

A IMPORTÂNCIA DA QUALIDADE DA INFORMAÇÃO PARA ETL

Christiano Roofn Menezes de Oliveira

<https://orcid.org/0009-0006-4520-0331>

Gleison Soares de Moura

<https://orcid.org/0009-0002-8229-5086>

Marcus Vinicius Freire Junior

<https://orcid.org/0009-0008-3132-0142>

Ricardo Esteves Kneipp

<https://orcid.org/0000-0003-4899-1731>

Shirley Nunes Costa Santos

<https://orcid.org/0009-0003-5859-4446>

Simara Terra de Souza Ferreira

<https://orcid.org/0009-0005-9029-953X>

RESUMO

O artigo trata da importância da qualidade da informação no momento que antecede a aplicabilidade da ETL (Extract, Transform and Load). Tem uma abordagem aplicada com bases teóricas através de investigação das principais características de qualidade da temática. Sendo assim, será realizado a partir de uma pesquisa exploratória para identificar informações relevantes que precisam de ações específicas no tratamento dos dados com a utilização das etapas do processo ETL. A Metodologia é mista, porém levando em consideração mais elementos qualitativos do que quantitativos acerca das literaturas existentes. Como resultado, insights que nortearão a coleta dos dados de maior relevância para extração, transformação e carregamento.

Palavras-chave: Qualidade de Informação. ETL. Extração de Dados. *Insights*.

1 INTRODUÇÃO

Com o avanço da necessidade de uma informação cada vez mais precisa, no momento certo e que consiga suportar por completo o que realmente precisamos, depara-se com uma dependência de dados cada vez mais assertivos e de qualidade.

A fase preliminar de manipulação de dados, que envolve a identificação de fontes de informação confiáveis, é crucial no contexto da tomada de decisão orientada a dados. Pesquisas, como as relacionadas no período seguinte, enfatizam a importância de dados de alta qualidade. No entanto, a disseminação rápida de informações muitas vezes supera a preocupação com sua veracidade comprometendo a integridade dos dados. Mesmo com a implementação eficaz de um processo de Extração, Transformação e Carregamento (ETL), a presença de dados falsos ou imprecisos pode resultar em conclusões errôneas, causando, principalmente, decisões equivocadas, em alguns casos, prejuízos financeiros.

Segundo Redman (2001), se as tecnologias da informação são os motores da Era da Informação, então os dados e a informação são os combustíveis. De acordo com Machado (2018), a gestão das informações está diretamente relacionada com a qualidade das informações. Gradativa, e mais intensamente nos últimos anos, as instituições estão

tomando consciência da importância da qualidade nas informações e, sobretudo, dos prejuízos que podem advir da falta dessa qualidade.

Corroborando com o raciocínio da possibilidade de prejuízo, na ótica de finanças, Miller (2002) ressalta que uma informação de baixa qualidade pode gerar quatro axiomas econômicos. Tendo em vista que a informação incompleta cria a incerteza; a incerteza, por sua vez, cria risco para os investidores e os credores; o risco faz com que a ausência da demanda de investidores e credores fique com uma alta taxa de atratividade. Uma alta taxa de atratividade para os investidores e credores faz com que aumente o custo de capital da empresa. Ou seja, uma informação com pouca qualidade afeta todos os segmentos de uma organização, culminando em perdas, inclusive financeiras.

2 DESENVOLVIMENTO

2.1 Dados, Informação e Qualidade

Para contextualizar e embasar o estudo é necessário obter um mínimo de entendimento dos termos e definições utilizadas no contexto tecnológico envolvendo Business Intelligence, contexto onde este assunto desenvolve-se. A seguir, faremos um breve relato das principais definições envolvidas. Tendo em vista que dados, informação e qualidade podem ser definidos por diversas visões, derivações e conceitos.

Segundo Redman (2001), existem dois componentes inter-relacionados, “modelos de dados” (se trata por diferentes aspectos) e “valores de dados” (atributos para entidades específicas).

Dados podem ser definidos como elementos com possibilidades de serem coletados, armazenados e processados em diferentes formatos advindos de diversas fontes, podendo ser qualitativos (características) e quantitativos (numéricos).

Informação pode ser definida como o resultado do processamento dos dados, ou seja, a interpretação estruturada dos dados, independentemente de ser novas, já conhecidas, subjetivas, factuais, ou seja, um recurso de extremo valor a ser utilizado principalmente nas tomadas de decisões;

Qualidade pode ser definida em relação às percepções, necessidades e resultados principalmente no que diz respeito à excelência, perfeição ou adequação em determinado padrão. Para o nosso objetivo, resume-se pela precisão e integridade de dados e ainda pela relevância.

2.2 Qualidade de dados *versus* qualidade da Informação

Apesar de terem aspectos similares na sua composição e um completar o outro, a qualidade de dados e a qualidade de informação possuem particularidades.

A qualidade dos dados visa atender a requisitos como precisão, integridade, validade, consistência, exclusividade, atualidade e utilidade. Essa característica dos dados permite a tomada de decisões mais precisa, de forma a identificar tendências, bem como prever resultados e melhorar a eficiência.

Em Redman (2001), a qualidade dos dados é definida como a adequação dos dados para fins definidos pelo usuário e sendo subjetiva tendo em vista atender a padrões estabelecidos pelas expectativas do usuário final.

A qualidade de informação refere-se à excelência, confiabilidade e utilidade da informação disponível, indo além dos dados brutos, considerando como os dados são processados, interpretados, com o intuito de produzir conhecimento significativo.

Segundo Machado (2013), a qualidade da informação é determinada pela informação que foi processada, levando em consideração tanto o valor esperado quanto o valor percebido e o valor agregado, todos apresentados de maneira classificada e contextualizada.

2.2 Dimensionamento da qualidade da informação

Pelo dimensionamento da qualidade da informação, temos a medição da qualidade dos dados como sendo o processo de avaliação e a definição do nível de qualidade necessário para satisfazer as necessidades das partes interessadas.

Refere-se à uma avaliação e gestão da qualidade da informação e dos dados em uma organização. Podendo ser implementada de diversas maneiras a depender da combinação dos dados das análises qualitativa e quantitativa, assim, identificando os pontos fortes e fracos dos repositórios.

Alguns autores dimensionam a qualidade da informação com base em diferentes técnicas; Manel *et al.* (2019) dimensiona pela duplicação, completude, precisão, interpretabilidade, concisão e consistência; para Lee (2003) o dimensionamento refere-se à integridade, interpretabilidade, acessibilidade e relevância.

Todos os identificadores de dimensão, no sentido direto de avaliação, dependem da quantidade total de dados gravados, por outro lado, entende-se que também dependem da interpretação de cada dimensionador, como por exemplo na tabela abaixo.

Tabela 1: Identificadores de dimensão

Dimensionamento	Problema
Atualidade	Informações desatualizadas
Completude	Dados incompletos
Confiabilidade	Dados não confiáveis
Conformidade	Fora dos padrões
Consistência	Registros com inconsistência
Credibilidade	Dados sem credibilidade
Disponibilidade	Não acessíveis quando necessário
Integridade	Tendo problemas de integridade
Legibilidade	Falta de clareza e compreensibilidade
Valor	Não valoração adequada
Oportunidade	Falta de pontualidade
Precisão	Registros imprecisos
Rastreabilidade	Sem poder rastrear a origem
Relevância	Falta de importância que merece
Singularidade	Registros não únicos
Validade	Dados inválidos

Fonte: Autoria própria (2023).

Desta forma, completa-se a dualidade quantidade mensurável e qualidade mensurável, deixando claro o objetivo em conseguir dados não passíveis de erro independentemente do tipo de dimensão aplicado.

2.3 Tratamento de dados não coesos

Enfrentar a problemática do tratamento de dados não coesos envolve desafios diversos, sendo a padronização uma questão central. A falta de uniformidade nos formatos e estruturas de bancos de dados prejudica uma análise coerente, destacando a necessidade de estabelecer padrões para assegurar a consistência da informação.

Em projetos de criação de bancos de dados, como por exemplo os alimentados por um ERP, a não adaptação do sistema para extração de dados frequentemente resulta na perda de rastreabilidade ao longo do tempo, gerando lacunas prejudiciais.

A falta de correlação entre informações, em diferentes bancos de dados, é outro obstáculo significativo, impedindo a obtenção de *insights* abrangentes. Superar essas complexidades requer não apenas tecnologias avançadas, mas também uma mudança cultural que priorize a integridade e a consistência dos dados em todas as fases dos processos corporativos. De acordo com Inmon, Strauss e Neushloss (2007), os dados estruturados são aqueles armazenados sempre da mesma forma e *layout*.

3 CONCLUSÃO

Portanto, torna-se evidente que a precisão e confiabilidade dos dados são cruciais para a eficiência e durabilidade das empresas na era da informação. A necessidade urgente de lidar com a qualidade dos dados desde sua origem até o processo decisório é destacada pela essencialidade de informações exatas, combinada com a rápida disseminação dessas informações.

A análise das implicações econômicas decorrentes da baixa qualidade dos dados, conforme apontado por Miller (2002), destaca significativas consequências. A informação de qualidade inferior gera incerteza, riscos para investidores e credores, resultando no aumento do custo de capital da empresa. Isso ressalta a premente necessidade de garantir a integridade e precisão dos dados ao longo do processo ETL.

As definições fornecidas por Redman (2001) para dados, informação e qualidade, contribuem para uma compreensão mais aprofundada da complexidade desses elementos. A diferenciação entre qualidade de dados e qualidade de informação enfatiza a importância de abordagens específicas para garantir a excelência em ambas as áreas.

A avaliação qualitativa e quantitativa do dimensionamento da qualidade da informação revela-se uma ferramenta valiosa para identificar pontos fortes e fracos nos repositórios de dados. Diferentes abordagens dimensionais, conforme proposto por Manel *et al.* (2019) e Lee (2003), oferecem uma visão abrangente, considerando não apenas a quantidade, mas também aspectos como duplicação, completude, precisão e relevância.

No enfrentamento de dados não coesos, a padronização surge como elemento central na busca por consistência e integridade. A falta de uniformidade nos formatos e estruturas de bancos de dados, aliada à perda de rastreabilidade em projetos complexos, destaca a necessidade de uma abordagem que una aspectos culturais e tecnológicos para assegurar a qualidade dos dados em todas as fases dos processos corporativos.

Em resumo, este estudo reitera a tese de que a qualidade dos dados é essencial para a sustentabilidade e sucesso organizacional por meio do processo ETL. A conscientização sobre a importância da qualidade dos dados deve ser incorporada como um princípio orientador em todas as esferas das operações empresariais. A busca incessante pela excelência na gestão de dados é, portanto, crucial para garantir a confiabilidade e eficácia das decisões estratégicas.

REFERÊNCIAS

LEE, Shuang-lin. **Data Quality in Organizational Context-- A Case Study.** A Master's paper for the M.S. in I.S. degree. May, 2003.

MACHADO, Osmar A. **Qualidade da Informação: Uma abordagem orientada para o contexto.** São Paulo: Novas Edições Acadêmicas, 2018.

MANEL, Souibguia et al. *Data quality in ETL process: A preliminary study.* In: **23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems**, 2019.

MILLER, Paul B. W.; BAHNSON, Paul R. **Quality Financial Reporting.** New York: Mc Graw-Hill, 2002.

REDMAN, T.C. **Data quality: the field guide.** Boston, MA: Digital Press, 2001.

INMON, William; STRAUSS, Derek; NEUSHLOSS, Genia. **DW 2.0: the architecture for the next generation of Data Warehousing.** 2007.