Analisador de Estruturas de Bases de Dados para o Agronegócio com Dimensões de Qualidade de Dados

Clovis S. Junior[1], **Carina F. Dorneles**[2]

[1]Instituto de Ciencias Exatas e Naturais - ICEN
Universidade Federal de Rondonólis - UFR/Rondonópolis-MT

[2]Departamento de Informatica e Estatística - INE
Universidade Federal de Santa Catarina - UFSC/Florianópolis-SC

clovis@ufr.edu.br, carina.dorneles@ufsc.br

Abstract. *This paper presents a tool for the data quality analysis of databases. The expected result is an interface to assist in the verification and analysis of data. The solution was developed by combining data dictionaries and stored data capabilities, enabling the investigation and analysis of relevant aspects based on data quality criteria. The objective is to provide qualitative results regarding the data structure, presenting faulty points such as integrity, objectivity, validity, and others.*

Resumo. *Este artigo apresenta uma ferramenta para análise da qualidade de dados de bases de dados, cujo objetivo é fornecer uma interface para auxiliar na verificação e análise dos dados. A solução foi desenvolvida combinando recursos de dicionário de dados e dos dados armazenados possibilitando a investigação e análise de aspectos relevantes para análise baseada em critérios de dimensões de qualidade de dados. Com isso, pretende-se fornecer resultados qualitativos referentes à estrutura na qual os dados estão armazenados indicando características inadequadas como integridade, objetividade e validade entre outras dimensões*[1].

1. Introdução

Pesquisas relacionadas à qualidade de dados na agricultura abordam várias questões como gerenciamento e estatísticas agrícolas relacionadas com produção, comercialização, estoques, logística e aplicações de insumos. Neste contexto, várias fontes de dados precisam ser consideradas, tais como arquivos contendo dados coletados por sensores, características da cultura, informações do solo e gestão de recursos humanos. A coleta e uso de dados são a origem de alguns problemas de qualidade que precisam ser considerados na geração de dados adequados para o agronegócio. Por exemplo, dados armazenados inadequadamente ou de forma incompleta, cobertura incompleta de dados e conceitos imprecisos são problemas comuns enfrentados na origem dos dados [Malaverri and Medeiros 2012].

O monitoramento da evolução de bases de dados quanto à qualidade usualmente é realizada de forma manual, seja estruturalmente ou analisando as interfaces responsáveis

[1] $Acessoao Screencast demonstrativohttps://youtu.be/iVRQh3x_UOI$

por entradas e saídas de dados. Nos dois casos há pouca automação, pois são processos que requerem intervenção humana. Normalmente há setores especıficos para essa tarefa relacionada ao controle de qualidade (CQ) [Knauer et al. 2020]. O conceito de dimensões em qualidade de dados está relacionado à identificação de medidas de qualidade relacionadas a elementos de dados, incluindo atributos, registros, tabelas, sistemas ou agrupamentos mais abstratos, como unidades de negócios, empresas ou gamas de produtos [Nasr et al. 2020]. Monitorar a qualidade das bases de dados depende das dimensões que serão monitoradas. Para avaliar a qualidade de dados em um determinado domınio é necessário analisar as dimensões que serão avaliadas, pois cada base de dados tem caracterısticas específicas indicando o uso de parˆametros sob medida para avaliação.

Este artigo apresenta um protótipo do projeto Hero[2], cujo objetivo é realizar o monitoramento da qualidade de dados utilizando dimensões obtidas após investigação de diferentes trabalhos na literatura, totalizando 53 diferentes dimensões de qualidade. Utilizar todas as dimensões poderia resultar em ineficiência, pois nem todas são adequadas para o agronegócio. Assim, foi realizada uma consulta a profissionais da área técnica associados ao agronegócio, sendo: dois gerentes de tecnologia, um coordenador de projetos de software e um desenvolvedor de aplicações. Foi elaborado questionário para coleta de dados por meio eletrônico, inicialmente sendo exposto que eles deveriam informar sua percepção em relação à relevância das dimensões ou requisitos de qualidade de dados relacionados ao agronegócio. O protótipo apresentado neste artigo tem por objetivo auxiliar o processo de monitoramento de bases de dados usando dimensões de qualidade de dados indicadas para o agronegócio, conforme discutido na Seção 3. O auxílio proporcionado pelo protótipo refere-se a relatórios para acompanhamento do comportamento dos dados e estrutura de armazenamento confrontada com as dimensões indicadas.

O artigo está organizado como segue. Na Seção 2 são descritos os conceitos de dimensões de qualidade de dados, usados para criar o protótipo. A Seção 2, apresenta uma visão geral da ferramenta com funcionalidades. A seção 3 apresenta detalhes técnicos referente a implementação e interfaces do protótipo. Na Seção 4, são apresentadas as conclusões e discutidos trabalhos futuros.

2. Metodologia utilizada

A fase atual do desenvolvimento do *Hero* foi feita a partir da coleta de dados em literatura específica e validações em reuniões realizadas junto a pessoal técnico. A verificação das dimensões de qualidade de dados foi realizada em três etapas, a primeira refere-se a conceitos obtidos em literaturas relacionadas com o agronegócio com propósito de identificar dimensões da forma mais abrangente possível independente do cenário que as mesmas são aplicadas. A segunda etapa foi realizada junto a técnicos do agronegócio, com o objetivo de identificar as dimensões mais adequadas para o agronegócio, conforme apresentado na sessão 3 sob uma respectivas técnicas de profissionais de tecnologia da informação. A terceira etapa refere-se a criação do protótipo utilizando as dimensões indicadas para o agronegócio. A versão atual do protótipo foi desenvolvida com a linguagem Embarcadero® Delphi 10.4.

[2]Alusão ao resgate da qualidade dos dados

3. Projeto Hero

A arquitetura do Hero é apresentada na Figura 1 e ilustra de forma geral os componentes da aplicação. A versão atual utiliza uma conexão direta com o sistema gerenciador de banco de dados, verificando as estruturas necessárias para a investigação de cada dimensão de qualidade de dados utilizando apenas as instâncias de dados e o dicionário de dados. O protótipo foi testado com três bases de dados contendo características distintas, conforme abordadas na Seção 3. O código do projeto Hero está disponível em https://github.com/clovissjunior/hero.git.

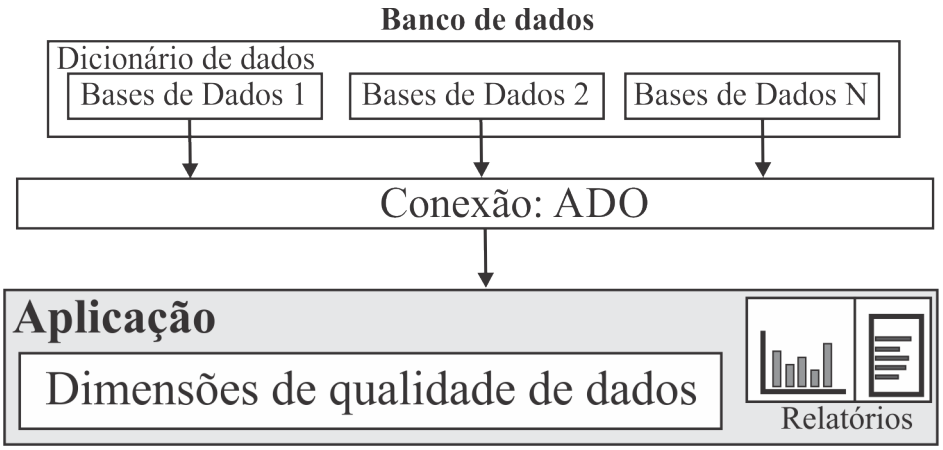

Figura 1. Arquitetura do protótipo.

3.1. Funcionalidades implementadas

As funcionalidades do protótipo foram implementadas usando as avaliações dos especialistas em relação às dimensões mais indicadas para o agronegócio, sob o ponto de vista deles. Durante a avaliação com os especialistas, as dimensões de qualidade de dados foram identificadas utilizando uma escala de valores entre 1% e 100%, para as 53 dimensões. Foi estabelecido como métrica que as dimensões com valores=90% seriam consideradas adequadas. Após a coleta de dados junto ao pessoal técnico foram identificadas 10 dimensões adequadas aos parâmetros definidos. As dimensões identificadas foram: atualização, auditabilidade, disponibilidade, credibilidade, consistência, fundamentos de integridade de dados, especificação de dados, eficiência, integridade e validade.As análises utilizadas para cada dimensão resultam em visualizações individuais com a possibilidade de criação de relatório com resumo de todas as análises, permitindo uma visão geral a respeito da situação da base de dados investigada e também a análise de cada dimensão. As dimensões utilizadas no Projeto *Hero* são:

- Atualização: verifica o grau de atualização em dias usando metadados do próprio banco de dados, períodos longos sem atualização são inadequados [Sidi et al. 2012].
- Auditabilidade: verifica o percentual de entidades habilitadas para auditoria, a verificação é feita calculando o percentual de tabelas com possibilidade de auditoria [Cai and Zhu 2015].
- Disponibilidade: verifica a quantidade de atributos obrigatórios e não obrigatórios, a quantidade excessiva de atributos não obrigatórios pode resultar em conjunto de dados pobres [Malaverri and Medeiros 2012].

- Credibilidade: verifica se os atributos possuem nomes indicando o seu significado. A verificação é feita em um conjunto pequeno de termos, mas está sendo expandido de acordo com a demanda identificada [Sidi et al. 2012].
- Consistência: verifica sobrecargas no armazenamento de atributos textuais comparando o tamanho utilizado em cada coluna em relação a capacidade de armazenamento do domínio definido [Sidi et al. 2012].
- Fundamentos de integridade de dados: verifica o grau de preenchimento de atributos não obrigatórios. O cálculo proposto verifica o percentual de atributos não obrigatório dicionário de dados em relação ao todo [Sidi et al. 2012].
- Especificação de dados: verifica a disponibilidade de regras de integridade em tabelas e seus atributos de acordo com a quantidade de restrições impostas às tabelas [Sidi et al. 2012].
- Eficiência: verifica a consistência de nomenclaturas na definição de atributos a partir de um conjunto de palavras reservadas para identificar a consistência de cada atributo disponível nos metadados do domínio [Sidi et al. 2012].
- Integridade: verifica a consistência relacionada a de nomes abreviados. A verificação é feita calculando o percentual de nomes abreviados presentes em atributos associados ao armazenamento de textos de forma geral [Cai and Zhu 2015].
- Validade: verifica a consistência de atributos quanto aos seus requisitos mínimos de existência [Sureddy and Yallamula 2020], foram utilizados 2 atributos para simulação: Nome=10 elementos CPF=11 elementos.

4. Estudo de Caso

A validação do protótipo foi realizada com três bases de dados reais, a primeira com foco no meio ambiente (Bd-Ambiental), a segunda referente a dados de propriedades rurais destinadas a agricultura familiar (Bd-Agricultura) e a terceira utilizada por um sistema comercial para ERP (Bd-Erp).

- Bd-Agricultura (ag). Trata-se de uma base de dados utilizada no projeto "Sistema para Coleta de Dados Socioeconômicos em Comunidade Rurais". A base de dados possui 74 tabelas e um volume de 13,19 MB de dados.
- Bd-Ambiental (am). Diz respeito à base de dados do projeto "Sistema para Gerenciamento de Planos de Recuperação em Áreas Degradadas, como estratégia de consolidação do Novo Código Florestal Brasileiro". A base de dados possui 62 tabelas e um volume de 8,25 MB de dados.
- Bd-Erp (er). A base de dados é utilizada por um ERP comercial com recursos como: negociações, logística, comercialização, contabilidade, fiscal, almoxarifado entre outros. A base de dados possui 892 tabelas e um volume de 178,16 MB de dados.

4.1. Interface de visualização

Nesta seção, são apresentadas as interfaces disponibilizadas para avaliação da qualidade de dados das dimensões propostas. As interfaces apresentadas nas Figuras 2, 3 e 4 mostram os resultados da avaliação por meio de gráficos comparativos entre as duas bases de dados testadas, Bd-ag, Bd-am e Bd-er. A implementação do protótipo foi feita utilizando interface gráfica (GUI), conforme observado nas figuras, com recurso para conexão direta para gerenciadores de banco de dados. As interfaces apresentam as dez dimensões de

qualidade de dados abordados na pesquisa, algumas dimensões foram implementadas utilizando mais de um critério, em alguns casos a adoção de uma fórmula única não tornaria possível a implementação de forma satisfatória. As dimensões são acionadas através de guias, e para cada uma delas são apresentados o objetivo, a fórmula usada para avaliação daquela dimensão e o gráfico correspondente ao resultado da avaliação sobre um determinado conjunto de dados.

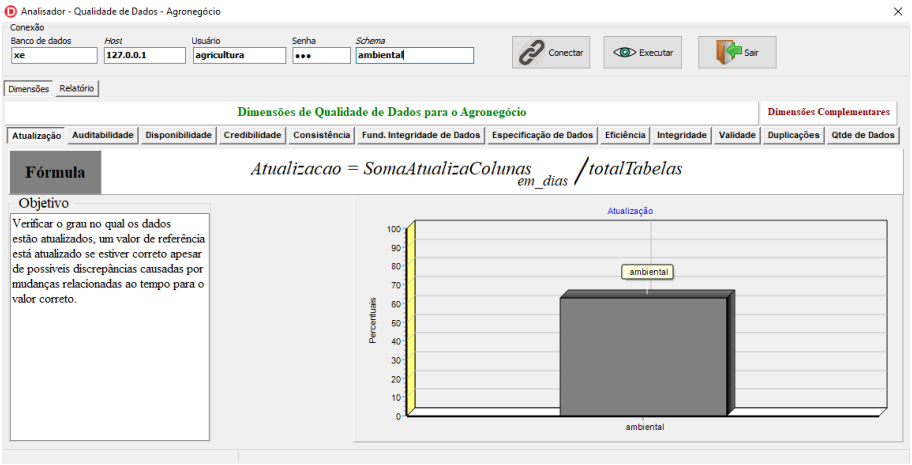

Figura 2. Interface principal para visualização dos resultados.

Além das informações gráficas disponibilizadas pelo protótipo, há também a implementação de relatório sintético de todas as dimensões avaliadas na base de dados indicada com os respectivos pontos positivos e negativos, conforme mostra a Figura 3.

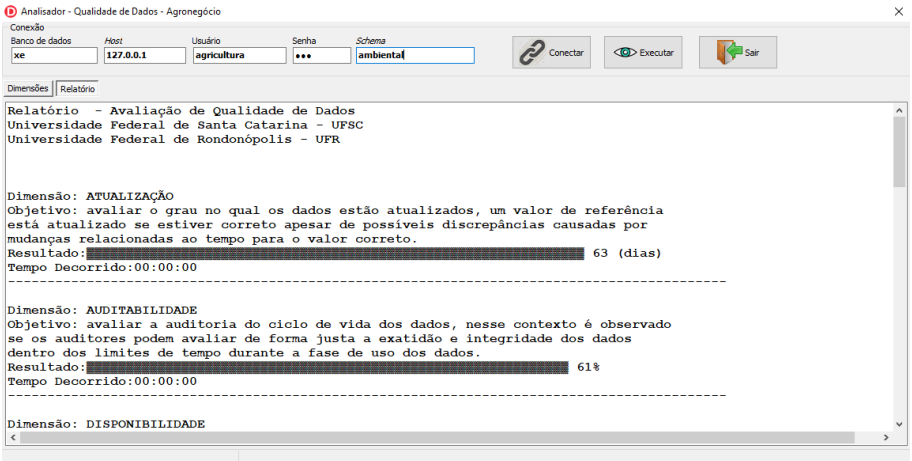

Figura 3. Relatório textual sintético.

Além das dimensões definidas com foco específico para o agronegócio foram implementados também duas dimensões complementares referente a duplicação de dados e quantidade de dados em uso. As dimensões complementares, mostradas na Figura 4, apesar de não terem sido indicadas na investigação inicial foram acrescentadas em razão de serem fortemente indicadas em diversas pesquisas como [Nasr et al. 2020] e [Malaverri and Medeiros 2012].

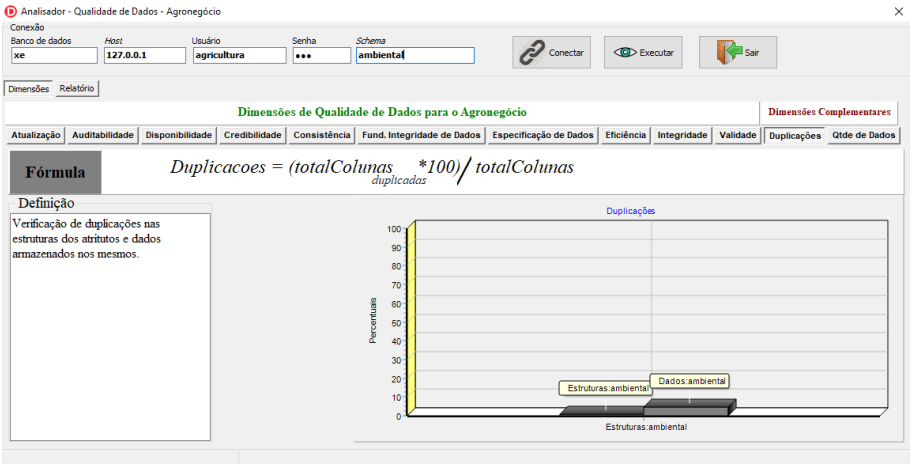

Figura 4. Dimensões complementares.

5. Conclusões e Trabalhos futuros

O protótipo está em fase avançada de desenvolvimento com resultados satisfatórios para realizar a avaliação proposta utilizando dimensões de qualidade de dados específicos para o agronegócio. Outras funcionalidades serão agregadas para melhorar os resultados de cada dimensão adicionando mais recursos. Como trabalhos futuros, pretende-se verificar a necessidade de agregar novas dimensões de qualidade de dados as análises atuais e novas análises as dimensões utilizadas, tornando o aplicativo mais preciso e eficiente quanto ao apontamento de melhorias ou enriquecimentos em bases de dados. Além disso, a versão atual do protótipo ainda limita-se a tipos específicos de banco de dados, entretanto o protótipo está evoluindo para se tornar uma ferramenta multibanco. Dessa forma, será possível a realização de análises nos principais gerenciadores de banco de dados. Essa limitação deve-se às características específicas de cada gerenciador como acesso aos dicionários de dados e outros objetos para realização de engenharia reversa e posteriormente execução de análises.

Referências

Cai, L. and Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14:2.

Knauer, T., Nikiforow, N., and Wagener, S. (2020). Determinants of information system quality and data quality in management accounting. *Journal of Management Control*, 31.

Malaverri, J. and Medeiros, C. (2012). Data quality in agriculture applications. *Proceedings of the Brazilian Symposium on GeoInformatics*, pages 128–139.

Nasr, M., Shaaban, E., and Gabr, M. I. (2020). Data quality dimensions. In *Internet of Things—Applications and Future*, pages 201–218. Springer.

Sidi, F., Shariat Panahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H., and Mustapha, A. (2012). Data quality: A survey of data quality dimensions. In *2012 International Conference on Information Retrieval Knowledge Management*, pages 300–304.

Sureddy, M. R. and Yallamula, P. (2020). Data quality architecture for data warehouses. *International Journal of Research Cultural Society*, 4(6):95–100.