JOÃO LUGARINHO MENEZES

INCORPORAÇÃO DE QUALIDADE DE DADOS NO PROCESSO DE ETL EM AMBIENTE DE ALTO VOLUME DE DADOS: UM EXPERIMENTO

Monografia apresentada ao Programa de Educação Continuada da Escola Politécnica da Universidade de São Paulo, para obtenção do título de Especialista, pelo Programa de Pós-Graduação em Engenharia de Dados e Big Data.

SÃO PAULO

2024

JOÃO LUGARINHO MENEZES

INCORPORAÇÃO DE QUALIDADE DE DADOS NO PROCESSO DE ETL EM AMBIENTE DE ALTO VOLUME DE DADOS: UM EXPERIMENTO

Monografia apresentada ao Programa de Educação Continuada da Escola Politécnica da Universidade de São Paulo, para obtenção do título de Especialista, pelo Programa de Pós-Graduação em Engenharia de Dados e Big Data.

Área de concentração: Tecnologia da Informação – Engenharia/ Tecnologia/ Gestão

Orientador: Leandro Mendes Ferreira

SÃO PAULO

2024

FICHA CATALOGRÁFICA

Menezes, João
INCORPORAÇÃO DE QUALIDADE DE DADOS NO PROCESSO DE ETL EM AMBIENTE
DE ALTO VOLUME DE DADOS: UM EXPERIMENTO / J. Menezes – São Paulo, 2023.
37 p.

Monografia (Especialização em Big Data & Engenharia de Dados) – Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia.

1.Big Data 2.Data Lakehouse 3.Qualidade de Dados 4.ETL I.Universidade de São Paulo. Escola Politécnica. PECE – Programa de Educação Continuada em Engenharia II.t.

AGRADECIMENTOS

Agradeço a meus familiares, meu orientador Prof. Leandro Mendes Ferreira, demais professores e amigos por tornarem este trabalho possível.

CURSO ENGENHARIA DE BIG DATA

Coordenadora: Profª. Drª Solange Nice Alves de Souza
Vice-Coordenadora: Profª. Drª Anarosa Alves Franco Brandão

Perspectivas profissionais alcançadas com o curso:

[*relate aqui se como resultado do curso, obteve alguma nova recolocação seja na própria empresa ou em outra. Para a continuidade e melhoria do curso é importante saber se cursar a pós-graduação trouxe benefícios diretos na carreira profissional.*]

RESUMO

Este estudo aborda a importância da qualidade dos dados no contexto de *Big Data*, especialmente em processos de Extração, Transformação e Carga (ETL), essenciais para a geração de vantagem competitiva e eficácia na tomada de decisão. Levando em consideração as características necessárias pelo *Big Data*, é conduzido um experimento que integra a qualidade dos dados aos processos de ETL em um ambiente de alto volume de dados, baseado em metodologias de trabalhos anteriores. O estudo destaca vantagens, como a prevenção proativa de problemas de qualidade e a capacidade de ajustes rápidos na implementação do ETL. Embora útil no *Data Lakehouse*, são necessárias adaptações para otimizar essa metodologia às características específicas do *Big Data*.

Palavras-chave: Big Data, Data Lakehouse, Qualidade de Dados, ETL.

ABSTRACT

This study addresses the importance of data quality in *Big Data*, particularly in Extract, Transform, and Load (ETL) processes, which are essential for generating competitive advantage and effectiveness in decision-making. Considering the characteristics needed by *Big Data*, an experiment is conducted integrating data quality into the ETL processes in a high-volume data environment, based on methodologies from previous works. The study highlights advantages such as proactive prevention of quality issues and the ability to make quick adjustments in ETL implementation. Although applicable in the *Data Lakehouse*, adaptations are necessary to optimize this methodology to the specific characteristics of *Big Data.*

Keywords: Big Data, Data Lakehouse, Data Quality, ETL

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nas últimas décadas, a geração e coleta de dados trouxeram a importância de se ter dados com qualidade (Cai e Zhu, 2015). Pois, os dados são utilizados para tomar decisões. Á medida que novas empresas se digitalizam e outras já surgem no ambiente digital, a forma com que os dados são tratados para gerarem informação passa a se tornar uma vantagem competitiva. Portanto é necessário que esses dados sejam confiáveis para a garantia da acurácia na tomada de decisão. Informações imprecisas, inconsistentes ou incompletas podem levar a resultados inadequados e impactar negativamente os objetivos organizacionais.

A qualidade de dados é um conceito essencial para qualquer atividade que busca tomar decisões e a relevância desse conceito tem se ampliado com o surgimento do *Big Data* (Cai e Zhu, 2015). A disciplina da qualidade de dados, considera não apenas suas diversas dimensões, mas também o contexto em que os dados são gerados, armazenados e utilizados. Portanto, o dado é considerado de alta qualidade quando atende as expectativas e necessidades do seu consumidor (DAMA-DMBOK, 2017).

Comumente o *Big Data* é definido e explicado através de 3Vs, os quais passam a ser desafios para as organizações: a) volume, b) variedade; c) velocidade. Alguns autores trabalham com definições mais amplas e incluem outros 2Vs, por exemplo, veracidade e valor como Abdullah et al. (2015) que se referem, respectivamente, a confiabilidade do dado e ao potencial de geração de valor para o negócio. Dessa forma, diversas tecnologias foram desenvolvidas para lidar de forma eficiente com o: a) volume: imenso volume de dados, b) velocidade: de forma a processá-lo eficientemente e dentro de estreitos requisitos de tempo; e c) variedade: compreendem dados estruturados, semiestruturados e não estruturados. Becker et al. (2015), cita que existe um aumento do problema de qualidade de dados proporcional ao aumento do volume de dados. Por isso, a veracidade e valor se tornam importantes no âmbito de *Big Data*.

Para lidar e armazenar a imensa quantidade de dados geradas e capturados por organizações, surgiu o *Data Lake* que consiste em um repositório de dados para

processamento e análise dos dados em seu estado inicial (dados brutos) (Ramchand e Mahmood, 2022). A aplicação do *Data Lake* e outras tecnologias de *Big Data* possibilitaram que empresas armazenassem grande volume de dados e fossem capazes de processá-los para possibilitar diferentes entendimentos de negócio e apoiar na tomada de decisão. Porém, é necessária a garantia de qualidade dos dados em todos os seus estágios de transformação para gerar informação de valor ao negócio.

Em sua pesquisa, Munawar (2021) apresenta uma metodologia de tratamento de qualidade de dados em *Data Warehouses* (DW) que propõe incorporar os processos de verificação de qualidade de dados durante o fluxo de ETL. A incorporação da qualidade de dados durante o processo de ETL é importante pois, segundo o autor, a maior parte da concentração do esforço para desenvolver um *Data Warehouse* está no ETL. A proposta do autor é focada exclusivamente em *DWs* e não é aplicada em nenhum contexto de *Big Data*. Uma das maneiras de atingir esse objetivo seria através da implementação desta metodologia a arquitetura de *Data Lake* conhecida como *Data Lakehouse*. Essa arquitetura tem o objetivo de solucionar os problemas presentes tanto nos *DWs* quanto em outras arquiteturas de *Data Lake* (Oreščanin e Hlupić, 2021). A proposta principal da arquitetura de Data Lakehouse é oferecer uma plataforma unificada o processamento e consulta de dados mantendo as características de desempenho e governança de um *Data Warehouse* com a flexibilidade e custo-benefício de *Data Lakes* (Errami et al., 2023).

Sendo assim, este trabalho tem como objetivo propor uma forma de garantir a qualidade de dados nas arquiteturas de *Data Lakehouse* a partir da verificação da qualidade de dados durante o processo de ETL focando-se na transformação do dado baseando-se no trabalho de Munawar. Este trabalho está divido nas seguintes seções: a) Fundamentação teórica: que irá descrever as características do *Data Lakehouse*, a qualidade de dados, os problemas mais comuns de qualidade de dados em *Big Data* e o trabalho proposto por Munawar; b) Experimento: seção a qual irá apresentar a metodologia e como o experimento foi realizado; c) Discussão: apresenta os resultados, discute a necessidade das metodologia no cenário de *Big Data* e sugere propostas de melhorias; e, por fim, d) Conclusão: um resumo do que foi apresentado, apresentação dos macros resultados, contribuições e sugestões de trabalhos futuros.

1.1 Problema de Pesquisa

O desenvolvimento desta pesquisa tem como foco principal a qualidade de dados em ambiente de *Big Data* e o processo de ETL. O problema pode ser expresso na seguinte pergunta: **como aplicar o processo de qualidade de dados no ETL em um ambiente de alto volume de dados?**

1.2 Objetivo

O objetivo deste trabalho é verificar a importância da inclusão do processo de qualidade durante o processo de ETL em ambiente de *Big Data* como o *Data Lakehouse*. A proposta dessa verificação deve ser busca responder as seguintes perguntas:

- A inclusão do processo de qualidade no ETL soluciona quais das principais causas de problemas de qualidade de dados?
- A inclusão do processo de qualidade no ETL agrava problemas conhecidos relacionados aos Vs do Big Data?
- A inclusão do processo de qualidade no ETL apresenta benefícios em um ambiente de alto de volume de dados?

1.3 Metodologia

Para responder as perguntas propostas no objetivo do trabalho foi realizado uma pesquisa-ação que segundo Engel (2000) busca desenvolver o conhecimento e a compreensão como parte da prática e visa intervir inovadoramente na prática durante o processo investigativo e não apenas ao final do projeto. A pesquisa-ação é composta por três fases sendo eles a coleta de informação, análise e interpretação de fatos e implementação para validar as ações (Krafta et. al, 2007) Portanto, para esta pesquisa foi realizada uma pesquisa de referências afim de mapear arquiteturas de Big Data e problemas de qualidade de dados. Além disso, foi realizada a construção do cenário selecionando a base de dados, tecnologias e desenvolvendo o script do experimento. Por fim, para validar as ações, foi executado o experimento em ambiente

controlado, os resultados foram analisados e quando necessário melhorias foram implementadas.

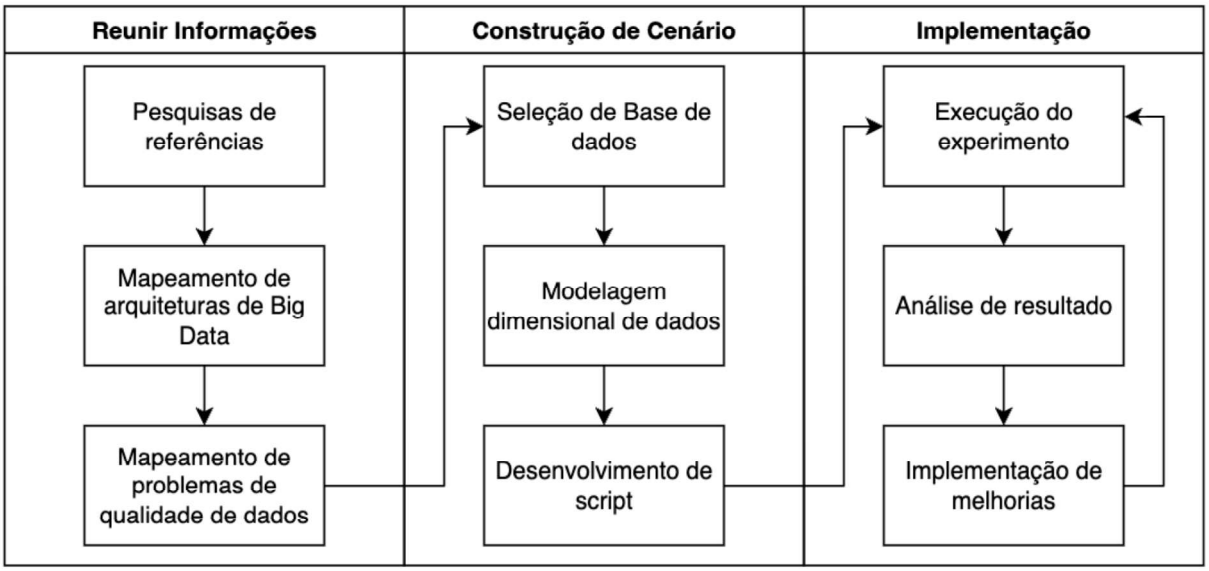

Figura 1 - Representação do processo de execução da pesquisa

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Data Lakehouse

Data Lake é definido como um sistema de armazenamento de baixo custo com arquivos com capacidade para armazenar dados em formatos genéricos (Armbrust et al., 2021). Conceito criado por Dixon em 2010[1], o *Data Lake* surgiu para solucionar problemas relacionados a nova realidade de dados na qual as empresas estavam vivenciando. Segundo o autor, os principais problemas mapeados estavam relacionados com o fato de que de 80% a 90% das empresas lidavam com dados semiestruturados e não-estruturados, em um cenário no qual os dados apresentavam um volume diário que não eram suportados tecnicamente e/ou economicamente em um Sistema Gerenciador de Banco de Dados Relacionais (SGBDRS).

Apesar de ser uma nova abordagem para lidar com os novos formatos de dados e, inicialmente, para substituir o *Data Warehouse*, os *Data Lakes* ainda assim apresentaram novos desafios para as arquiteturas de dados. No trabalho de Armbrust et al. (2021) é pontuado que os *Data Lakes* eram, essencialmente, arquiteturas *schema-on-read* o que possibilitou a velocidade de armazenamento em baixo custo. Porém, isso trouxe novos problemas associados a governança e qualidade dos dados. Harby e Zulkernine (2022) refletem acerca de outros problemas como a descoberta de informações relevantes nos dados armazenados justamente pelo alto volume, falta de estrutura de dados compreensível e falta de catálogo dos dados o que transforma os Data Lake em *Data Swamps* - “Pântanos de Dados” em tradução livre.

Conforme essa abordagem foi-se aperfeiçoando, novas formas de organizar os *Data Lakes* foram propostas para que fosse possível mitigar ou reduzir significativamente os novos problemas associados a essa arquitetura. Hlupić et al. (2022) faz uma revisão bibliográfica acerca dos modelos propostos de organização de Data Lake, sendo eles:

---

[1] https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

a) *"Two Layered Architecture"*: Segue a definição de um Data Lake como repositório de dados brutos que são armazenados permanentemente, acrescentando duas camadas de dados. A primeira camada recebe os dados em seu formato original e são mantidos temporariamente e a segunda camada mantém os dados de forma permanente.
b) "*Data Lakehouse Architecture*": Combina as funcionalidades de Data Lake e Data Warehouse. Possui uma camada de dados em seu formato original, uma camada que normaliza o formato do dado bruto em um formato aberto de tabela e demais camadas de dados processados
c) "*Data Pond Architecture*": Divide os dados nas camadas do Data Lake a partir do propósito que será utilizado como, por exemplo, dados textuais, tabulares, imagens etc. Cada camada é logicamente separada das outras e tem um foco específico na estrutura e utilização dos dados.
d) "*Multi-Layered Architecture*": Divide as camadas do Data Lake a partir das funções empregadas no ciclo de processamento do dado como, por exemplo, ingestão, armazenamento, transformação e consumo.
e) *"Zaloni Zone Architecture"*: Organiza os dados em diversas camadas (zonas) para diferentes etapas de processamento como, por exemplo, zona de dados brutos, zonas de dados confiáveis com qualidade e conformidade verificadas, zonas refinadas para dados modelados conforme as necessidades de usuários finais.
f) "*Data Vault Based Zone Architecture*": Usa o método *Data Vault* para estruturar dados em zonas distintas. Cada zona é modelada para uso específico, incluindo zonas para armazenamento de dados brutos, zonas para dados estruturados, além de zonas de entrega e exploração para análises *de dados.*

Armbrust et al. (2021) propuseram o conceito de *Data Lakehouse*, o qual representa um sistema de gerenciamento de dados que fornece características de gerenciamento e desempenho de um SGBDR como princípio ACID, as quais garantem confiabilidade, isolamento e consistência mantendo a integridade do dado, versionamento, indexação, armazenamento intermediário (*caching*) e otimização de consultas sobre *Data Lakes*. Desta forma, a proposta do *Data Lakehouse* pretende combinar a eficiência e gerenciamento do *Data Warehouse* com o baixo custo,

flexibilidade e suporte a uma grande variedade de formato de dados do *Data Lake* (Errami et al. 2023).

O *Data Lakehouse* surgiu a partir dos desafios apresentados na era dos *Data Warehouses, Data Lakes* e suas primeiras concepções. Ainda, no trabalho de Armbrust et al. (2021), os autores justificam o surgimento do *Data Lakehouse* a partir do que chamam de primeira e segunda geração *Data Analytics Platform*. A primeira, a qual refere-se a era do *Data Warehouse*, teve como principal desafio o alto volume de dados e os diferentes formatos e estruturas de dados que os sistemas de *DWs* não conseguiam consultar. Já a segunda geração, que tem relação com os *Data Lakes* que foram inseridos nas arquiteturas de plataforma de dados para solucionar os problemas dos *DWs*, teve como principais desafios a consistência entre o *Data Lake* e *Data Warehouse*, os novos pontos de falha criados a partir dos processos que atualizariam os *DWs* a partir do *Data Lake*, obsolescência do dado entre os dois sistemas entre outros.

Errami et al. (2023) citam que as principais características do *Data Lakehouse*, que visam endereçar as limitações encontradas no *Data Warehouse* e *Data Lake*, são:

a) **Gerenciamento:** A implementação do princípio ACID e ambiente unificado, característica herdada do *DW*, garantem confiança no ecossistema dos dados pois há controle de transações do que é inserido, atualizado, deletado ou lido.
b) **Otimização:** Por conta da requisição de um formato específico de arquivo para operar no *Lakehouse* conhecido como *Open Table Format* (OTF), o sistema pode-se valer de técnicas que leem os metadados do arquivo para realizar, por exemplo, uma leitura de dados otimizada para garantir o desempenho de consultas.
c) **Armazenamento:** O sistema de armazenamento e processamento são completamente independentes. Isso garante a possibilidade de diversos sistemas se conectarem e usarem os dados do *Data Lakehouse*.
d) **Metadados:** O *Data Lakehouse* implementa uma camada de armazenamento transacional de metadados para lidar com problemas de concorrência para leituras e gravações de dados, também permitirá o controle de versão das tabelas entre outras características propostas pelos sistemas que operam no *Lakehouse*.

2.2 Qualidade de Dados

A qualidade de dados é definida e caracterizada na literatura, de forma geral, como um dado ou conjunto de dados que está adequado para ser utilizado pelo seu consumidor (Strong, Wang, 1996). Taleb et al. (2016) argumentam que importância desse conceito está relacionada com o fato de que decisões estratégicas são baseadas a partir de *insights* produzidos a partir dos dados. Além disso, também pontuam que muitos problemas de discrepâncias e inconsistências nos dados são gerados por diversos fatores, incluindo a intervenção humana no ciclo de vida do dado. Por exemplo, Strong et al. (1997) apresentam uma análise de problemas de qualidade em três diferentes empresas e indicam que, em uma dessas empresas que é do ramo da saúde, médicos que utilizavam formulários com opções definidas (*checkboxes*) geravam dados mais confiáveis do que os que usavam formulários com campos abertos. Ainda assim, a combinação desses dados em sua fonte de origem gerava um conjunto de dados de baixa qualidade.

Firmani et al. (2016) observaram que a qualidade de dados é um conceito multifacetado e que são necessárias diversas dimensões para explicá-la. As dimensões de qualidade são critérios pelos quais a qualidade pode ser avaliada através de propriedades e suas respectivas métricas (Carlo et al., 2011). As definições de dimensões aplicadas a qualidade de dados não possuem uma unificação podendo ser vistas dimensões e métricas diferentes na bibliografia científica. Para este trabalho foram adotadas dimensões definidas no DAMA-DMBOK (2017), sendo as seguintes:

- Completude: A proporção dos dados armazenados em relação ao potencial de 100%
- Unicidade: Nenhuma entidade é armazenada mais de uma vez baseado em como a instância é identificada
- Pontualidade: O grau com que o dado representa a realidade do momento requerido
- Validade: Os dados são válidos se estiverem em conformidade com a sintaxe (formato, tipo e intervalo) de sua definição
- Precisão: O grau com que o dado descreve corretamente o evento ou objeto da realidade sendo descrito.

- Consistência: Ausência de diferenças quando comparado duas ou mais representações de um evento ou objeto a sua definição.

Cada dimensão de qualidade de dados é associada a métricas especificas que são métodos ou fórmulas estabelecidas para quantificar e classificar as dimensões. Toda métrica oferece uma maneira de avaliar a dimensão seja a partir de fórmulas simples ou expressões multivariadas mais complexas as quais possibilitam uma avaliação precisa e detalhada da qualidade dos dados (Taleb et al. 2018). As métricas são importantes para detectar problemas ao passo que são inseridas nos sistemas analíticos (Abdullah et al., 2015). Portanto, qualquer métrica deve ser capaz de distinguir se um dado respeita ou não determinado atributo definido de qualidade (Taleb et al., 2016).

2.2.1 Origens de problemas em Qualidade de dados

Após a compreensão do que constitui a qualidade de dados, suas propriedades e métricas que a definem, é importante explorar as causas subjacentes que frequentemente levam a problemas na qualidade de dados. Dessa forma, o problema de qualidade de dados pode ser abordado de uma forma tradicional, que considera os dados como conceitos intrínsecos e independentes, entretanto uma abordagem moderna sobre a ótica de qualidade de dados visa compreender que a qualidade de dados está presente no contexto em que os dados são produzidos e utilizados. Dessa forma, a alta qualidade do dado é relacionada com a utilidade para o consumidor. Neste sentido, Strong et al. (1997) propuseram que a qualidade seja avaliada pelas categorias: intrínseca, acessibilidade, contextual e representacional. Cada uma destas categorias compõe um grupo de dimensões elencadas pelos autores. Sendo assim, os problemas relatados que reduzem a usabilidade do dado são, por exemplo, incompatibilidade entre fontes do mesmo dado, dados incompletos ou mal definidos por resultado de decisões técnicas, problemas de integração e análise que afetam a capacidade de uso dos dados, recursos de sistemas insuficientes e questões de segurança no acesso aos dados.

Ainda aplicando a abordagem de qualidade de dados presente no contexto em que os dados são produzidos e utilizados, vemos em sistemas analíticos que o

processo de criação, integração e transformação de dados para esses sistemas podem causar problemas de qualidade de dados. Em Chen et al. (2009) vemos que um dos principais problemas mapeados em qualidade de dados têm relação com a fonte de dados e o processo de ETL desse *DW*. A fonte de dados, segundo os autores, produzia alguns dados de baixa qualidade e que eram inseridos no *DW*. Já no processo de ETL, erros surgiram durante a extração e conversão de dados, além de o carregamento inadequado amplificar essas inconsistências. De forma semelhante, vemos em Idris e Ahmad (2011) uma discussão sobre a gestão da qualidade dos dados no processo de desenvolvimento de *DW*, contrastando a abordagem convencional, que faz o tratamento da qualidade posteriormente.

A proposta dos autores é utilizar uma metodologia baseada na ISO 9001:2008 e Gestão Total da Qualidade de Dados (GTQD). Esta abordagem visa identificar e resolver problemas de qualidade nas fases iniciais, minimizando assim o impacto negativo nos sistemas de apoio à decisão. Entre as principais causas dos problemas identificados estão: seleção específica de fontes de dados, falta de rotinas de validação, alterações inesperadas nos sistemas de origem, representações de dados inconsistentes, registros duplicados e atrasos na atualização das fontes.

2.2.2 Qualidade de dados em Big Data

O contexto de *Big Data* é definido através das suas principais características, ou seja, pela Velocidade, Volume, Variedade, Valor e Veracidade de dados (5Vs) (Taleb et al., 2018), desta forma a qualidade de dados em *Big Data* deve ser avaliada também por estas características. A nova realidade dos dados em *Big Data* traz formatos diversos, maior volume e maior velocidade de entrega e alteração. Dentre as principais causas de problemas de qualidade de dados, a diversidade de fontes de dados, estruturas complexas e dificuldade de integração. O volume de dados também é uma questão que desafia a avaliação da qualidade em um prazo razoável e a rapidez na mudança dos dados a qual impõe requisitos mais altos para as tecnologias de processamento (Cai e Zhu, 2015).

Além das questões relacionadas aos 3Vs do *Big Data*, Becker et al. (2015) apresentaram outras perspectivas relacionadas a qualidade em um ambiente de alta

variedade e volume de dados. Em seu trabalho, o qual faz um estudo com quatro diferentes entidades que lidam com *Big Data*, observaram problemas relacionados aos *pipelines* de dados que podem introduzir erros não intencionais aos dados ou são utilizados, em grande parte, para compensar problemas de qualidade das fontes. Também pontua que no contexto de *Big Data* erros manuais são mais difíceis de encontrar do que os dos erros gerados por dispositivos ou sistemas. Abaixo, na Tabela 1, é apresentado uma sumarização das informações coletadas apresentando as origens das fontes dos problemas de dados e os principais problemas mapeados de Qualidade de Dados (QD):

Tabela 1 - Origens e Problemas de Qualidade de Dados

Origem	Problemas
Fonte de Dados	• Incompatibilidade entre dados semelhantes de fontes diferentes • Dados incompletos • Integração limitada • Seleção inadequada de dados • Alterações inesperadas • Estruturação inadequada dos dados
Manipulação dos Dados	• Extração incompleta • Transformação de dados incorreta • Carregamento de dados incorretos
Infraestrutura	• Recursos insuficientes para processamento • Acesso limitado (segurança)
Gestão	• Falta de rotinas de validação

2.2.3 Incorporação de Qualidade de Dados no ETL

O processo de Extração, Transformação e Carga (ETL) é um procedimento que extrai dados de fontes e aplica regras de negócio que incluem adequações, limpezas e agregações para que estes possam ser carregados em suas tabelas finais, em geral utilizado para criação de sistemas analíticos e DW (Ferreira, 2015). Machado et al. apresentaram que a atividade de ETL e limpeza de dados podem representar até um terço do orçamento total de um projeto de *DW*, podendo consumir até 80% do tempo total de desenvolvimento.

No trabalho proposto por Munawar (2021) é discutida a importância dos processos de ETL em projetos de *DW* com foco na integração de dimensões de qualidade para melhorar os resultados de análises e suporte a decisão, garantindo que o *DW* seja confiável para os usuários de negócio. Para atingir esse objetivo é proposto um *framework* para incorporar todas as dimensões importantes de QD nos processos de ETL, com o objetivo de manter altos níveis de qualidade de dados. A partir do framework proposto por Munawar (2021) apresentamos na Figura 2 a interpretação do diagrama. Onde destacamos em partes: (a) a representação da tabela e suas principais características e (b) o processo de transformação associado a implementação de qualidade de dados.

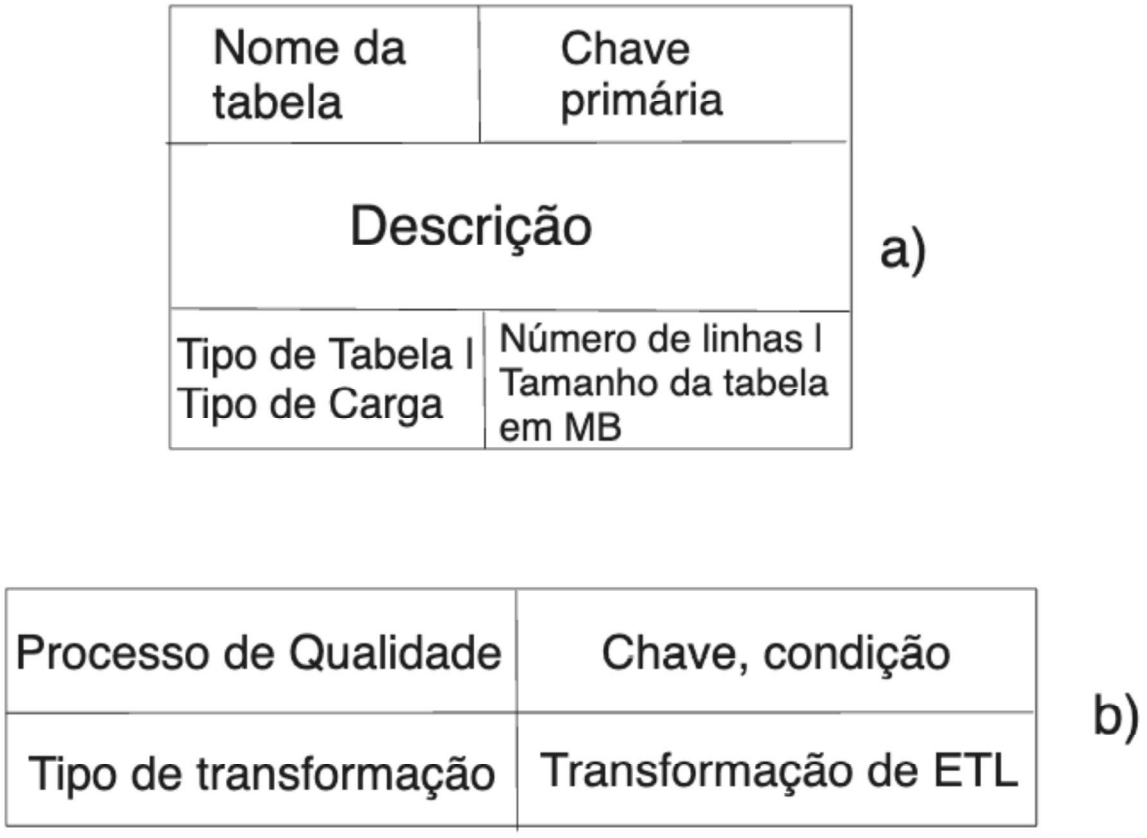

Figura 2 - Representação de Tabela (a) e Processo de transformação de ETL (b) – Fonte: Adaptado de Munawar (2021)

Na Figura 3 é apresentado um exemplo de implementação deste framework. Este exemplo mostra o procedimento de ETL associada a execução de testes de qualidade no qual duas tabelas são utilizadas como fonte para gerar uma terceira tabela.

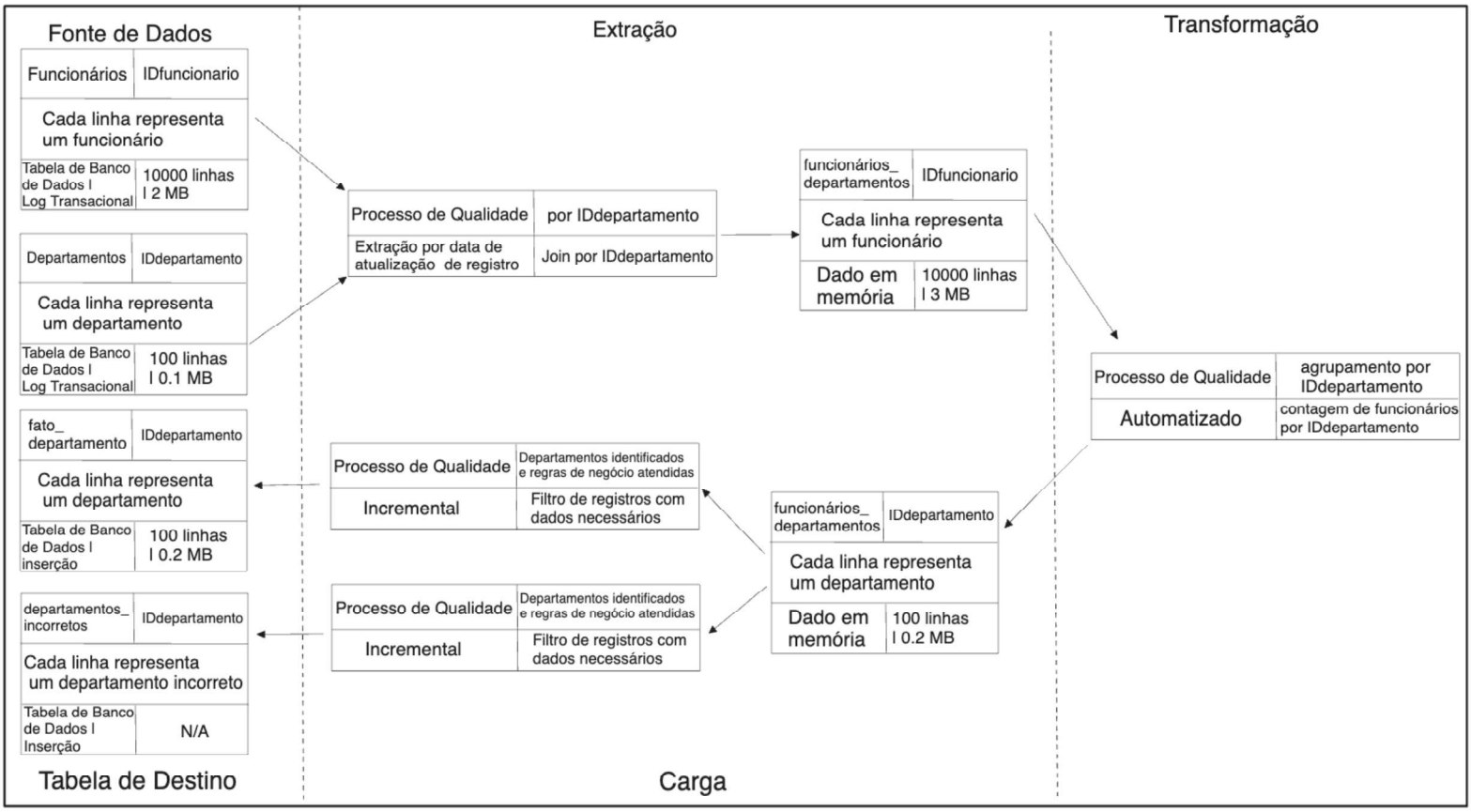

Figura 3 - Visualização de processo de ETL implementado com Framework proposta – Adaptado de Munawar (2021)

Para total compreensão do exemplo, a Figura 4 representa as tabelas que são utilizadas, sendo (a) as fontes de dados e (b) a tabela que será criada a partir do processo de ETL. O fluxograma de ETL apresentado pode ser interpretado como uma operação que se inicia extraindo os dados da tabela Funcionários e Departamentos pela data de atualização e realiza um *JOIN* utilizando o IDdepartamento para então realizar verificações de qualidade como completude. Depois, este dado é agregado pelo IDdepartamento para que a contagem de funcionários por departamento seja feita. Por fim, os dados são carregados na tabela de destino “fato_departamento” de forma incremental caso nenhuma regra de negócio e qualidade tenha sido quebrada durante o processo de ETL, do contrário os dados são inseridos na tabela “departamentos_erros” que guardará dados com problemas.

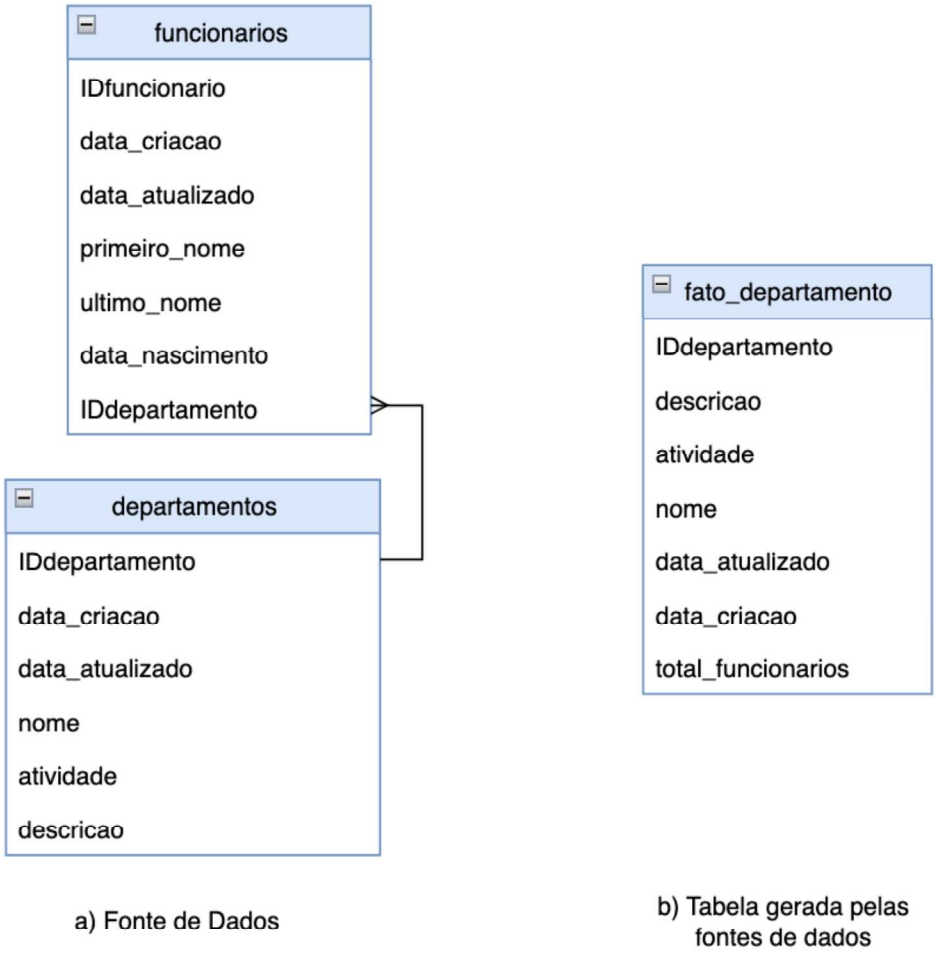

Figura 4 - Representação de tabelas fonte (a) e tabela gerada a partir de fontes (b)

Sendo assim, Munawar (2021) discute a eficácia dessa abordagem proposta e segure que seja mais compreensível para pessoas não técnicas e personalizável em comparação com outras abordagens. Também cita como vantagem a inclusão dos testes de qualidade desde a extração até o carregamento dos dados. Para o autor, o design do ETL juntamente do processo de qualidade deve ser tratado como uma parte integral do desenvolvimento de DW visto sua importância durante o desenvolvimento desse sistema analítico.

3 EXPERIMENTO

O experimento proposto realiza uma validação acerca da incorporação de procedimentos de qualidade de dados no ETL em ambiente de alto volume de dados, ou seja, em ambientes de *Big Data*. Portanto, as seções seguintes apresentam a arquitetura proposta para o experimento, fonte de dados, tecnologias utilizadas e, por fim, a implementação do algoritmo que incorpora a QD no ETL.

3.1 Arquitetura Proposta

A arquitetura selecionada para o experimento corresponde a de *Data Lakehouse* representada na Figura 5.

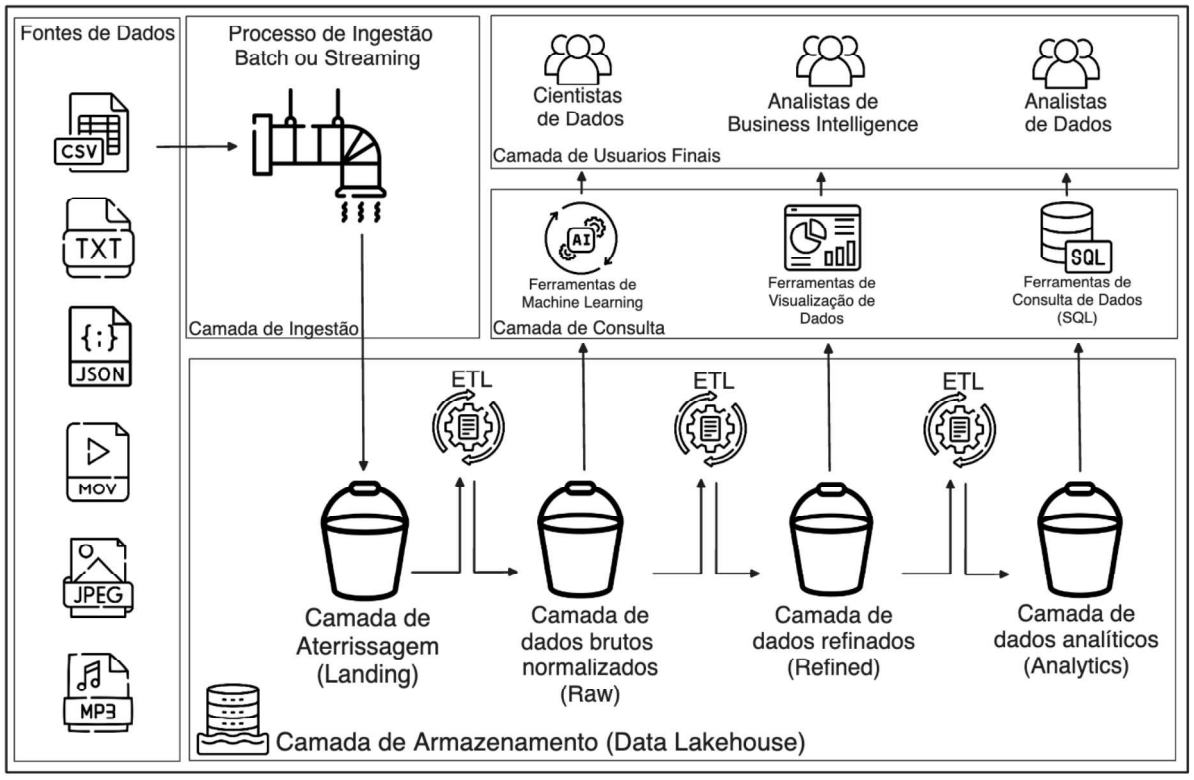

Figura 5 - Arquitetura de Data Lakehouse

As fontes de dados, as quais podem ter diversos formatos, são extraídas, processadas e inseridas no *Data Lakehouse* em seu formato original através de um processo de ingestão o qual pode ser em *batch* e *streaming*, ou seja, por operações em lote ou em quase tempo real. Após esse processo, os dados são inseridos na

camada de Armazenamento que é o próprio *Data Lakehouse*. Essa camada necessita ser operada através de um sistema que implementa as características propostas pelo *Data Lakehouse;* como por exemplo; a possibilidade de realizar operações ACID e a criação e gerenciamento dos metadados das tabelas. A camada de armazenamento é divida fisicamente e logicamente por zonas que delimitam o estágio e função do dado naquela zona. Sendo assim, as camadas são definidas como:

- Camada de Aterrissagem (*Landing*): Os dados provenientes do processo de ingestão são armazenados nesta camada em seu estado e formato original.
- Camada de dados brutos normalizados (*Raw*): Os dados brutos da camada Landing passam por um processo de normalização de formato e passam a adotar o formato requisitado pelo *Data Lakehouse* conhecido como Formato Aberto de Tabela (Open Table Format) que é correspondente ao implementado pelo Apache Parquet, por exemplo. É válido ressaltar que nesse estágio dispensa quaisquer outras alterações que não seja o formato do dado, a integridade e conteúdo original dos dados deve ser mantida.
- Camada de dados refinados (*Refined*): Esta camada é responsável por armazenar os dados brutos normalizados que sofreram alterações regidas por regras de negócio para fins analíticos. No caso deste trabalho, irá armazenar as tabelas que representam dimensões e fatos dos dados selecionados para o experimento.
- Camada de dados analíticos (*Analytics*): Esta camada destina-se a armazenar tabelas que são criadas exclusivamente para fins analíticos como visualizações de dados, relatórios e cálculo de indicadores. Os dados são provenientes da camada de dados refinados e, normalmente, passaram por processos de agregação e desnormalização.

Por fim, a camada de consulta representa sistemas e serviços que acessam os dados do *Data Lakehouse* a partir da camada de dados brutos normalizados. Esses serviços são, por exemplo, ferramentas de relatório e de consulta de dados através de SQL. A camada de usuários mostra exemplos de agentes que podem acessar o *Data Lakehouse* através da camada de consulta. Essas camadas não foram exploradas neste trabalho.

3.2 Fonte de Dados

Como fonte de dados para o projeto foi utilizada uma proveniente do Instituto Nacional de Estudos e Pesquisas Educacionais Anisio Teixeira (INEP)[2], sendo esta referente a informações do Exame Nacional do Ensino Médio (ENEM). Dentre os conjuntos de dados disponíveis, foi utilizado um nomeado como “microdados” o qual possui informações do candidato e as provas realizadas. Foram utilizados dados de 2019 a 2022 totalizando 7.5 GB de dados. O conjunto de dados utilizado possui diversas informações referente a prova realizada assim como algumas informações do candidato. Sendo importante ressaltar que nenhuma das informações utilizadas era possível identificar o candidato. Dessa forma, as informações utilizadas foram divididas em seis diferentes categorias, sendo elas:

- Dados do participante: Compreende número de inscrição, ano que a prova foi realizada, escolaridade, idade e gênero.
- Dados da escola: Informações referente a localização, dependência administrativa e situação de funcionamento
- Dados do Local de Aplicação da Prova: Dados de localização da aplicação da prova como estado e município.
- Dados da Prova Objetiva: Refere-se aos dados das provas como suas notas, código, cor da prova, língua estrangeira selecionada e gabaritos de cada prova.
- Dados da Redação: Contém dados referente a avaliação da prova de redação realizada pelo candidato como notas em cada critério avaliado, nota final e situação da redação do participante.
- Dados do Questionário Socioeconômico: Resposta do questionário submetido aos candidatos referente a características sociais e econômicas.

A modelagem de dados que foi realizada para o experimento é apresentada na Figura 6. Os dados criados segundo o modelo apresentado na Figura foram inseridos nas camadas do *Data Lakehouse* e validados com a incorporação da QD durante o ETL.

---

[2] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

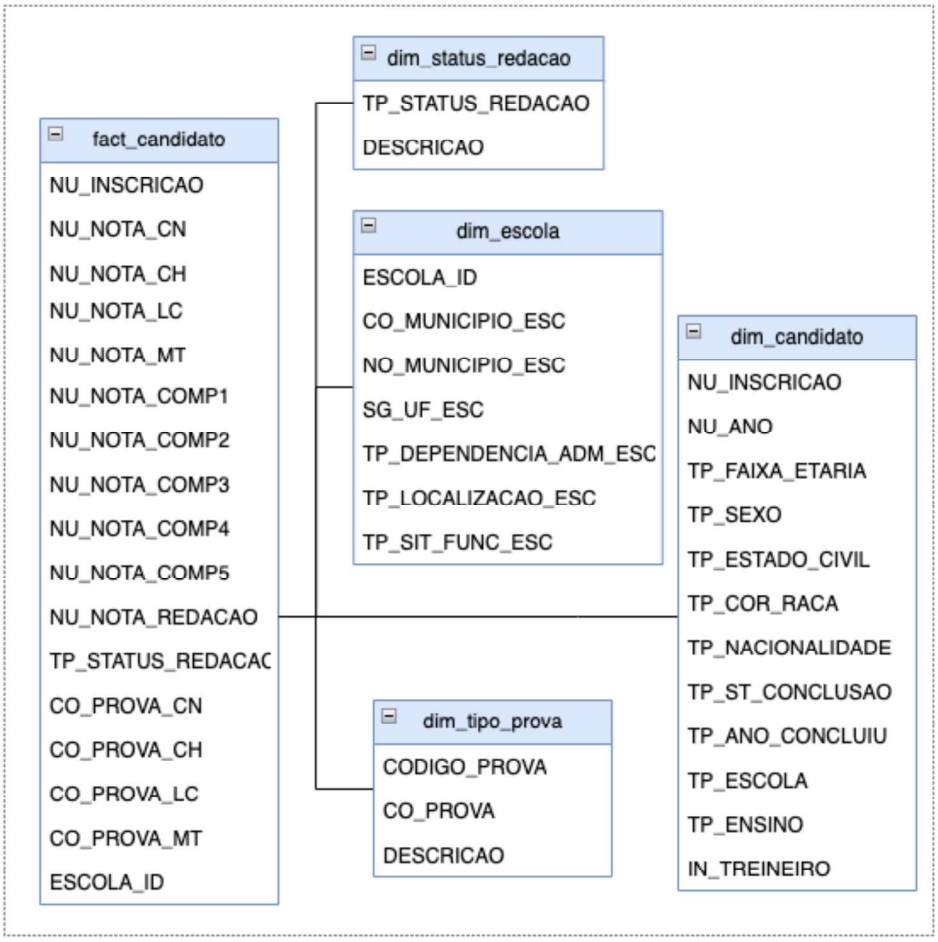

Figura 6 - Modelagem Dimensional dos Microdados do ENEM

A partir da modelagem foram construídas tabelas, como pode ser visto a seguir:

- dim_escola: Dimensão referente a escola em que o candidato cursou o ensino médio com as informações referente as características e localização.
- dim_candidato: Dimensão referente ao candidato e suas características sociais.
- dim_tipo_prova: Dimensão que se refere ao tipo da prova aplicada no ENEM, cor da prova e descrição
- dim_status_redacao: Dimensão referente a situação da redação feita pelos candidatos
- fact_candidato: Tabela fato com demais informações referente a resultados individuais, gerais e da redação realizada pelos candidatos. Importante ressaltar, que a partir dessas informações não era possível identificar o candidato.

3.3 Tecnologias Utilizadas

Nesta seção serão apresentadas as tecnologias utilizadas e as ferramentas escolhidas para o gerenciamento do *Data Lakehouse*. O esquema de implementação com as tecnologias para a realização e execução do experimento pode ser observado na Figura 7.

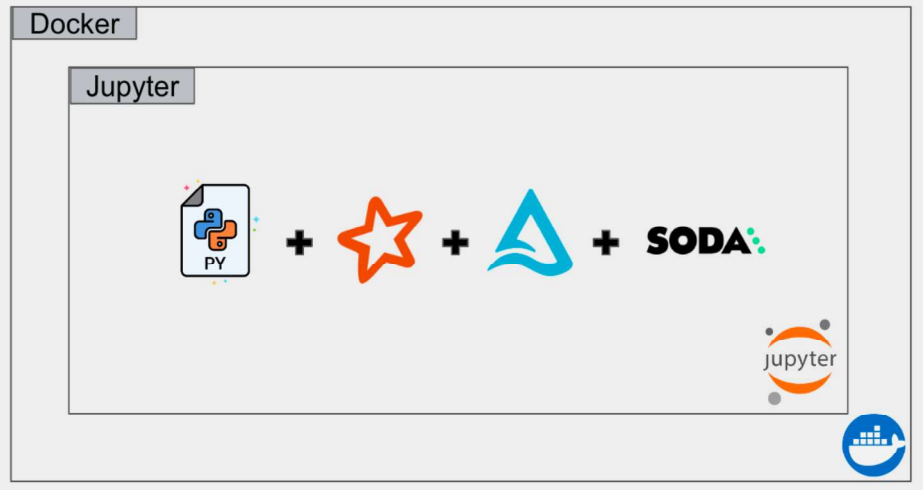

Figura 7 - Representação da Implementação das Tecnologias Utilizadas

Como pode ser visto na Figura 7, o experimento se deu a partir das seguintes tecnologias:

- Docker: Plataforma de código aberto para desenvolver, gerenciar e executar aplicações em containers. O Docker foi configurado para utilizar 12 CPU, 12 GB de RAM e 1 GB de SWAP. A versão utilizada corresponde a 24.0.6
- Jupyter: Aplicação web, baseada em Python, para executar códigos de forma interativa. Esta aplicação em questão será executada e gerenciada pelo Docker.
- Python: Linguagem de programação de alto nível, interpretada e de propósito geral. Todos os códigos serão desenvolvidos e executados por esta linguagem. A versão utilizada corresponde a 3.10.10.
- PySpark: Pacote que implementa o Apache Spark, mecanismo de processamento de dados em larga escala, no Python. Será utilizado para fazer a manipulação dos dados. A versão desse pacote corresponde a 3.3.2

- Delta Spark: Ferramenta que implementa e gerencia as características de Data Lakehouse no Data Lake. Será utilizado para gerenciar os arquivos criados no Data Lake. A versão desse pacote corresponde a 2.3.0
- Soda: Será utilizado para realizar os testes de qualidade nos dados do projeto. A versão desse pacote corresponde a 3.1.0

3.4 Implementação

O processo de incorporação de QD durante o ETL foi implementado segundo o diagrama disposto na Figura 8, que implementa o framework de Munawar (2021) de acordo com do exemplo da Figura 2. Essa implementação foi desenvolvida para que seja genérica e reaproveitável em qualquer estágio da camada do Data Lakehouse. O algoritmo desta execução funciona da seguinte forma:

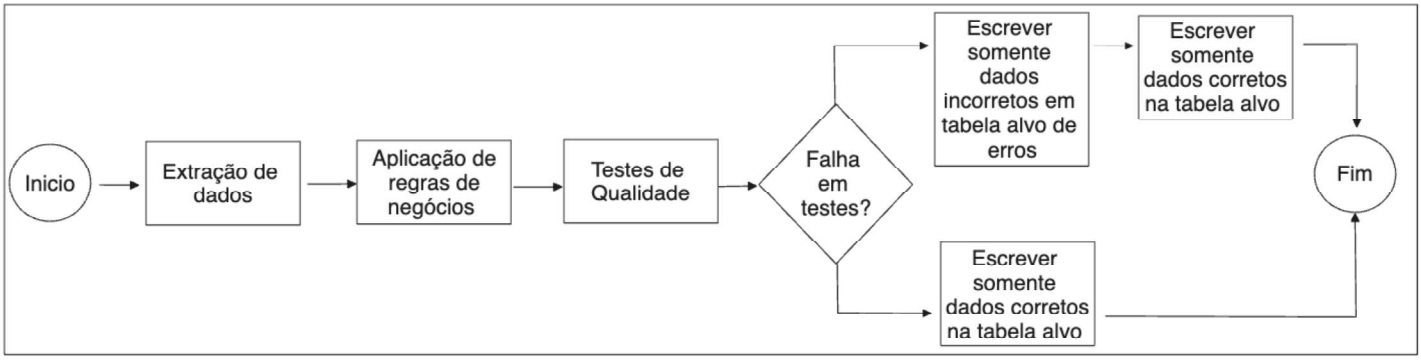

Figura 8 - Processo de Incorporação de QD no ETL

- Extração de dados: Os dados são lidos de uma camada alvo do Data Lake utilizando o gerenciador do Data Lakehouse. No caso deste experimento, o Delta Lake.
- Aplicação de regras de negócios: Este estágio realiza a transformação dos dados. Portanto, são aplicadas normalização ou desnormalizações, agregações, implementação de regras de negócios especificas, junção de tabelas. Este estágio utilizava exclusivamente o Spark para as transformações.
- Testes de Qualidades: Ainda em memória, os dados que foram transformados na fase anterior têm sua qualidade testada a partir das especificações informadas pelo usuário. Nesta fase, é utilizado a ferramenta SODA para realizar os procedimentos de qualidade.

- Verificação de resultado dos testes de qualidade: Por fim, depois que os dados são testados na fase anterior é realizada uma verificação do resultado dos testes para saber se foram encontrados erros no conjunto de dados.
    - Existem inconsistências: Caso existam inconsistências, os dados incorretos são separados dos dados corretos. Após a separação, os dados incorretos são destinados a uma tabela alvo de erros específica para o conjunto de dados manipulado. Já os dados corretos são escritos na tabela alvo específica para o conjunto de dados manipulado.
    - Não existem inconsistências: Os dados são destinados tabela alvo específica para o conjunto de dados manipulado.

O código para este trabalho foi desenvolvido em Python e está disponível em repositório de código aberto (Github), sobre licença open-source e pode ser acessado através do link: https://github.com/joaoluga/data_lakehouse_quality.

A execução do experimento compreendeu um teste de performance entre a execução do processo de ETL com a incorporação da QD e a execução tradicional do ETL sem a realização de verificações de qualidade. Este teste visa verificar questões relacionadas ao desempenho da metodologia proposta por Munawar (2021) no ambiente de alto volume de dados e se há agravamento de problemas de conhecidos no contexto de *Big Data.* Foram verificadas o tempo de criação de camada do *Data Lakehouse* e suas respectivas tabelas a partir da camada de dados brutos normalizados (*raw)*. Além disso, também foi analisado o impacto na qualidade de dados em cada camada do *Data Lakehouse* e suas tabelas após a criação de todas as camadas. Buscou-se compreender os benefícios dessa metodologia e quais origens de problemas conhecidos de qualidade de dados são solucionados. Os testes de qualidade realizados na fonte de dados buscaram considerar que um dado sem qualidade é aquele que não apresenta um alto índice completude. Além disso, foram realizados testes de precisão nos campos normalizados dos microdados do ENEM seguindo as definições presentes no dicionário de dados fornecido pelo INEP. Por fim, foram realizados testes de validade, unicidade e consistência que visavam verificar se o dado entregue está corretamente se referindo ao ano que a prova foi realizada, se os dados não estão duplicados e se os dados não apresentam inconsistência entre si. Essas mesmas verificações foram submetidas aos fluxos de ETL que aplicam regras

de negócios de transformação efetiva nos dados, porém levando em consideração o resultado esperado de cada processo de transformação para regular os critérios do teste de qualidade.

4 DISCUSSÃO DOS RESULTADOS

4.1 Apresentação dos Resultados

Os resultados do desempenho de execução do fluxo de ETL com e sem a incorporação de processos de qualidade podem ser observadas na Tabela 2 na qual é apresentado o tempo, em segundos, de criação de cada camada com suas respectivas tabelas. Já a Tabela 3 apresenta o resultado do processo de qualidade aplicada separadamente em cada tabela demonstrando demais características de cada tabela bem como os registros e colunas que foram considerados validos ou não.

Processo	DQ incorporado no ETL (segundos)	ETL sem DQ (segundos)
Criação da camada raw	4320	306
Criação da camada refined	720	222
Criação da camada analytics	10	30

Tabela 2 - Desempenho de execução dos processos de criação das Camadas do Data Lakehouse

Tabela	**Camada**	**Colunas**	**Registros**	**Registros válidos**	**Registros Incorretos**	**Colunas incorretas**
enem_microdados	Raw	76	17744217	2703996	15040221	17
dim_candidatos	Refined	12	2703996	2703996	0	0
dim_escola	Refined	7	11202	11202	0	0
dim_status_redacao	Refined	2	8	8	0	0
dim_tipo_prova	Refined	3	192	192	0	0
fact_candidato	Refined	17	2703996	2703996	0	0
fact_candidato_desnormalizada	Analytics	33	2703996	2703996	0	0

Tabela 3 - Resultado de processos de qualidade por tabela

4.2 Discussão dos Resultados

A inclusão do processo de qualidade no ETL mostrou-se efetiva na identificação de alguns problemas que ocasionam queda na qualidade dos dados em relação a fonte de dados e ao processo de manipulação de dados. Nas fontes de dados, por exemplo, foi capaz de antecipar problemas de dados incompletos e alterações inesperadas nas fontes de dados. Essas alterações inesperadas normalmente afetam os processos de manipulação de dados ocasionando falhas nas implementações de ETL. Já os problemas relacionados a manipulação de dados, também foi observada a efetividade em evitar ocorrências de aplicação de regras de dados inconsistentes que potencialmente podem inserir dados sem qualidade nos sistemas analíticos. No caso dos dados do ENEM, cada conjunto de dados referente ao ano que a prova foi realizada possui informações distintas quanto ao código e a cor da prova, por exemplo. Assim, os testes de qualidade foram importantes para o processo de desenvolvimento do ETL para melhor atender as regras de negócio associadas a processos de desnormalização dos dados do ENEM.

Cerca de 80% dos dados provenientes da fonte de dados utilizada foram removidos por não atender os requisitos de completude dos dados. Ou seja, o conjunto de dados fornecido pelo ENEM possui muitos campos nulos. As informações nulas em sua grande maioria são referentes a dados da escola. Essa informação seria relevante em um cenário que fosse desejado compreender o resultado das provas por escola e sua localização em relação a políticas públicas aplicadas em cada estado ou município. Em todo caso, os dados válidos e incorretos foram separados da forma esperada em suas respectivas tabelas. Porém, foi analisado que no conjunto de dados incorretos somente 17 colunas de 76 haviam infringido alguma regra de qualidade. Isso ocasiona um impacto para quaisquer outras análises que poderiam utilizar outras colunas que atendem os requisitos de qualidade impostos.

O processo de ETL tradicional teve mais desempenho e demonstrou-se ser 14 vezes mais rápido que o processo que considera a QD durante a execução do ETL. A queda no desempenho foi ocasionada não pelo processo de execução da qualidade de dados em si, mas pelos passos posteriores que sugerem uma separação entre dados válidos e inválidos gerando dois carregamentos de dados em tabelas distintas.

Isso demonstra que há um agravamento proporcionado pelo contexto de Big Data, aplicado ao cenário de QD, nas características de volume e velocidade que tem relação com a eficácia ao avaliar os dados em um prazo razoável frente a velocidade que os dados são recebidos, demandando requisitos mais altos para tecnologias de processamento.

Ainda assim, a abordagem tradicional tende a ofertar maior dificuldade em explorar e ajustar os problemas de qualidade quando tratados posteriormente ao processo de ETL. Isso porque os dados, no caso do *Data Lakehouse*, são separados por camadas que aplicam alterações de negócios e o mesmo dado pode ser utilizado para produzir diversas tabelas dificultando o processo de descoberta da origem do problema de qualidade de dados.

Os benefícios observados no emprego da QD no processo de ETL tem relação com a separação dos dados inconsistentes dos dados de qualidade. A maior ocorrência de erros de qualidade ocorreu na camada de dados brutos normalizados mantendo somente os dados com qualidade provenientes da fonte e que não sofreram intervenção de qualquer regra de negócio por conta da natureza dessa camada. Isso garantiu que as camadas subsequentes não apresentassem erros de qualidade provenientes da fonte de dados apesar de terem sido detectados problemas, mas que tinham relação com as regras de negócio aplicadas na transformação do dado e foram corrigidos durante o desenvolvimento e testes do ETL. Em um cenário tradicional em que a qualidade de dados é implementada somente após a criação das tabelas que compõe os produtos de dados para fins analíticos, em uma arquitetura como a do *Data Lakehouse*, poderia ser complexa a detecção dos agentes redutores da qualidade visto a característica de múltiplas zonas para armazenar os dados.

5 CONCLUSÃO

Conforme as organizações aderem a tecnologias de suporte a nova realidade imposta pelo *Big Data*, a veracidade e validade dos dados assumem um papel central para obter-se vantagem competitiva e guiar de forma mais eficiente o processo de tomada de decisão. Este estudo se propôs a responder questões relacionadas a qualidade dos dados em um ambiente de alto volume de dados, com ênfase no processo de ETL através da metodologia proposta por Munawar (2021). Conforme a proposição, a partir da análise teórica e aplicação prática no *Data Lakehouse* destacou-se as principais vantagens e desafios dessa implementação. De um lado foi apresentada capacidade de prevenção a problemas de qualidade e possibilidade de ajustes imediatos nos processos de ETL para melhor atender as particularidades oferecidas pelos dados quando são alterados em seu ciclo de vida. De outro lado, observou-se que a aplicação estrita dessa metodologia em ambientes de *Big Data* não oferece desempenho e potencialmente pode agravar problemas relacionados à velocidade e volume em que os dados são recebidos e precisam ser processados. Por fim, compreende-se que essa aplicação possui importância para o ambiente de *Big Data*, especialmente o *Data Lakehouse*, mas sugere-se melhorias para melhor atender as características necessárias para se trabalhar com esses dados nesse ambiente.

5.1 Contribuições do trabalho

O estudo demonstrou que a metodologia proposta inicialmente por Munawar (2021) e aplicada sobre um ambiente de *Big Data* atende de forma eficaz as questões relacionadas aos problemas de qualidade de dados. Através dessa metodologia foi possível identificar problemas de qualidade na fonte de dados como dados incompletos. A incorporação da QD no ETL se destaca como uma ferramenta potencial para integração contínua, aumentando a confiabilidade na implementação de regras de negócio durante a transformação dos dados, apresentando assim a importância da inclusão do processo de qualidade de dados no processo de ETL em ambiente de *Big Data* como o *Data Lakehouse*.

A incorporação da QD no ETL e a separação de dados com e sem qualidade aumenta a confiabilidade das camadas subsequentes no *Data Lakehouse* visto que a fonte de dados de uma determinada camada sempre é a anterior. Essa prática evita a propagação de dados sem qualidade nas camadas do *Data Lake* que poderiam ofertar um cenário complexo de detecção da causa do problema de qualidade. Em ambientes de Big Data, que aplicam a QD após o processo de ETL, pode proporcionar casos onde a correção da qualidade de dados seja impraticável, devido ao grande volume de dados, transformações em diversas camadas e necessidade de processamento.

No entanto, este procedimento que separa dados com e sem qualidade pode impactar negativamente em outras atividades analíticas que poderiam se valer de colunas que não infringem regras de qualidade. Além disso existe um impacto no desempenho do processamento dos dados, acrescentando um tempo adicional para entrega dos dados com qualidade na camada final, agravando assim problemas impostos pelo cenário de Big Data que tem relação com a velocidade de recebimento do dado e volume de dados.

5.2 Proposta de Melhoria

A utilização de sistemas de *Big Data*, em especial as tecnologias que interagem com os sistemas de arquivos, poderiam ser utilizadas para aumentar a eficiência da metodologia proposta por Munawar (2021) em ambientes de *Data Warehouse*. Portanto, é sugerido que em ambientes de *Big Data* haja somente uma separação lógica ao invés de física entre os dados com e sem qualidade. Ao invés de gerar uma tabela com o objetivo de armazenar dados inconsistentes, há a oportunidade de utilizar técnicas de particionamento de tabelas para esse objetivo. Dessa forma, a metodologia proposta poderia ter mais eficiência no ambiente de *Big Data* através da simplificação do processo de carregamento dos dados, evitando a operação de separação e carregamento dos dados corretos e incorretos. Essa otimização também evitaria o problema de privação do uso dos dados em relação a outras atividades que poderiam valer-se de colunas que não infringem regras de qualidade.

Outra sugestão que contrapõe a regra estrita de remover a linha inteira do conjunto de dados quando encontrado um problema de qualidade pode ser traduzida na remoção da coluna do conjunto de dados. Dessa forma, poderia ser estabelecido um limite percentual que uma coluna pode apresentar problemas de qualidade para que seja propagada ou não para camadas subsequentes do *Data Lakehouse*. No estudo proposto foram identificadas colunas que apresentavam 80% de problemas de qualidade em relação a completude. Essas colunas poderiam ser removidas e armazenadas em outro repositório para análise posterior da origem dessas ocorrências. Assim, seria evitado o problema de privação de dados com qualidade úteis para atividades analíticas em outras camadas.

Por fim, há a necessidade de se estabelecer uma tratativa para os dados que são armazenados nas tabelas de erros. No trabalho proposto por Munawar (2021) não há menção do que deve ser feito com estes dados após sua escrita nas tabelas de erros. Portanto, é necessário estabelecer um processo de avaliação, tratativa e, caso haja possibilidade, reinserção desses dados nas suas tabelas com dados corretos. Isso atenuaria o problema de privação das colunas com dados de qualidade que poderiam ser utilizados para demais atividades analíticas visto que a metodologia propõe a remoção total da linha do dado com alguma coluna sem qualidade.

5.3 Trabalhos futuros

Para trabalhos futuros propõe-se submeter essa metodologia em cenários que podem abranger outras características de *Big Data* não contempladas neste trabalho. Portanto, valer-se de volumes maiores de dados, dados não estruturados e cenários de processamento em *near-real-time*. A execução desta metodologia em outras arquiteturas e sistemas de *Big Data* afim de ampliar a compreensão de sua eficácia e adaptabilidade também se mostra necessária. Além disso, propõe-se estudar a viabilidade da separação lógica dos dados com e sem qualidade, ou seja, a através do particionamento de tabelas para verificação da performance dessa metodologia em ambientes de alto volume de dados. Por fim, o aprofundamento na criação e utilização de ferramentas automatizadas que poderiam facilitar a incorporação da qualidade dos dados em processos de ETL em *Big Data* e aumentando sua eficiência de execução.

REFERÊNCIAS BIBLIOGRÁFICA

CAI, L.; ZHU, Y. **The Challenges of Data Quality and Data Quality Assessment in the Big Data Era**, CODATA Data Science Journal, 14, 2, p. 1-10. 2015.

HENDERSN, D.; EARLEY, S., Data Administration Management Association **DAMA-DMBOK: data management body of knowledge**, Second edition. ed. Technics Publications, Basking Ridge, New Jersey, 2017. 624p.

ABDULLAH, N.; ISMAIL, S.; SOPHIAYATI, S.; SAM, S. **Data Quality in Big Data: A Review**, Int. J. Advance Soft Compu. Appl, Vol. 7, No. 3, p. 16-27. 2015.

BECKER, D.; KING, T.D.; MCMULLEN, B. **Big data, big data quality problem**, 2015 IEEE International Conference on Big Data (Big Data), pp. 2644–2653. 2015.

RAMCHAND, S.; MAHMOOD, T. **Big data architectures for data lakes: A systematic literature review**, 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 1141–1146. 2022.

MUNAWAR, **Extract Transform Loading (ETL) Based Data Quality for Data Warehouse Development**, 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), pp. 373–378. 2021.

ORESCANIN, D.; HLUPIC, T. **Data Lakehouse - a Novel Step in Analytics Architecture**, 44th International Convention on Information, Communication and Electronic Technology (MIPRO), pp. 1242–1246. 2021.

ERRAMI, S.; HAJJI, H.; KADI, K.; BADIR, H. Spatial big data architecture: From Data Warehouses and Data Lakes to the LakeHouse. **Journal of Parallel and Distributed Computing 176**, p. 70–79. 2023.

ENGEL, G.I. **Pesquisa-ação**, Educar, Curitiba, n. 16, p. 181-191. 2000.

ARMBRUST, M.; GHOODSI, A.; XIN, R.; ZAHARIA, M. **Lakehouse**: **A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics**, 11th Annual Conference on Innovative Data Systems Research (CIDR ’21), January 11–15, 2021.

HARBY, A.A.; ZULKERNINE, F. **From Data Warehouse to Lakehouse: A Comparative Review**, 2022 IEEE International Conference on Big Data (Big Data), pp. 389–395. 2022.

HLUPIC T.; ORESCANIN D.; RUZAK D.; BARANOVIC M. **An Overview of Current Data Lake Architecture Models**, 2022 45th Jubilee International Convention on

Information, Communication and Electronic Technology, MIPRO 2022 – Proceedings, p. 1082-1087. 2022.

WANG, R.Y.; STRONG, D.M. **Beyond accuracy: what data quality means to data consumers**, Journal of Management Information Systems, Vol.12, No.4, pp5-34. 1996.

TALEB I.; KASSABI H.; SERHANI M.; DSSOULI R.; BOUHADDIOUI C. **Big Data Quality: A Quality Dimensions Evaluation,** 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, p. 759-765. 2016.

STRONG, D.M.; LEE, Y.W.; WANG, R.Y. **Data Quality In Context**, Commun. ACM 40, p. 103-110. 1997.

FIRMANI, D.; MECELLA, M.; SCANNAPIECE, M.; BATINI, C. **On the Meaningfulness of "Big Data Quality"** (Invited Paper). Data Sci. Eng. 1, p. 6–20. 2016.

CARLO, B.; DANIELE, B.; FEDERICO, C.; SIMONE, G. **A Data Quality Methodology for Heterogeneous Data**. IJDMS 3, p. 60–79. 2011.

TALEB, I.; SERHANI, M.A.; DSSOULI, R. **Big Data Quality Assessment Model for Unstructured Data**, 2018 International Conference on Innovations in Information Technology (IIT), p. 69–74. 2018.

BING, C.; XUCHU, W.; BEIZHAN, W.; XUEQIN H. **Analysis and solution of data quality in data warehouse of Chinese materia medica**, 2009 4th International Conference on Computer Science & Education, p. 823-827. 2009.

AHMAD, K.; IDRIS, N. **Managing Data Source Quality for Data Warehouse in Manufacturing Services**, 2011 International Conference on Electrical Engineering and Informatics 17-19 July 2011, Bandung, Indonesia, 2011.

FERREIRA, L.M. **MODELO DE PROCESSO PARA CRIAÇÃO DE BI EM BANCO DE DADOS NOSQL ORIENTADO A COLUNAS,** Conferências Ibero-Americanas WWW/Internet e Computação Aplicada (CIACA 2015), p. 315-320. 2015.

MIRANDA, M.; FERREIRA, J.; ABELHA, A.; Machado, J. **O Processo ETL em Sistemas Data Warehouse**, INForum 2010 - II Simpósio de Informática, Luís S. Barbosa, Miguel P. Correia (eds), 9-10 Setembro, 2010, p. 757–765. 2010.

KRAFTA, L.; FREITAS, H.; MARTENS, C.D.P.; ANDRES, R. **O método da pesquisa-ação: um estudo em uma empresa de coleta e análise de dados**, 2007 Revista Quanti & Quali. Disponível em: http://www.

faccat.br/download/pdf/posgraduacao/profaberenice/09pesquisa_ acao_2009_3.pdf. Acesso em: 21 Jan. 2024.