incluem Vendas, que abrange a performance de vendas em geral; E-commerce, que engloba a performance das operações online; Estoque, para acompanhar os níveis de disponibilidade de produtos; Marketing, que avalia o impacto das estratégias de marketing; Atendimento ao Cliente, para mensurar a satisfação dos consumidores; Operações de Loja, abrangendo a eficiência das lojas físicas; e Recursos Humanos, para monitorar a gestão de pessoal. Estes KPIs são estruturados em três níveis: Geral, que fornece uma visão holística da organização; Filial, que se concentra em avaliar o desempenho de filiais específicas; e Hierárquico, que permite uma análise granular, avaliando colaboradores em níveis hierárquicos. Ao todo, foram identificadas 163 métricas distintas que são a fonte de dados desses KPIs, resultando em um conjunto de 55 indicadores-chave de desempenho.

3.3 Coleta de dados

No processo de coleta de dados, foi adotada uma abordagem dividida em duas categorias: quantitativa e qualitativa. Na coleta quantitativa, focou-se na exploração dos logs estatísticos da operação de geração de KPIs ao longo de um período de 30 dias, para obter dados mensuráveis relacionados ao desempenho, tempo de execução e incidência de erros durante as operações de integração e geração de KPIs. Paralelamente, na condução desta avaliação, adotou-se a estratégia de coleta de dados qualitativos, os quais foram coletados por um especialista interno no projeto de desenvolvimento de integração de dados para o ambiente de banco de dados centralizado na organização. A estratégia de coleta de dados qualitativos obtida por esse especialista permite obter conhecimentos que não seriam facilmente conclusivos para alguém de fora da organização [13]. Com isso, foram realizadas entrevistas com membros da comunidade estudada e analisados processos internos e sistemas da empresa, proporcionando uma visão abrangente do projeto e dos resultados obtidos nos dois cenários de teste. A combinação de métodos de coleta de dados qualitativos permite uma compreensão mais completa e profunda do fenômeno investigado, enquanto os métodos quantitativos são úteis para medir e quantificar variáveis e relações entre elas, os métodos qualitativos são úteis para explorar e compreender as experiências, percepções e significados dos participantes [12].

3.4 Análise de dados

Na subseção de análise de dados da metodologia, será realizada uma abordagem combinada para avaliar abrangentemente os resultados dos cenários de teste. Para os dados quantitativos, empregaremos uma abordagem descritiva, que nos permitirá quantificar variáveis específicas relacionadas aos resultados dos testes, incluindo tempo de execução e taxas de erro. Isso fornecerá uma base sólida para avaliar de forma objetiva o desempenho das abordagens de integração de dados.

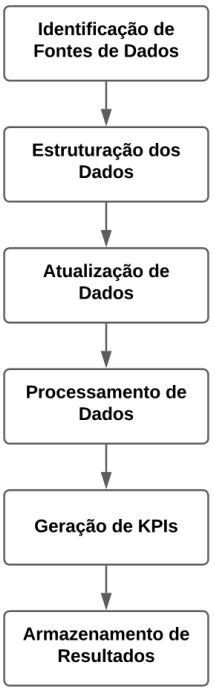
Além disso, para os dados qualitativos, adotaremos uma abordagem de análise qualitativa. Aqui, nossa ênfase será na exploração e compreensão das experiências e percepções dos participantes envolvidos nos cenários de teste. Isso

será realizado por meio de entrevistas, observação participante e análise de sistemas da empresa, permitindo-nos capturar informações subjetivas e nuances que complementam a análise quantitativa. A combinação dessas abordagens qualitativas e quantitativas garantirá uma compreensão holística dos resultados e das perspectivas envolvidas nos cenários de teste, enriquecendo nossa avaliação geral do projeto.

3.5 Regra de Negócio

Esta seção aborda a metodologia adotada no estudo, que se alinha à abordagem da regra de negócio centralizada na integração, processamento e geração de Indicadores-Chave de Desempenho (KPIs) a partir de diversas fontes de dados. Isso inclui bancos de dados operacionais de Recursos Humanos e informações operacionais de lojas. A base da metodologia é capacitar a tomada de decisões informadas e aprimorar o sucesso operacional através de uma análise abrangente dos KPIs. O processo metodológico compreende etapas bem definidas, desde a identificação das fontes de dados relevantes e a estruturação dos dados até a replicação desses dados para o ambiente de análise, o processamento necessário para o cálculo dos KPIs, a geração dos indicadores e o subsequente armazenamento dos resultados para análises futuras. O principal objetivo é fornecer uma compreensão do desempenho organizacional por meio de métricas tangíveis e contextualizadas.

Figura 4 – Processo metodológico da replicação de dados para geração de KPI.



Fonte: Elaborado pelo autor

- **Identificação de Fontes de Dados:** O processo envolve a identificação das fontes de dados relevantes, como bancos de dados operacionais do RH e dados operacionais de lojas;
- **Estruturação dos Dados:** Os dados são estruturados e organizados de acordo com as necessidades de geração de KPIs. Isso inclui a seleção de tabelas específicas e a categorização de tabelas com base em grupos de contexto e outras características relevantes. O Quadro 1 apresenta a categorização de tabelas do RH;

Quadro 1 – Categorias das tabelas do RH.

Grupo	Qtde de Tabelas	Descrição
Gestão por Desempenho	18	Plataforma Gestão de desempenho organizacional, abrangendo relação de cargo, métrica e KPI
Avaliação de Desempenho	5	Processos de avaliação de desempenho dos colaboradores
Detalhamento Cargo	5	Informações sobre os diferentes cargos e funções
Detalhamento Colaboradores	1	Informações sobre os colaboradores da empresa
Imagem Colaboradores	1	Imagens dos colaboradores da organização

Fonte: Elaborado pelo Autor

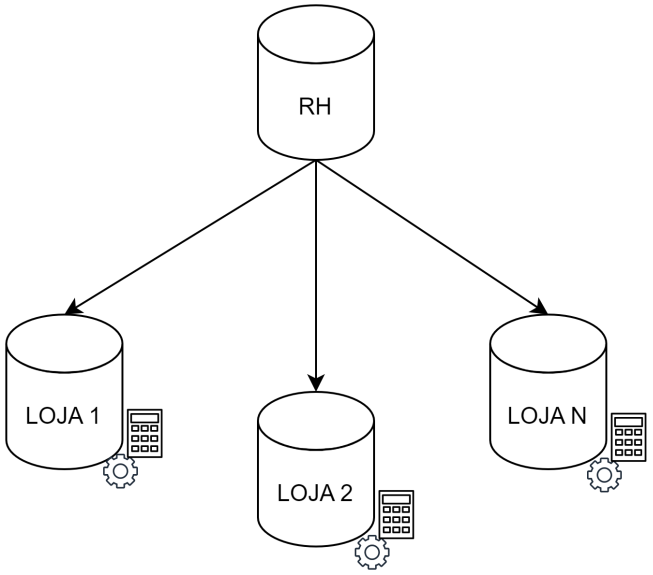
- **Atualização de Dados:** os dados são replicados dos endpoints de origem para os endpoints que irão realizar o processamento dos dados. Isso pode envolver diferentes abordagens, como a replicação de dados via Database Link (no cenário 1) ou o uso de ferramentas de replicação de dados (no cenário 2);
- **Processamento de Dados:** após a replicação, os dados são processados para calcular os KPIs de desempenho. Isso pode incluir cálculos complexos, agregações e outras transformações necessárias para gerar os indicadores desejados;
- **Geração de KPIs:** com base nos dados processados, os KPIs são calculados e gerados. Isso envolve a aplicação de fórmulas e métricas específicas para avaliar o desempenho de diferentes aspectos do negócio;
- **Armazenamento de Resultados:** os resultados dos KPIs calculados são armazenados no BD centralizado ou na plataforma de Gestão de Desempenho para análise e consulta posterior.

4 CENÁRIOS DE TESTE

4.1 Database Links e Procedures (abordagem descentralizada)

No cenário de teste 1, o processo descentralizado de replicação de dados é delineado pela replicação das tabelas do banco de dados de Recursos Humanos (RH) para as tabelas operacionais das lojas (Figura 5). Esse processo abrange a transferência estruturada de dados, garantindo que as informações relevantes sejam disponibilizadas nos bancos de dados de destino associados às respectivas lojas. Após a realização da replicação, o próximo passo envolve a realização do cálculo das Indicadores-Chave de Desempenho (KPIs) diretamente nos bancos de dados de destino.

Figura 5 – Fluxo de replicação de dados no cenário 1.



Fonte: Elaborado pelo autor

4.1.1 Implementação técnica

Na primeira etapa do processo de replicação de dados, a identificação dos Database Links ativos é realizada através da execução de uma consulta SQL que

retorna os links de banco de dados das lojas que estão ativos. Esses links são responsáveis por estabelecer uma conexão com os endpoints distribuídos para a replicação de dados, garantindo uma disponibilização uniforme dos dados entre os diferentes bancos de dados das lojas.

A procedure exe_remote_sql cria e controla um cursor anônimo para executar o SQL remotamente. Um cursor anônimo é uma estrutura de programação que permite a manipulação e iteração sobre os resultados de uma consulta SQL de forma dinâmica. A característica "anônima" refere-se ao fato de que esse tipo de cursor não está explicitamente associado a uma consulta SQL nomeada e predefinida. Em vez disso, ele é criado dinamicamente no código PL/SQL, o que o torna flexível para executar consultas variáveis em tempo de execução. A Figura 6 mostra a Procedure que executa remotamente a instrução SQL para a filial 2.

Figura 6 – Procedure que executa remotamente um código SQL.

```
PROCEDURE exe_remote_sql_02 (
    p_sql    IN    VARCHAR2
)
AS
    exec_cursor       INTEGER DEFAULT dbms_sql.open_cursor@IND.FIL02.COM ;
    rows_processed    NUMBER  DEFAULT 0;
BEGIN
    dbms_sql.parse@IND.FIL02.COM  (exec_cursor, p_sql, dbms_sql.native);
    rows_processed := dbms_sql.execute@IND.FIL02.COM  (exec_cursor);
    dbms_sql.close_cursor@IND.FIL02.COM (exec_cursor);
END exe_remote_sql_02;
```

Fonte: Elaborado pelo autor

Além disso, o código realiza uma busca nas informações das colunas de uma tabela original para criar uma nova tabela com a mesma estrutura em bancos de destino (Figura 7). A nova tabela resultante terá a mesma estrutura da tabela original, assegurando a replicação precisa das colunas e seus tipos nos bancos de destino. Também é executada uma instrução INSERT INTO para copiar os dados da tabela original para a nova tabela nos bancos de destino, garantindo a transferência de dados.

Figura 7 – Código SQL de preparação de criação de uma tabela.

```
SELECT listagg(Column_Name||' '||Data_Type||
       CASE WHEN Data_Type='VARCHAR2'
        THEN '('||Data_Length||')'
       END, ',') WITHIN GROUP (ORDER BY COLUMN_ID)
  INTO l_sql
  FROM All_Tab_Columns
 WHERE Table_Name = T.Nm_Tabela;
  l_sql := 'CREATE TABLE '||T.Nm_Tabela||'('||l_sql||')';
  c_tw_processa_indicadores.exe_remote_sql (l_sql, i.dblink);
COMMIT;
  l_sql := 'INSERT INTO '||T.Nm_Tabela||' SELECT * FROM '||T.Nm_Tabela||'@RH.COM';
  c_tw_processa_indicadores.exe_remote_sql (l_sql, i.dblink);
COMMIT;
```

Fonte: Elaborado pelo autor

4.1.2 Desafio 1: concorrência com a operação das lojas

A concorrência com a operação diária das lojas, incluindo operações de caixa, monitoramento de níveis de estoque, e pedido de compra e venda de produtos, surge como o principal desafio no processo de replicação de dados por meio de database links em bancos de dados descentralizados. Isso se deve ao fato de que, para garantir a conformidade com os princípios ACID de um banco de dados relacional, as tabelas são frequentemente bloqueadas temporariamente para realizar operações varejistas, o que resulta em bloqueios temporários. Por outro lado, consultas SQL mais pesadas também afetam a disponibilidade das tabelas para a operação diária das lojas, prejudicando o principal objetivo da rede varejista, que é proporcionar o melhor atendimento possível aos clientes. Esse equilíbrio delicado entre a replicação de dados e a operação cotidiana das lojas representa um desafio técnico e organizacional significativo.

4.1.3 Desafio 2: baixa tolerância a erros

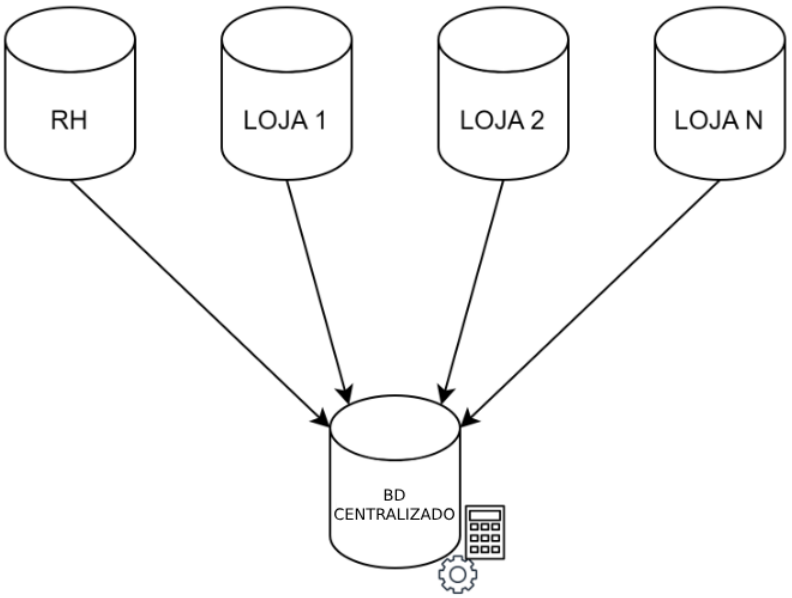
Ademais, no processo de replicação de dados por meio de database links para bancos de dados distribuídos, um dos desafios técnicos mais significativos é a baixa tolerância a erros. Isso se deve ao fato de que qualquer interrupção momentânea na conexão com o endpoint de destino pode resultar em inconsistências na integração de dados para o endpoint afetado. Com isso, a

integridade dos dados replicados é comprometida devido a quedas de conexão temporárias. Essa vulnerabilidade exige o esforço humano de reprocessamento da integração dos dados, dado que a estratégia de tratamento de falhas não é robusta o suficiente para garantir a confiabilidade e a consistência da replicação de dados em ambientes descentralizados.

4.2 Ferramenta Qlik Replicate (abordagem centralizada)

No cenário de teste 2, o processo de replicação de dados é caracterizado pela replicação das tabelas presentes nos bancos de dados de Recursos Humanos (RH) e nos bancos de dados operacionais das lojas para o ambiente do Banco de Dados centralizado (Figura 8). Esse procedimento envolve a coleta e integração das informações de diversas fontes, centralizando os dados em um local unificado para posterior análise. Após a conclusão da replicação, a etapa subsequente consiste em realizar o cálculo das KPIs no próprio banco de dados centralizado

Figura 8 – Fluxo de replicação de dados no cenário 2.



Fonte: Elaborado pelo autor

4.2.1 Implementação técnica

Na criação das tarefas de replicação, foram consideradas as particularidades das diferentes fontes de dados. Para tabelas do RH, a estrutura foi baseada na definição do owner e do grupo correspondente, onde 18 tabelas foram definidas no modelo de replicação incremental e 12 no modelo de carga total. Já para as tabelas operacionais das lojas, a abordagem levou em conta o owner, a origem das tabelas e o modo de replicação, com 420 tabelas definidas no modelo de replicação incremental e 10 no modelo de carga total. No processo de replicação de dados de carga total, cada tarefa pode carregar até cinco tabelas simultaneamente, otimizando o tempo de replicação e permitindo cálculos de KPIs antecipados. Por outro lado, nas tarefas de replicação de dados incrementais, foi adotado o modo Batch Optimized Apply, que agrupa transações de forma eficiente, priorizando o desempenho. Embora isso possa resultar em breves lapsos temporários na integridade transacional, para o ambiente do banco de dados centralizado, essa abordagem é adotada de maneira válida, considerando que a prioridade é otimizar o desempenho do processo de replicação.

Na Figura 9, apresentam-se as estatísticas de uma das tarefas do modelo carga total, revelando que cerca de 11 milhões de registros foram replicados para o endpoint de destino em aproximadamente 10 minutos.

Figura 9 – Estatísticas de uma tarefa do modelo carga total

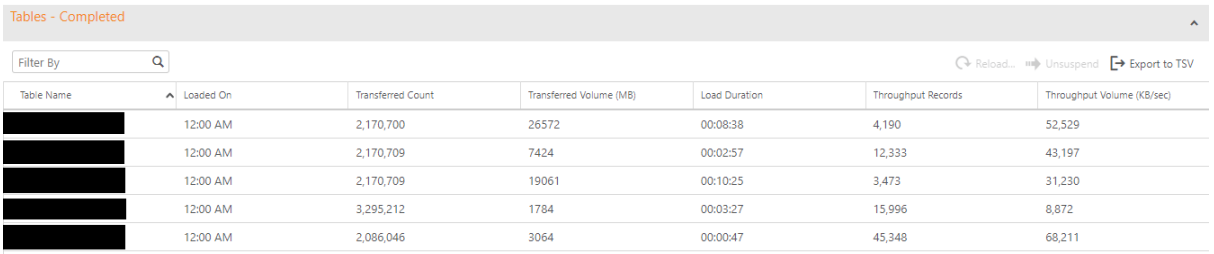


Table Name	Loaded On	Transferred Count	Transferred Volume (MB)	Load Duration	Throughput Records	Throughput Volume (KB/sec)
[redacted]	12:00 AM	2,170,700	26572	00:08:38	4,190	52,529
[redacted]	12:00 AM	2,170,709	7424	00:02:57	12,333	43,197
[redacted]	12:00 AM	2,170,709	19061	00:10:25	3,473	31,230
[redacted]	12:00 AM	3,295,212	1784	00:03:27	15,996	8,872
[redacted]	12:00 AM	2,086,046	3064	00:00:47	45,348	68,211

Fonte: Elaborado pelo autor

No caso das tarefas do modo incremental, o Qlik Replicate opera de forma mais ágil. Ele monitora continuamente o redo log do endpoint de origem, verificando as mudanças nos dados. A leitura desses logs é realizada a cada 5 segundos, permitindo que as atualizações sejam capturadas quase em tempo real e replicadas

para os bancos de destino de forma eficiente. Na Figura 10, é exibido o log da tarefa de replicação do modelo incremental, que indica o início do processamento de um redo log online.

Figura 10 – Log de execução de uma tarefa do modelo incremental

```
[SOURCE_CAPTURE  ]I:  Start processing archived Redo log sequence 56084 thread 1 name /u03/oradata/fctam/archive/1_56084_1109889353.dbf  (oradcdc_redo.c:914)
[SOURCE_CAPTURE  ]I:  Start processing online Redo log sequence 56085 thread 1 name /u03/oradata/fctam/onlinelog/redo01a.log  (oradcdc_redo.c:914)
[SORTER          ]I:  Task is running  (sorter.c:716)
[SORTER          ]I:  Task is running  (sorter.c:716)
[SOURCE_CAPTURE  ]I:  Start processing online Redo log sequence 56086 thread 1 name /u03/oradata/fctam/onlinelog/redo03a.log  (oradcdc_redo.c:914)
```

Fonte: Elaborado pelo autor

4.2.2 Desafio 1: Complexidade do Ambiente de Negócios

Dada a complexidade do ambiente de negócios no setor varejista, que engloba diversas dimensões, como logística, compras, vendas e atendimento ao cliente, a implantação da integração de dados na abordagem centralizada apresentou-se como um obstáculo devido a ausência de metadados de negócios, que fornecem informações sobre o significado e o contexto dos dados no âmbito empresarial. Para tornar possível a integração e a geração de métricas abrangendo essas diversas áreas, o primeiro passo foi a realização de um mapeamento das tabelas utilizadas em cada dimensão, incluindo suas definições de regras de transformação e critérios de filtragem de dados.

4.2.3 Desafio 2: Volumetria de dados

A replicação de dados no ambiente centralizado enfrentou um desafio relacionado à volumetria dos dados. Ao observar exclusivamente as tabelas utilizadas no processamento dos KPIs, os bancos de dados no ambiente distribuído tinham uma média de 322 milhões de registros (desvio padrão de 114 milhões), equivalente a cerca de 74 gigabytes de dados. Entretanto, ao centralizar esses dados em um único banco de dados, a volumetria saltou para cerca de 1,14 bilhão de registros, o que equivale a aproximadamente 328 gigabytes de dados. Isso resultou em um desafio significativo em termos de desempenho, especialmente no que diz respeito ao tempo necessário para a execução de consultas e cálculos. No

entanto, vale ressaltar que, apesar do salto na volumetria no banco de dados centralizado, a disponibilização desses dados é importante para a exploração dos analistas de dados e facilitar a implementação de novas KPIs. Isso evita trabalho posterior de sincronia ao adicionar novas colunas ou novos registros que não contemplam o banco centralizado.

O Quadro 2 apresenta informações extraídas do "explain plan" do Oracle, que é uma ferramenta que fornece uma representação do plano de execução de uma consulta SQL. Com isso, ele descreve como o Oracle planeja realizar a consulta, incluindo a ordem das operações, o uso de índices, tabelas envolvidas e estimativas de custo relativas, onde o custo representa o uso estimado de recursos para esse plano. Quanto menor o custo, mais eficiente se espera que o plano seja. O modelo de custo do otimizador leva em conta os recursos de IO (Input e Output), CPU e rede que serão usados pela consulta [16].

Quadro 2 – Comparação de Desempenho entre BD Centralizado e BD Distribuído

Métrica	**Tempo de execução em segundos**		**Custo**	
	BD centralizado	**BD distribuído**	**BD centralizado**	**BD distribuído**
M1	50	9	47,4 milhões	185 mil
M2	74	4	3,1 milhões	95 mil
M3	15	5	1,7 milhões	133 mil
M4	66	26	1,3 milhões	40 mil
M5	65	16	100 mil	27 mil

Fonte: Elaborado pelo autor

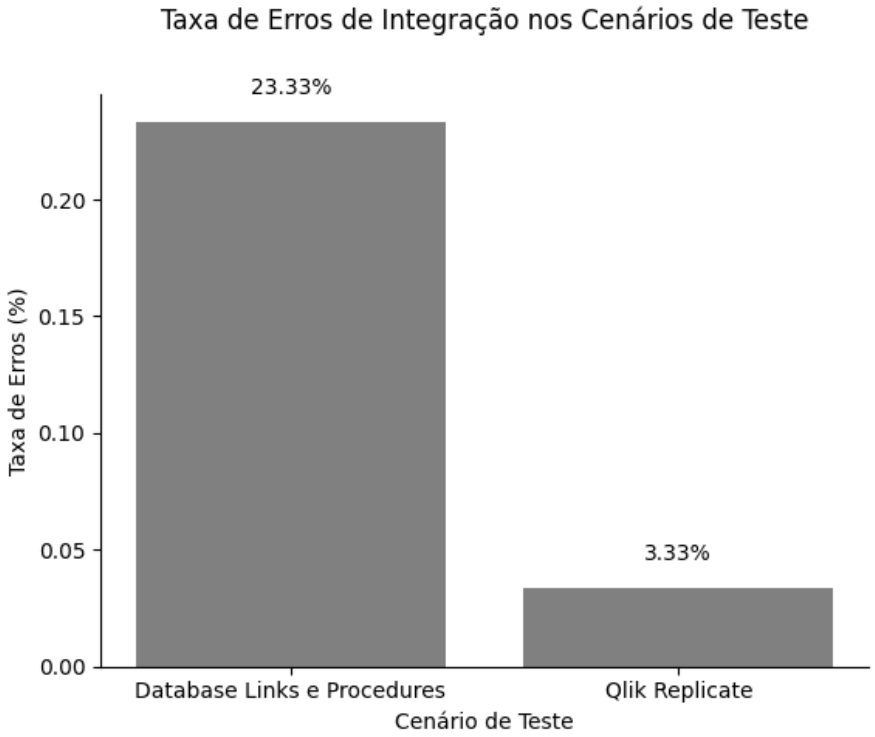
Para enfrentar essa questão, foi empreendido um esforço significativo, culminando no desenvolvimento de um total de 67 índices no banco de dados centralizado. Essa iniciativa tinha como objetivo otimizar o acesso aos dados e acelerar as operações de consulta. No resultado desse esforço, considerando apenas os índices de tabelas que estão contidas no processo de replicação de dados, houve um aumento de aproximadamente 10% no número de índices nas tabelas citadas. Além disso, algumas consultas tiveram que ser refatoradas, adaptando-as ao ambiente centralizado e garantindo um desempenho adequado.

5 RESULTADOS

5.1 Taxa de Erro de Integração

Os resultados, obtidos a partir de uma amostra de 30 dias de observações, na análise da taxa de erros no processo de integração de dados entre os dois cenários (Gráfico 1), o cenário "Database Links e Procedures" apresentou uma maior incidência de erros em comparação com o cenário "Qlik Replicate". Destaca-se que a maioria dos erros no cenário "Database Links e Procedures" foi atribuída a problemas relacionados ao fim do arquivo no canal de comunicação. Este tipo de erro é comum nesse cenário devido a lapsos de lentidão no tráfego de rede, decorrente do sistema de banco de dados distribuídos entre as lojas. Essa lentidão ocasionalmente afetou a integridade da transmissão de dados, resultando em erros. Por outro lado, no cenário "Qlik Replicate", houve apenas uma ocorrência de erro. Isso pode ser atribuído à capacidade da ferramenta Qlik Replicate de restabelecer a quedas de conexões que ocorrem no dia-a-dia, garantindo uma maior robustez na integração de dados.

Gráfico 1 – Taxa de erros de integração nos cenários de teste



Fonte: Elaborado pelo autor

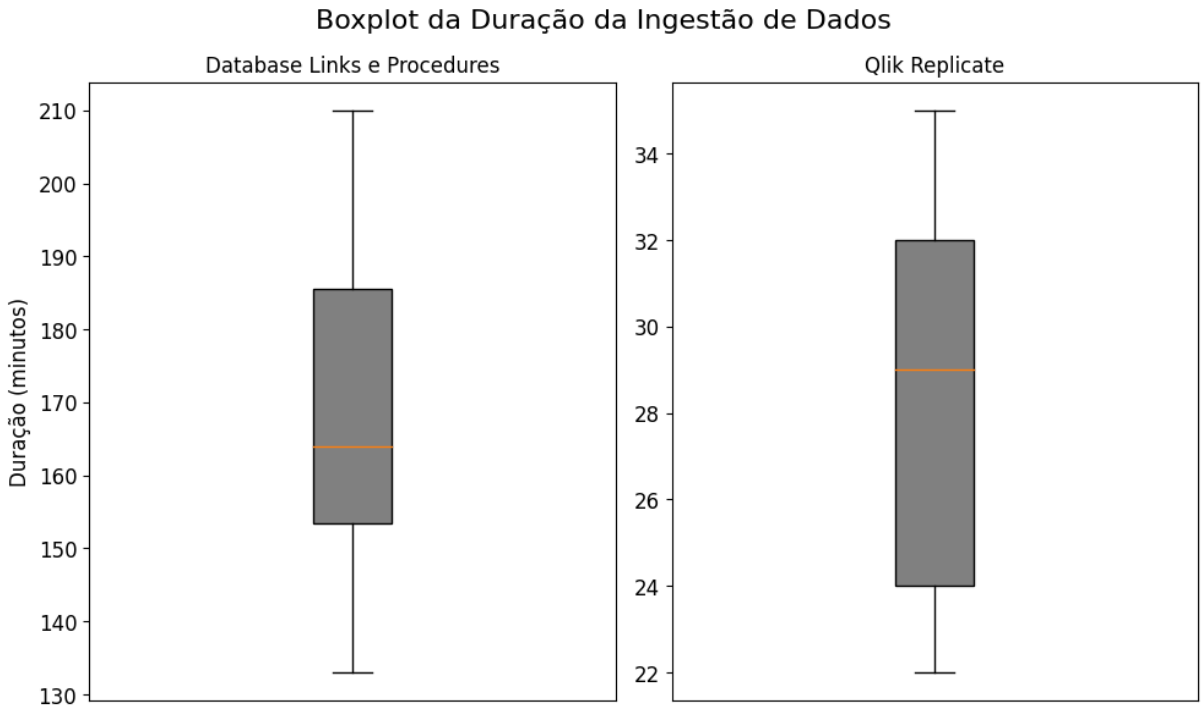
5.2 Duração da ingestão de dados

Os resultados obtidos nos dois cenários de teste, "Database Links e Procedures" e "Qlik Replicate", a partir de uma amostra de 30 dias de observações, revelam diferenças em relação à métrica de "Duração da ingestão de dados" (Gráfico 2).

No cenário "Database Links e Procedures", com base na leitura dos logs de execução, é possível analisar que a duração da ingestão de dados foi de em média 169,1 minutos, com um desvio padrão de 21,8 minutos. Isso indica que, em média, o processo de ingestão de dados nesse cenário leva mais tempo, e os tempos variam consideravelmente ao redor dessa média, com uma dispersão relativamente alta. Devido ao modelo de datacenters distribuídos adotado pela empresa varejista, essa dispersão relativamente alta pode ser atribuída à imprevisibilidade da situação do tráfego de rede na comunicação dos servidores de banco de dados. Os quartis também mostram uma ampla variação, desde o mínimo de 133 minutos até o máximo de 210 minutos.

Por outro lado, no cenário "Qlik Replicate", a média da duração da ingestão de dados no modelo carga total foi substancialmente menor, registrando 28,8 minutos, com um desvio padrão menor de 4,6 minutos. Isso sugere que a ingestão de dados no cenário "Qlik Replicate" é mais rápida em média e mais consistente, com tempos de ingestão menos variáveis. Os quartis também indicam uma menor dispersão, com um mínimo de 22 minutos e um máximo de 35 minutos.

Gráfico 2 – Duração da ingestão de dados



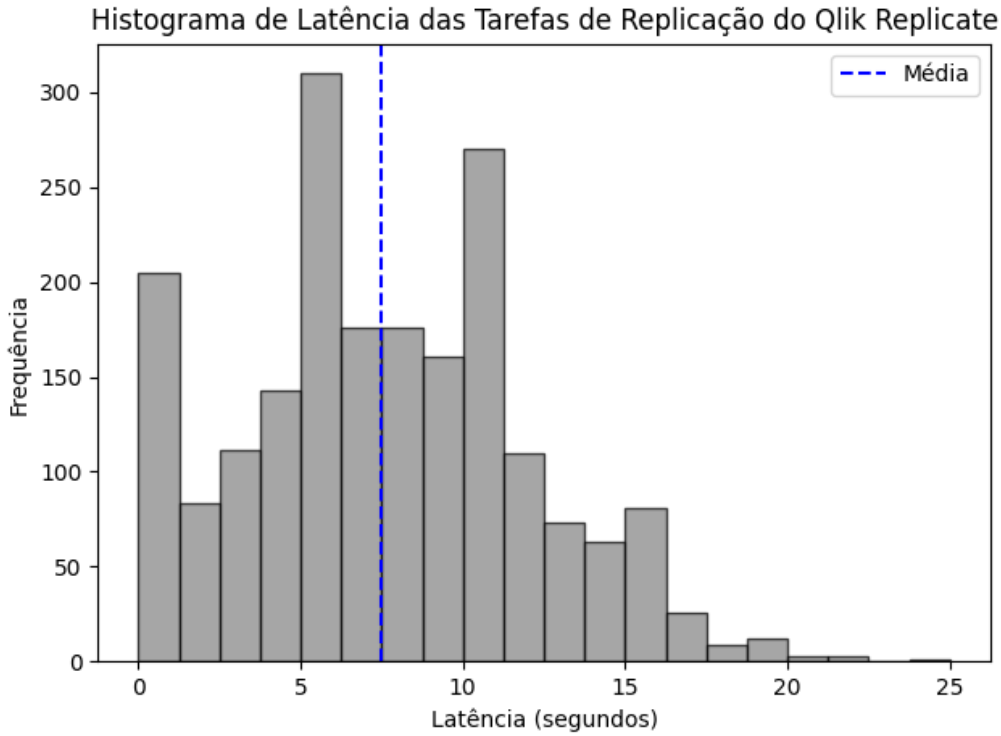
Fonte: Elaborado pelo autor

5.3 Latência das replicações do modelo incremental no Qlik Replicate

Com base no resultado da amostra de latência das tarefas de replicação extraídos ao longo de um período de 7 dias, com granularidade de cinco minutos (Gráfico 3), o desempenho da latência das tarefas de replicação no Qlik Replicate pode ser avaliado de maneira positiva. A média de latência registrada foi de 7 segundos, indicando um tempo de conclusão das tarefas de replicação que se mostra adequado para garantir a disponibilidade oportuna dos dados essenciais para análise e criação de KPIs. Além disso, os quartis revelam uma distribuição equilibrada dos dados, com o primeiro quartil (25%) em 4 segundos, a mediana (50%) em 7 segundos e o terceiro quartil (75%) em 10 segundos. Essa distribuição sugere que a maioria das tarefas de replicação é realizada em tempo aceitável, com poucas ocorrências de latências extremamente altas. No geral, esses resultados indicam que o Qlik Replicate mantém um desempenho estável e previsível na latência das tarefas de replicação ao longo do período avaliado de sete dias. É importante ressaltar que essa métrica de latência das tarefas de replicação não é

aplicável ao cenário 1, onde se utiliza Database Links e Procedures, uma vez que não há a implementação do modo incremental nesse cenário.

Gráfico 3 – Histograma de latência das tarefas de replicação do Qlik Replicate



Fonte: Elaborado pelo autor

5.4 Tempo de processamento dos KPIs

No contexto dos cenários de teste para o processamento de KPIs, a partir de uma amostra de 30 dias de observações, os resultados das métricas revelam diferenças significativas entre o Cenário 1, que representa o processamento nos bancos de dados operacionais da loja, e o Cenário 2, que representa o processamento no banco de dados centralizado (Gráfico 4).

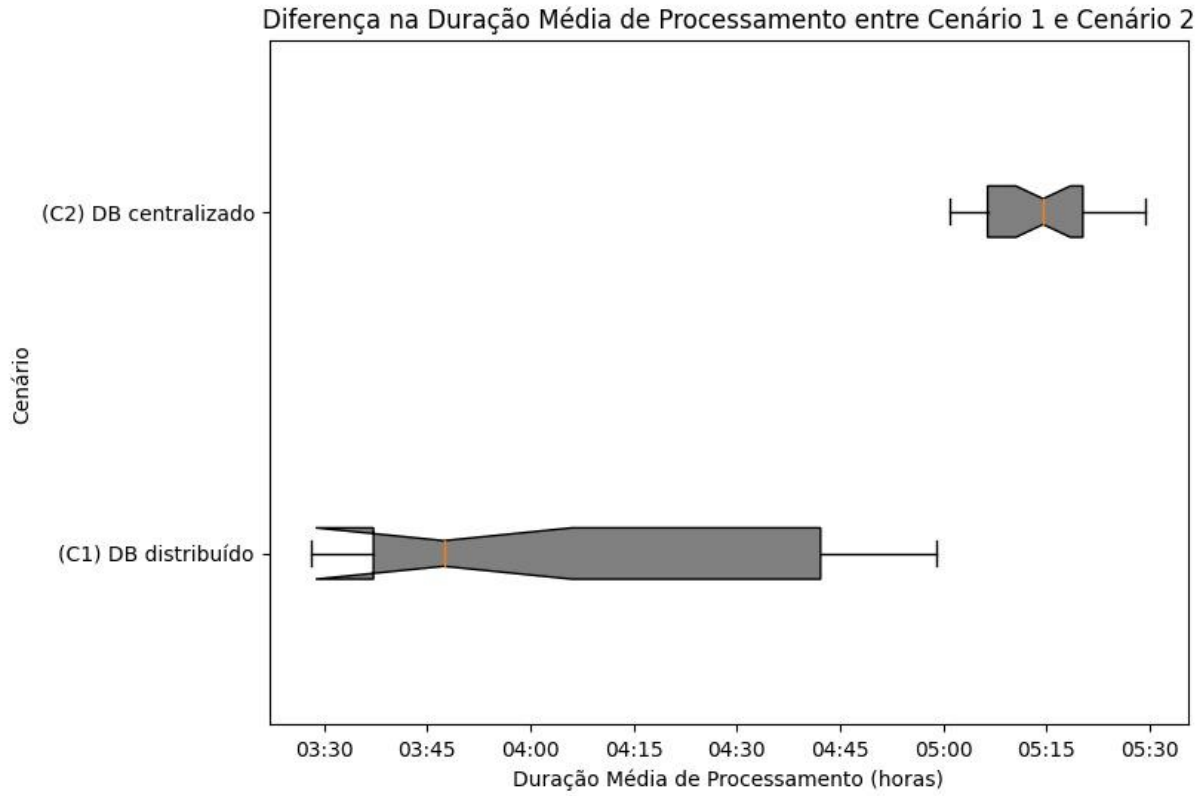
Primeiramente, é notável que o Cenário 1 apresenta um tempo médio de processamento mais rápido, com uma média de 04 horas e 06 minutos. No entanto, vale observar que o desvio padrão do tempo de processamento é relativamente alto, atingindo 30 minutos. Uma possível explicação para essa variabilidade no tempo de processamento do Cenário é que o processamento das KPIs nos bancos de dados operacionais pode entrar em conflito com a utilização das tabelas para o

funcionamento diário das lojas, resultando em flutuações no tempo de processamento.

Por outro lado, o Cenário 2 apresenta um tempo médio de processamento ligeiramente mais longo, com uma média de 5 horas e 12 minutos. No entanto, o destaque aqui é o desvio padrão notavelmente menor, que é de apenas 6 minutos. Isso indica que o tempo de processamento no Cenário 2 é mais previsível. Essa consistência pode ser atribuída ao fato de que o banco de dados centralizado foi projetado para atender aos requisitos de um banco de dados voltado para análise, proporcionando um ambiente mais estável e menos sujeito a flutuações.

Portanto, apesar do Cenário 1 ser mais rápido em termos de tempo médio de processamento, sua alta variabilidade pode resultar em desafios na previsibilidade do desempenho. Em contrapartida, o Cenário 2, com seu tempo de processamento ligeiramente mais longo, oferece uma consistência valiosa, tornando-o uma escolha sólida para as operações de análise de dados no banco de dados centralizado.

Gráfico 4 – Boxplot da duração média de processamento entre Cenário 1 e Cenário 2



Fonte: Elaborado pelo autor

5.5 Eficiência do Gerenciamento de Mudanças no Segundo Cenário de Teste

O gerenciamento de mudanças no segundo cenário se mostrou mais eficaz. Isso pode ser atribuído ao fato de que o processamento dos KPIs é realizado em uma única base de dados. Essa abordagem proporcionou uma agilidade significativa quando se trata de correções de erros e implementações de novos indicadores. A centralização dos dados e do processo de geração de KPIs no segundo cenário contribuiu para uma resposta mais rápida às necessidades em constante evolução da empresa, demonstrando sua eficácia no contexto da integração de dados em uma abordagem centralizada.

5.6 Privacidade de Dados

Foi observado um aprimoramento significativo no controle de privacidade de dados no contexto do segundo cenário de teste. Isso se deve ao fato de que dados críticos do departamento de Recursos Humanos não estão mais sendo replicados para vários datacenters distintos, o que reduziu os pontos potenciais de acesso indevido, bem como a exposição a ataques externos. Esse refinamento na gestão de dados sensíveis mostra a importância da integração de dados em um banco de dados centralizado para garantir a segurança e a proteção das informações críticas da empresa.

5.7 Análises e Eficiência Operacional

O processamento centralizado transformou a maneira como as métricas são geradas e as análises são conduzidas nos dias de hoje. Anteriormente, muitas dessas observações eram custosas e complexas, exigindo recursos significativos para coletar e integrar dados dispersos. No entanto, com essa nova abordagem, abriram-se portas para a realização de análises que antes eram consideradas desafiadoras.

Um dos benefícios mais evidentes é a melhoria nos processos que envolvem a operação das várias filiais. Essa é uma vantagem que não pode ser subestimada. A capacidade dos bancos distribuídos de performar de forma mais eficiente as operações dessas unidades torna-se uma realidade, uma vez que os dados agora

estão centralizados e acessíveis em um único local. Dessa forma, essa transformação possibilita não apenas otimizar a tomada de decisões, mas também permite uma alocação mais eficiente de recursos com base em dados sólidos e integrados.

6 CONCLUSÃO

A utilização de KPIs construídos a partir dos dados integrados no banco de dados centralizado traz benefícios distintos em comparação com o método utilizando Database Links. Em primeiro lugar, a centralização dos dados no banco de dados centralizado permite que os KPIs sejam construídos a partir de uma fonte única e confiável, o que garante maior precisão e consistência nas análises. Além disso, a integração dos dados de diferentes áreas da empresa possibilita uma visão mais abrangente e integrada do desempenho, permitindo identificar tendências e padrões que seriam difíceis de detectar com estrutura distribuída de banco de dados.

Ademais, a utilização de KPIs construídos a partir da abordagem centralizada permite uma análise mais granular do desempenho, possibilitando identificar gargalos e oportunidades de melhoria em áreas específicas da empresa. Isso permite uma alocação mais eficiente de recursos e uma tomada de decisão mais estratégica. É válido mencionar que a integração dos dados no banco de dados centralizado também contribui para uma maior segurança e privacidade dos dados, uma vez que reduz a exposição a ataques externos e minimiza os pontos potenciais de acesso indevido.

Os desafios técnicos e organizacionais são uma preocupação comum em ambos os cenários de teste avaliados neste estudo de caso. No primeiro cenário, que utiliza Database Links e Procedures de banco de dados Oracle para a integração de dados, o principal desafio é a concorrência com a operação diária das lojas. Isso ocorre porque a replicação de dados por meio de database links em bancos de dados distribuídos pode afetar o desempenho do sistema e, consequentemente, prejudicar a operação das lojas.

No segundo cenário, que emprega a ferramenta de replicação de dados Qlik Replicate, o principal desafio técnico está relacionado à ausência de metadados de negócios, que são necessárias para o mapeamento das regras de filtro e transformação. Além disso, a volumetria de dados aumentou consideravelmente ao centralizar os dados, demandando estratégias de otimização, como índices eficientes e distribuição de carga, para garantir o desempenho adequado nas consultas e cálculos dos KPIs.

Além dos desafios técnicos, é importante destacar que ambos os cenários de teste enfrentam desafios organizacionais. No entanto, no primeiro cenário, os

desafios organizacionais são mais críticos devido ao impacto direto no atendimento ao cliente e ao funcionamento dos bancos de dados das lojas. É necessário garantir que as lojas estejam alinhadas com a estratégia de integração de dados e que os funcionários estejam cientes dos impactos da replicação de dados em seus sistemas. Isso é fundamental para evitar interrupções no atendimento ao cliente e garantir a continuidade das operações comerciais.

Por outro lado, no segundo cenário, embora ainda sejam necessários esforços para alinhar a equipe responsável pela configuração e manutenção da ferramenta com a estratégia de integração de dados, os desafios organizacionais não têm um impacto direto no atendimento ao cliente. O foco está mais na capacitação técnica da equipe para garantir o funcionamento eficaz da ferramenta de integração de dados.

Em suma, este estudo demonstra que a utilização de KPIs baseados nos dados integrados no banco de dados centralizado apresenta uma solução robusta para as necessidades de análise em uma empresa de varejo. A centralização dos dados e a análise detalhada proporcionam vantagens significativas. Além disso, a segurança e a privacidade dos dados são aprimoradas. Os desafios técnicos e organizacionais, embora presentes em ambos os cenários de teste, destacam-se no primeiro cenário devido ao impacto direto no atendimento ao cliente e no funcionamento das lojas. É importante alinhar as estratégias de integração e conscientizar os funcionários sobre esses impactos. Portanto, a escolha entre esses cenários dependerá das prioridades da empresa em relação ao equilíbrio entre eficiência, estabilidade e impacto organizacional.

7 REFERÊNCIAS

[1] PETERSON, Eric T. **The big book of key performance indicators**. Web analytics demystified, 2006.

[2] SCHRÖER, Christoph; FRISCHKORN, Jonas. Decentralized and Microservice-Oriented Data Integration for External Data Sources. In: **Innovation Through Information Systems: Volume III: A Collection of Latest Research on Management Issues**. Springer International Publishing, 2021. p. 55-60.

[3] ATMAJA, I. Putu Medagia et al. Implementation of change data capture in ETL process for data warehouse using HDFS and apache spark. In: **2017 International Workshop on Big Data and Information Security (IWBIS)**. IEEE, 2017. p. 49-55.

[4] KAKISH, Kamal; KRAFT, Theresa A. ETL evolution for real-time data warehousing. In: **Proceedings of the Conference on Information Systems Applied Research ISSN**. 2012. p. 1508.

[5] STEIN, Brian; MORRISON, Alan. The enterprise data lake: Better integration and deeper analytics. **PwC Technology Forecast: Rethinking integration**, v. 1, n. 1-9, p. 18, 2014.

[6] COUTO, Júlia Colleoni; RUIZ, Duncan Dubugras. An overview about data integration in data lakes. In: **2022 17th Iberian Conference on Information Systems and Technologies (CISTI)**. IEEE, 2022. p. 1-7.

[7] NARGESIAN, Fatemeh et al. Data lake management: challenges and opportunities. **Proceedings of the VLDB Endowment**, v. 12, n. 12, p. 1986-1989, 2019.

[8] RAVAT, Franck; ZHAO, Yan. Metadata management for data lakes. In: **New Trends in Databases and Information Systems: ADBIS 2019 Short Papers, Workshops BBIGAP, QAUCA, SemBDM, SIMPDA, M2P, MADEISD, and Doctoral Consortium, Bled, Slovenia, September 8–11, 2019, Proceedings 23**. Springer International Publishing, 2019. p. 37-44.

[9] KAKISH, Kamal; KRAFT, Theresa A. ETL evolution for real-time data warehousing. In: **Proceedings of the Conference on Information Systems**

Applied Research ISSN. 2012. p. 1508.

[10] DORAN, Mark; POTINENI, Padmaja; BHATIYA, Rajesh. **Oracle Database Administrator's Guide, 21c**. F31835-14. Copyright © 1996, 2022, Oracle and/or its affiliates.

[11] **QLIK REPLICATE SETUP AND USER GUIDE**. Qlik ReplicateTM. Maio de 2022. Copyright © 1993-2023 QlikTech International AB. Todos os direitos reservados.

[12] ABUHAMDA, Enas; ISMAIL, Islam Asim; BSHARAT, Tahani RK. Understanding quantitative and qualitative research methods: A theoretical perspective for young researchers. **International Journal of Research**, v. 8, n. 2, p. 71-87, 2021.

[13] KINITZ, David J. The emotional and psychological labor of insider qualitative research among systemically marginalized groups: Revisiting the uses of reflexivity. **Qualitative Health Research**, v. 32, n. 11, p. 1635-1647, 2022.

[14] DURBIN, Jason. **Oracle 8 Distributed Database Systems**. Copyright © 1996, 1997, Oracle and/or its affiliates.

[15] Science Research Society. (2023). **Distributed Database Systems for Large-Scale Data Management**. Turkish Online Journal of Qualitative Inquiry, 11(4).

[16] **The Oracle Optimizer Explain the Explain Plan**. Copyright © 2017, Oracle and/or its affiliates.