Ferramentas open-source de qualidade de dados para licitações públicas: Uma análise comparativa

Gabriel P. Oliveira[1], Arthur P. G. Reis[1], Bárbara M. A. Mendes[1], Clara A. Bacha[1], Lucas L. Costa[1], Gabriel L. Canguçu[1], Mariana O. Silva[1], Victor Caetano[1], Michele A. Brandão[1,2], Anisio Lacerda[1], Gisele L. Pappa[1]

[1]Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG, Brasil

[2]Instituto Federal de Minas Gerais (IFMG) – Ribeirão das Neves, MG, Brasil

gabrielpoliveira@dcc.ufmg.br, {arthurpetrocchi,barbaramit}@ufmg.br
{clarabacha,lucas-lage}@ufmg.br, {gabriel.lima,mariana.santos}@dcc.ufmg.br
{victor.caetano,michele.brandao,anisio,glpappa}@dcc.ufmg.br

Abstract. *Data have been increasingly used as decision support in different contexts. For these decisions to be reliable, it is necessary to ensure data quality. In this context, this work presents a brief comparison of eight open-source data quality tools. We then choose one tool for analyzing an actual data warehouse formed by public bids. Finally, our analyses show that the Great Expectations tool has relevant characteristics to generate good data quality indicators, thus ensuring that public bidding data can help in the decision-making process.*

Resumo. *Dados são cada vez mais utilizados como suporte à decisão em diferentes contextos. Para que essas decisões sejam confiáveis e precisas, é necessário garantir a qualidade dos dados. Nesse contexto, este trabalho apresenta um breve comparativo de oito ferramentas open-source de qualidade de dados, e justifica a escolha de uma delas para análise de um armazém de dados reais e volumosos formado por licitações públicas. Finalmente, nossas análises mostram que a ferramenta Great Expectations possui características relevantes para gerar bons indicadores de qualidade de dados, garantindo assim que os dados de licitações públicas possam auxiliar na tomada de decisões.*

1. Introdução

A afirmação "Dados são o novo petróleo" (*"Data is the new oil"*) feita pelo matemático britânico Clive Humby[1] diz muito sobre a importância da criação e manutenção de dados com qualidade. Cada vez mais decisões são tomadas com base em dados, especialmente em uma realidade que grandes volumes de dados são disponibilizados constantemente na Web[2]. Assim, para que essas decisões sejam mais assertivas e precisas, é importante que os dados sejam confiáveis [Medeiros et al. 2020, Junior and Dorneles 2021].

É nesse contexto que se encontra a área de qualidade de dados. Apesar de haver várias definições para esta área, um consenso é que ela está sempre associada a um

[1]*Data is the new oil*: https://www.forbes.com/sites/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing/?sh=65cd33973045

[2]Um minuto na internet: https://www.statista.com/chart/25443/estimated-amount-of-data-created-on-the-internet-in-one-minute/

contexto específico [Junior and Dorneles 2021]. Ou seja, os dados podem ser adequados para um cenário, mas não para outro. Por isso, muitos trabalhos analisam qualidade em um domínio específico [Cichy and Rass 2019]. Outra definição está relacionada a considerar múltiplas dimensões, identificadas por atributos, que representam características específicas dos dados [Scannapieco and Catarci 2002, Medeiros et al. 2020].

Portanto, este trabalho objetiva identificar a qualidade de dados em um armazém de dados provenientes de licitações públicas. A principal motivação é a identificação de incongruências que possam impactar nas análises e auditorias feitas sobre essas licitações. Dessa forma, são utilizados diversos indicadores de qualidade, *i.e.*, implementações de métricas que avaliam regras específicas nos dados. Por exemplo, em relação a uma coluna que armazena valores percentuais, um indicador pode avaliar se todos os registros dessa coluna estão no intervalo entre 0% e 100%.

Neste trabalho, são consideradas oito ferramentas *open-source* que consideram diferentes dimensões de qualidade de dados. Após a comparação de suas funcionalidades, selecionamos a ferramenta *Great Expectations* (GE) como a mais adequada ao contexto analisado por verificar problemas de qualidade e reportá-los aos usuários responsáveis de maneira automatizada. Tal ferramenta possui diversos indicadores implementados nativamente, além de permitir o desenvolvimento de indicadores personalizados, pelos quais é possível implementar regras de negócio específicas do contexto dos dados analisados. A GE também possui um componente para geração de uma interface gráfica interativa com os resultados dos indicadores. Após a seleção da ferramenta, é proposta uma metodologia para a tarefa de análise da qualidade de dados utilizando a GE.

O restante deste artigo está organizado da seguinte maneira. Os trabalhos relacionados são apresentados na Seção 2. Em seguida, a Seção 3 descreve a análise comparativa das ferramentas de qualidade de dados. A Seção 4 apresenta os passos da metodologia para análise da qualidade de dados utilizando a ferramenta GE. A Seção 5 apresenta os resultados obtidos na análise da qualidade de dados de licitações públicas. Finalmente, na Seção 6, são apresentadas as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

O termo "qualidade de dados" está relacionado a um conjunto de características que os dados devem possuir. Essas propriedades são denominadas de dimensões, que incluem, por exemplo, precisão, completude e consistência [Scannapieco and Catarci 2002, Medeiros et al. 2020]. O processo de gerenciamento dessa qualidade compreende quatro práticas, sendo elas: (*i*) a geração de perfis para elaboração de um panorama dos dados e identificação de como estão armazenados [Cichy and Rass 2019]; (*ii*) a medição da qualidade que consiste, por exemplo, na identificação de dados faltantes, *outliers* e informações corrompidas [Lee et al. 2002, Ehrlinger and Wöß 2018]; (*iii*) a limpeza que faz a remoção de dados não desejáveis [Elmagarmid et al. 2007]; e (*iv*) o monitoramento que objetiva manter a doutrina de qualidade de dados em uma equipe, e a criação/uso de ferramentas e processos a serem aplicados nas etapas anteriores [Pipino et al. 2002, Laranjeiro et al. 2015].

A qualidade de dados deve ser definida no seu contexto de uso, pois o mesmo conjunto de dados pode precisar de diferentes tipos de indicadores, dependendo das necessidades dos seus usuários [Ballou and Pazer 1985]. A importância da qualidade de

dados foi observada em vários contextos distintos, incluindo cartografia [Chrisman 1983] e medicina [Goudar et al. 2015, Zöllner et al. 2016]. Além disso, dada a necessidade de treinamento de modelos de inteligência artificial, Sessions e Valtorta [2006] apresentam uma análise dos efeitos da qualidade de dados em algoritmos de aprendizado de máquina, demonstrando a importância da aplicação destes conceitos.

Dessa forma, para analisar a qualidade em grandes volumes de dados em diferentes contextos, são necessários métodos automatizados, resultando em um amplo mercado de ferramentas com essa finalidade. Nesse sentido, existem trabalhos na literatura cujo objetivo é comparar ferramentas de qualidade de dados. Por exemplo, Pushkarev et al. [2010] avaliam sete ferramentas *open-source* ou com períodos gratuitos de teste utilizando critérios como conectividade, gerenciamento, interface e funcionalidades. Gao et al. [2016] analisam oito ferramentas comerciais em relação às suas funcionalidades. Além disso, Altendeitering e Tomczyk [2022] propõem uma taxonomia para qualidade e analisam 18 ferramentas nesse contexto. Por fim, um estudo mais extensivo é realizado por Ehrlinger e Wöß [2022], que analisam 667 ferramentas distintas de qualidade. Os autores utilizam um conjunto de critérios de exclusão para selecionar 13 ferramentas (oito comerciais e cinco *open-source*) para uma comparação mais detalhada.

Apesar de a qualidade de dados ser um tópico de pesquisa bastante estudado em diferentes contextos, não foram encontrados trabalhos que fazem esse tipo de análise em dados de licitações públicas. Dessa forma, neste trabalho, são analisadas oito ferramentas *open-source* que podem ser aplicadas nesse contexto. A qualidade dos dados obtidos de licitações é fundamental para aplicações finalísticas, como a detecção de fraudes nesses processos e outras tarefas de predição e recomendação.

3. Análise de Ferramentas de Qualidade de Dados

Esta seção apresenta uma análise comparativa de ferramentas de qualidade de dados, selecionadas a partir de critérios pré-definidos. Na Seção 3.1, os critérios de seleção das ferramentas são descritos, bem como as ferramentas selecionadas. Em seguida, na Seção 3.2, os resultados da comparação das funcionalidades de cada ferramenta são apresentados. Por fim, na Seção 3.3, a ferramenta *Great Expectations* é descrita em detalhes, sendo a ferramenta que melhor atende os critérios de seleção definidos.

3.1. Ferramentas de Qualidade de Dados Selecionadas

Para a realização da análise comparativa, dois estudos foram usados como base para a seleção de ferramentas de qualidade de dados. O primeiro trabalho, proposto por Ehrlinger e Wöß [2022], contém uma revisão sistemática de 667 ferramentas, onde são aplicados critérios de exclusão para a seleção de 13 ferramentas consideradas em análises comparativas. Tais critérios verificam, principalmente, se as ferramentas são dedicadas a tarefas e domínios específicos, e se elas estão disponíveis publicamente de forma gratuita ou se são privadas, mas com um período de testes gratuito. O segundo trabalho, proposto por Foidl et al. [2022], explora três ferramentas adicionais, que não estão presentes no escopo do primeiro trabalho.

Com um conjunto de 16 ferramentas pré-selecionadas a partir dos critérios de exclusão dos dois trabalhos considerados [Ehrlinger and Wöß 2022, Foidl et al. 2022], também foi incluído um critério adicional, que avalia se uma ferramenta é *open-source* e

visa garantir a possibilidade de utilizar a ferramenta facilmente e reproduzir a metodologia aqui proposta. Após a inclusão do critério adicional, o processo de seleção resultou em um conjunto final de oito ferramentas de qualidade de dados. A seguir, cada uma delas é brevemente descrita com a referência do artigo fonte onde a ferramenta foi descoberta. Nas próximas seções, são apresentados os resultados das análises comparativas entre as ferramentas selecionadas.

Aggregate Profiler (AP)[3]**.** Ferramenta dedicada à preparação e qualidade de dados. A ferramenta AP fornece uma plataforma integrada de gerenciamento de dados que, além de recursos voltados para a preparação de dados, também fornece limpeza de dados, análises estatísticas, correspondência de padrões e a criação de *profiling* de dados [Ehrlinger and Wöß 2022].

Apache Griffin (AG)[4]**.** Ferramenta voltada para *big data*, dedicada a medir continuamente a qualidade de dados em lote ou em streaming. A ferramenta AG oferece um conjunto de modelos de domínio de qualidade de dados bem definidos, que cobre diferentes problemas de qualidade de dados em geral [Ehrlinger and Wöß 2022].

Great Expectations (GE)[5]**.** Biblioteca em código aberto para validar, documentar e caracterizar dados. Seu funcionamento se baseia no conceito de testes automatizados da Engenharia de Software, sendo possível atestar a qualidade de dados a partir do que se espera [Foidl et al. 2022].

MobyDQ[6]**.** Ferramenta para automatizar verificações de qualidade de dados durante o processamento de dados, capturando medições e resultados de métricas, e acionando alertas em caso de anomalias. A ferramenta foi inspirada em um projeto interno da Ubisoft Entertainment para medir e melhorar a qualidade dos dados de sua Enterprise Data Platform. No entanto, a versão *open-source* foi reformulada para melhorar seu design e remover dependências técnicas com software comercial [Ehrlinger and Wöß 2022].

OpenRefine & Metric (ORM)[7]**.** Ferramenta dedicada à limpeza e transformação de dados, operando em dados estruturados (linhas e colunas), semelhante à maneira como as tabelas de banco de dados relacional operam. Especificamente, projetos ORM consistem em uma tabela, cujas linhas podem ser filtradas usando critérios definidos [Ehrlinger and Wöß 2022].

PyDeequ[8]**.** API em Python para a Amazon Deequ, uma biblioteca cuja finalidade é executar "testes de unidade" para dados, ou seja, medir a qualidade dos dados de acordo com regras e condições pré-estabelecidas [Foidl et al. 2022].

Talend Open Studio (TOS)[9]**.** A empresa Talend oferece dois produtos para qualidade de dados: Talend Data Management Platform e Talend Open Studio (TOS). O primeiro

[3]Aggregate Profiler: https://sourceforge.net/projects/dataquality/
[4]Apache Griffin: https://griffin.apache.org/
[5]Great Expectations: https://greatexpectations.io/
[6]MobyDQ: https://ubisoft.github.io/mobydq/
[7]OpenRefine & Metric: https://openrefine.org/
[8]PyDeequ: https://github.com/awslabs/python-deequ
[9]Talend Open Studio: https://www.talend.com/products/talend-open-studio/

é a versão paga, que requer assinatura, enquanto o segundo é uma ferramenta gratuita *open-source*. Ambos os produtos (Open Studio e Enterprise) oferecem um bom suporte para análise de Big Data, como Spark ou Hadoop, e uma variedade de funcionalidades de *profiling* e limpeza de dados [Ehrlinger and Wöß 2022].

Tensorflow Data Validation (TFDV)[10]**.** Biblioteca para explorar e validar dados de aprendizado de máquina. A ferramenta TFDV foi projetada para ser altamente escalável e funcionar bem com o TensorFlow e o TensorFlow Extended (TFX) [Foidl et al. 2022].

3.2. Comparação de Funcionalidades

Nesta seção, as oito ferramentas *open-source* são comparadas em relação às suas respectivas funcionalidades disponíveis. Para guiar a análise comparativa, seguindo uma metodologia similar a [Ehrlinger and Wöß 2022], um catálogo de requisitos para avaliação foi definido e listado na primeira coluna da Tabela 1. O objetivo é classificar o atendimento de cada requisito a partir de quatro categorias possíveis: (✓) atendido, (✗) não atendido, (p) parcialmente atendido, (c) customização de indicadores disponível. Em particular, a categoria c indica a possibilidade de implementar indicadores customizados, dessa forma é possível customizar qualquer requisito indisponível na ferramenta.

Para realizar a análise comparativa e definir o catálogo de requisitos, avaliamos apenas as documentações de cada ferramenta. Portanto, qualquer funcionalidade que não tenha sido citada nas documentações, não foi incluída na comparação. O catálogo final de requisitos inclui funcionalidades relacionadas às seguintes categorias (**#1**) formatação de tabelas, como tamanho e existência de linhas/colunas; (**#2**) restrições sobre valores; (**#3**) intervalo de valores; (**#4**) casamento de padrões em *strings*; (**#5**) dados em formato de data ou JSON; (**#6**) funções de agregação de dados; (**#7**) operações multi-coluna; (**#8**) funções relacionadas a distribuições de probabilidade; e (**#9**) funções relacionadas a arquivos. Além das nove categorias, também incluímos uma adicional (**#10**) relacionada à possibilidade de customização de indicadores[11].

A Tabela 1 revela que as funcionalidades mais básicas (1–4) são cobertas pela maioria das ferramentas, seja de forma completa ou parcial. Já a maioria das funcionalidades mais sofisticadas (5–9), como funções relacionadas a distribuições de probabilidade e arquivos, são mais incomuns. Como exceção, as funções de agregação, apesar de também serem funcionalidades mais complexas, são parcialmente cobertas pela maioria das ferramentas. Por fim, em relação à funcionalidade **#10**, referente à disponibilização de indicadores customizados, observa-se que 50% das ferramentas apresentam tal requisito. Assim, mesmo que tais ferramentas não apresentem certos indicadores específicos de forma nativa, é possível implementá-los de forma customizada.

No geral, as ferramentas que menos atendem aos requisitos listados são a Apache Griffin, OpenRefine & Metric, PyDeequ e Tensorflow Data Validation. Além de apresentarem poucas funcionalidades, em comparação com as demais ferramentas, elas também não fornecem indicadores customizados. De forma contrária, as que melhor atendem as dez funcionalidades são as ferramentas Great Expectations, MobyDQ, Aggregate Profiler e Talend Open Studio. Todas as quatro ferramentas fornecem a aplicação de regras

[10]Tensorflow Data Validation: https://github.com/tensorflow/data-validation

[11]O detalhamento de cada categoria de funcionalidades está contido no Material Suplementar disponível em https://doi.org/10.5281/zenodo.7007428.

Tabela 1. Comparativo de funcionalidades entre as ferramentas.

#	Funcionalidades	Aggregate Profiler	Apache Griffin	Great Expectations	MobyDQ	OpenRefine & Metric	PyDeequ	Talend Open Studio	Tensorflow Data Validation
1	Formato da tabela	c	p	✓	c	p	p	p	✓
2	Restrições sobre valores	c	p	c	c	p	p	c	p
3	Intervalo de valores	p	p	c	c	p	p	c	p
4	Casamento de strings	c	p	p	✗	p	p	p	p
5	Timestamp e JSON	✗	✗	✓	p	p	✗	p	✗
6	Funções de agregação	p	p	p	p	✗	p	p	p
7	Operações multi-coluna	p	✗	p	p	✗	p	✗	✗
8	Funções relacionadas a distribuições de probabilidade	✗	✗	p	p	✗	✗	✗	p
9	Funções relacionadas a arquivos	p	✗	✓	c	p	✗	✗	✗
10	Indicadores customizados	✓	✗	✓	✓	✗	✗	✓	✗

Tabela 2. Ranqueamento das melhores ferramentas.

#	Ferramentas	✓	c	p	✗	Componentes Adicionais
1	Great Expectation	40%	20%	40%	0%	Profiler, Interface gráfica
2	MobyDQ	10%	40%	40%	10%	Interface gráfica
3	Aggregate Profiler	10%	30%	40%	20%	Interface gráfica
4	Talend Open Studio	10%	20%	40%	30%	–

de negócio, visto que elas disponibilizam customização de indicadores. Em particular, tal funcionalidade é de suma importância para o domínio analisado neste estudo (*i.e.*, licitações públicas), dado que tal contexto requer regras de negócio específicas.

A partir do conjunto final das quatro melhores ferramentas identificadas, foi realizado um ranqueamento delas, em relação ao melhor atendimento dos requisitos e o fornecimento de componentes adicionais. A Tabela 2 apresenta esse ranqueamento, com a porcentagem de cada categoria de atendimento e uma lista dos respectivos componentes adicionais, se houverem. Observa-se que a ferramenta que melhor se enquadrou em toda a análise comparativa foi a *Great Expectations*, que além de ter superado as demais tarefas em relação às funcionalidades, fornece componentes adicionais relevantes, incluindo o *Profiler* e uma interface gráfica, chamada de *Data Docs*, com os resultados dos indicadores executados. Na seção a seguir, tais componentes e funcionalidades principais da ferramenta são descritos em detalhes.

3.3. A Ferramenta *Great Expectations*

A *Great Expectations* (GE) é uma ferramenta *open-source* de qualidade de dados que utiliza um mecanismo similar aos testes de unidade para validação dos dados, onde cada validação é feita por um módulo chamado *expectation*. Aqui, as *expectations* são chama-

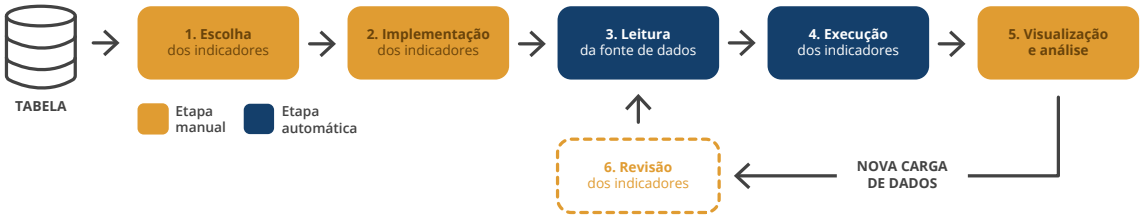

Figura 1. Metodologia para análise da qualidade de dados.

das indicadores. A GE fornece vários indicadores nativos que fazem validações genéricas de dados, como, por exemplo, as checagens de: tipos de campos, faixas de valores, registros nulos, entre outras. Além disso, a GE oferece a possibilidade da criação de indicadores personalizados, que permitem implementar regras de negócio especificas para cada tabela. Tais indicadores são codificados em linguagem Python e integrados à estrutura da ferramenta, podendo ser utilizados em conjunto com os indicadores nativos. A seguir, os principais componentes e funcionalidades disponíveis na GE são brevemente descritos.

Expectations. Correspondem a um conjunto de asserções expressas em linguagem declarativa e utilizadas para validação de dados. A GE verifica tais asserções nas colunas da tabela desejadas e retorna como resultado o sucesso ou a falha da verificação. As *expectations* são os indicadores para avaliar a qualidade dos dados, e sua implementação pode ser feita sobre *dataframes* Pandas, Spark e SQLAlchemy. Por fim, os resultados podem ser retornados em formato estruturado (JSON), permitindo a realização de qualquer tipo de pós-processamento.

Profiler. Este componente realiza uma pré-analise e retorna uma caracterização dos dados, assim como uma coleção de indicadores que mais se adéquam aos dados analisados. Esses indicadores retornados podem ser vistos como uma recomendação de quais as melhores validações devem ser feitas sobre esses dados.

Data Docs. Este componente é responsável por exibir os resultados dos indicadores executados sobre os dados. Ele fornece uma interface gráfica interativa, no formato de página HTML, onde o usuário pode navegar pelos resultados.

Em resumo, para o contexto dos dados utilizados neste trabalho, a ferramenta GE foi escolhida pelo fato de: (*i*) os indicadores customizados permitirem a implementação de indicadores de qualidade de dados voltados a problemas específicos do negócio; (*ii*) o *Data Docs* gera uma interface gráfica contendo os resultados dos indicadores executados, facilitando a análise dos resultados por parte do usuário; e (*iii*) o *Profiler* faz uma pré-análise da estrutura dos dados armazenados, e em seguida mostra um panorama dos dados com alguns indicadores interessantes de serem implementados sobre eles.

4. Metodologia para Análise da Qualidade de Dados com a GE

Após a escolha da ferramenta de qualidade de dados, foi desenvolvida uma metodologia para a tarefa de análise da qualidade de dados utilizando a *Great Expectations* (GE), ilustrada na Figura 1. Esse fluxo consiste em cinco passos principais e um opcional, cobrindo desde a escolha dos indicadores específicos para cada tabela até a visualização e análise dos resultados por especialistas. A seguir, são detalhadas as etapas do fluxo.

Escolha de indicadores para análise. Esta é a primeira etapa após a seleção da tabela e

consiste em uma inspeção manual de sua estrutura e de seu conteúdo para a definição dos indicadores de qualidade que serão implementados. É importante que a pessoa condutora desta etapa tenha conhecimento técnico e do negócio para que os indicadores escolhidos sejam adequados. Nesta etapa é interessante a utilização do componente *Profiler* da GE, pois quando o componente realiza a pré-análise dos dados, ele verifica quais indicadores nativos são mais indicados para serem executados sobre os dados analisados.

Implementação dos indicadores. Esta etapa se refere à implementação em código dos indicadores escolhidos utilizando a ferramenta *Great Expectations*.

Leitura da fonte de dados. Neste passo, é feita a leitura de cada tabela do banco de dados a ser avaliada. Vale destacar que é necessária a leitura de todo o conteúdo da tabela para que os indicadores sejam executados.

Execução de indicadores de qualidade. Após a leitura da tabela, são executados os indicadores implementados e os resultados são apresentados em uma interface gráfica interativa, gerada pelo componente *Data Docs*.

Visualização e análise dos indicadores. A última etapa da metodologia corresponde à visualização e análise dos resultados dos indicadores na interface gráfica interativa. A partir desta análise, é possível verificar casos que indicam erros na carga e/ou no formato dos dados, bem como tomar as ações necessárias para a correção.

Revisão dos indicadores (opcional). Caso haja necessidade, esta etapa pode ser executada logo após as novas cargas de dados nas tabelas avaliadas, e compreende o processo de reavaliação e de implementação de novos indicadores de acordo com necessidades e demandas que possam surgir.

As próximas seções apresentam a aplicação da metodologia proposta para análise da qualidade de dados utilizando a GE em dados reais de licitações públicas. Os indicadores utilizados também são descritos nas próximas seções.

5. Aplicação em Dados Reais de Licitações Públicas

Esta seção apresenta a aplicação de uma ferramenta de qualidade de dados em um ambiente *big data* com dados reais de licitações públicas. Conforme discutido na Seção 3.2, foi escolhida a ferramenta *Great Expectations* (GE) por ser mais adequada ao contexto considerado. Além disso, foi utilizado o fluxo de qualidade de dados proposto na Seção 4 para escolher e gerar os indicadores de qualidade mais adequados aos dados. Assim, esta seção está organizada da seguinte maneira: primeiro, é apresentada uma descrição dos dados de licitações públicas sobre os quais serão gerados os indicadores de qualidade (Seção 5.1); em seguida, são discutidos os principais resultados gerados (Seção 5.2).

5.1. Descrição dos Dados e Escolha de Indicadores

Para realização deste trabalho, são considerados dados de licitações ocorridas no estado de Minas Gerais nos âmbitos estadual e municipal. As licitações municipais provêm do portal do Sistema Informatizado de Contas dos Municípios (SICOM) do Tribunal de Contas do Estado de Minas Gerais[12] e as estaduais são do Portal da Transparência do

[12] https://portalsicom1.tce.mg.gov.br/

Tabela 3. Indicadores da Great Expectations utilizados na análise.

	Indicador (*expectation*)	Descrição
Nativo	expect_column_values_to_not_be_null	Valores da coluna devem ser não-nulos
	expect_column_values_to_be_unique	Não devem haver valores duplicados na coluna
	expect_column_min_to_be_between	O menor valor da coluna deve estar dentro do intervalo [min, max]
	expect_column_values_to_be_in_type_list	Valores da coluna devem ser do tipo especificado
	expect_column_values_to_be_in_set	Valores da coluna devem pertencer a um conjunto de valores
	expect_column_values_to_be_between	Valores da coluna devem estar no intervalo [min, max]
	expect_column_values_to_match_regex	Valores da coluna devem seguir uma determinada expressão regular
Customizado	expect_value_less_revenue	Licitações devem ter valor menor ou igual à receita do ente
	expect_table_fato_licitacao_to_have_guests_if_invite	Licitações na modalidade convite devem ter licitantes convidados
	expect_dates_to_match_across_tables	Datas reportadas devem estar em ordem cronológica válida
	expect_only_one_year_of_activity	Licitações devem possuir apenas um único ano de atividade
	expect_sum_of_item_values_to_match_fato_licitacao	Soma dos valores dos itens deve coincidir com o valor da licitação

Governo de Minas Gerais[13]. Assim, o conjunto de dados final contém informações sobre 378.137 licitações que compreendem 12.522.661 itens licitados e 103.858 licitantes (pessoas físicas ou jurídicas) durante os anos de 2014 a 2021. Os dados estão armazenados em um ambiente *big data* utilizando o armazém de dados Apache Hive[14] versão 2.0.0. Vale destacar que nessa versão do Hive não há suporte para verificação de restrições de integridade nos dados. No entanto, utilizou-se essa versão para mostrar que a GE, ao detectar os registros problemáticos, consegue mitigar a falta de tais restrições.

Em relação aos indicadores de qualidade da ferramenta *Great Expectations* para essa fonte de dados, foram utilizados indicadores nativos e customizados (conforme Seção 3.3). Os primeiros são regras padrões e genéricas, implementadas internamente na ferramenta, como, por exemplo, validar domínio dos dados, verificar se os dados seguem uma expressão regular ou um intervalo de valores. A Tabela 3 apresenta a lista dos indicadores nativos utilizados na análise dos dados de licitações públicas realizada neste trabalho.[15]

Já os indicadores customizados servem para validar alguma regra de negócio específica do domínio dos dados. Por exemplo, ao analisar manualmente os valores das licitações, foram observados números muito discrepantes: uma única licitação possuía o valor quase 200 vezes maior do que toda a arrecadação de seu município naquele ano. Possíveis causas dessa anomalia são um erro de digitação por quem inseriu esses dados na fonte ou uma falha na extração desses valores a partir dos documentos dos processos licitatórios. Assim, foi implementado um indicador customizado que compara o valor da licitação com o valor total arrecadado pelo município ou estado no ano da licitação. Ao todo, foram implementados cinco indicadores customizados, também descritos na Tabela 3. Tais indicadores foram implementados de forma codificada seguindo os padrões de nomenclatura e organização dos indicadores nativos.

5.2. Análise da Qualidade de Dados

Nesta seção, são apresentados os principais resultados dos indicadores de qualidade da ferramenta *Great Expectations* (GE) em dados reais de licitações públicas. Nesta análise, foram consideradas as cinco principais tabelas que reúnem os dados de licitações: (*i*) informações gerais das licitações; (*ii*) licitantes habilitados a participar dos processos lici-

[13] https://www.transparencia.mg.gov.br/compras-e-patrimonio/compras-e-contratos

[14]Hive: https://hive.apache.org/

[15]Lista de indicadores nativos existentes na GE https://greatexpectations.io/expectations

Tabela 4. Estatísticas gerais da execução dos indicadores na GE.

Tabela	Sucessos	Falhas	Total
Licitação	88 (68,22%)	41 (31,78%)	129
Licitantes habilitados	25 (71,43%)	10 (28,57%)	35
Licitantes homologados	32 (43,24%)	42 (56,76%)	74
Itens licitados	54 (65,06%)	29 (34,94%)	83
Comissões	47 (83,93%)	9 (16,07%)	56

Tabela 5. Quantidade de falhas capturadas pelos indicadores de qualidade.

Erro / Tabela	Licitação	Lic. Habilitado	Lic. Homologado	Item Licitado	Comissão
Valores nulos	18	1	15	6	0
Valores fora do esperado	12	3	10	5	1
Tipo de dados incoerente	8	2	3	8	7
Valores duplicados	0	1	3	1	0
Outros	3	3	11	9	1
Total	**41**	**10**	**42**	**29**	**9**

tatórios; (*iii*) licitantes homologados como vencedores em licitações; (*iv*) itens licitados; e (*v*) comissões de licitação (órgãos instituídos para atuar em licitações).

Para cada tabela, foram escolhidos indicadores nativos específicos, ou seja, que fizessem sentido no contexto de cada uma. Além disso, foram implementados indicadores customizados conforme as regras de negócio pré-definidas. A Tabela 4 apresenta o número de sucessos e falhas nos indicadores para cada tabela analisada, bem como o total de indicadores implementados.

A Tabela 5 apresenta os erros mais comuns detectados nas tabelas analisadas. Um dos erros mais frequentes é a presença de valores nulos em colunas onde não são permitidos segundo as regras de negócio. Por exemplo, não é esperado que campos contendo o ano de exercício de licitações apresentem valores nulos. Outros erros comuns incluem a presença de valores fora do padrão e/ou intervalo esperado e tipo incoerente de dados. Além disso, algumas tabelas possuem registros duplicados, erro que pode ocorrer por duas razões: (*i*) problemas na carga de dados e (*ii*) o armazém de dados utilizado não suporta restrições de integridade para evitar essa duplicidade. No entanto, como a GE detecta esses registros duplicados, é possível mitigar tal limitação do armazém de dados.

Além disso, os indicadores customizados da GE permitem verificar regras de negócio mais complexas e que não podem ser verificadas por indicadores nativos. A Tabela 6 apresenta uma parte do resultado do indicador *expect_values_less_revenue* na tabela com informações de licitações. É possível verificar que três licitações possuem valores

Tabela 6. Indicador customizado que verifica licitações cujo valor é maior que o valor arrecadado pelo município no ano exercício.

Licitação	Ano exercício	Nome da entidade	Valor da licitação	Valor arrecadado
A	2014	Município X	R$ 59.415.748.800,00	R$ 11.912.844,54
B	2015	Município Y	R$ 16.880.000,00	R$ 13.124.280,52
C	2020	Município Z	R$ 262.029.682,50	R$ 240.799.958,79

Tabela 7. Indicador customizado que verifica quantos registros desrespeitam a ordem cronológica de datas das licitações (Data 1 $\leq$ Data 2).

Data 1	Data 2	Qtd. Registros	%
Data do edital	Data publicação do edital	7.672	2,03
Data do edital	Data de publicação no veículo	8.728	2,31
Data publicação do edital	Data prevista recebimento documentação	2.186	0,58

discrepantes quando comparados à arrecadação total do município naquele ano (também obtida a partir do banco de dados). Por exemplo, a licitação A possui um valor mais de 4.000 vezes superior à arrecadação do município. No entanto, este valor não é o que está presente na pesquisa de preços presente no edital da licitação, indicando um provável erro no processo de extração e/ou carga de dados.

Outra regra de negócio verificada por um indicador customizado verifica a ordem cronológica de campos de data presentes nos registros de licitação, pois as datas devem respeitar a ordem do processo licitatório. Por exemplo, a data do edital da licitação deve ser anterior à sua publicação, pois a elaboração do edital é a primeira etapa do processo, e o recebimento da documentação só acontece apos o edital ser publicado. A Tabela 7 apresenta a quantidade de casos que não respeitam essa ordem na tabela de licitações. Analisando a quantidade de registros nessa situação, é possível que tenha havido algum problema na imputação ou carga dos dados. Tal resultado reforça a necessidade de uma análise minuciosa dos processos de extração, tratamento e carga desses dados.

6. Conclusão

Este artigo apresentou uma análise comparativa de oito ferramentas *open-source* de avaliação de qualidade de dados, bem como resultados de uma análise em um ambiente *big data* com dados reais de licitações públicas utilizando a ferramenta *Great Expectations* (GE). Tal ferramenta foi escolhida por possuir componentes de interface gráfica que auxiliem na visualização dos resultados e por possibilitar a criação de indicadores customizados que permitem verificar regras de negócio complexas, não verificadas por indicadores nativos. Em outras palavras, tais indicadores permitem a implementação de validações específicas para os dados que estão sendo analisados. Vale destacar que a GE auxilia na identificação de registros que possuem problemas causados pela impossibilidade de implementar restrições de integridade nos dados pelo armazém de dados considerado. Finalmente, os resultados revelam que a escolha da melhor ferramenta depende de sua dinâmica de utilização, especificamente, dos dados e do contexto de uso. Como trabalhos futuros, planeja-se a análise da qualidade de outros domínios de dados, que podem exigir a aplicação de ferramentas diferentes da GE.

Agradecimentos. Ao Ministério Público de Minas Gerais (MPMG) pelo apoio através do Projeto Capacidades Analíticas. Ao CNPq, CAPES e FAPEMIG pelo apoio aos pesquisadores envolvidos.

Referências

Altendeitering, M. and Tomczyk, M. (2022). A functional taxonomy of data quality tools: Insights from science and practice. In *Wirtschaftsinformatik*.

Ballou, D. P. and Pazer, H. L. (1985). Modeling data and process quality in multi-input, multi-output information systems. *Management Science*, 31(2):150–162.

Chrisman, N. R. (1983). The role of quality information in the long-term functioning of a geographic information system. In *Auto-Carto*, pages 303–312.

Cichy, C. and Rass, S. (2019). An overview of data quality frameworks. *IEEE Access*, 7:24634–24648.

Ehrlinger, L. and Wöß, W. (2018). A novel data quality metric for minimality. *QUAT*, 1:1 – 15.

Ehrlinger, L. and Wöß, W. (2022). A survey of data quality measurement and monitoring tools. *Front. Big Data*, 5.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Trans. Knowl. Data Eng.*, 19(1):1–16.

Foidl, H., Felderer, M., and Ramler, R. (2022). Data smells: Categories, causes and consequences, and detection of suspicious data in ai-based systems. In *arXiv*.

Gao, J. Z., Xie, C., and Tao, C. (2016). Big data validation and quality assurance - issuses, challenges, and needs. In *SOSE*, pages 433–441. IEEE Computer Society.

Goudar, S. et al. (2015). Data quality monitoring and performance metrics of a prospective, population-based observational study of maternal and newborn health in low resource settings. *Reproductive Health*, 12(2):1–10.

Junior, C. S. and Dorneles, C. F. (2021). Avaliação de dimensões de qualidade de dados para o agronegócio. In *SBBD*, pages 283–288. SBC.

Laranjeiro, N., Soydemir, S. N., and Bernardino, J. (2015). A survey on data quality: Classifying poor data. *PRDC*, pages 179 – 188.

Lee, Y. W. et al. (2002). Aimq: a methodology for information quality assessment. *Information & Management*, 40(2):133 – 146.

Medeiros, G. F. d., Degrossi, L. C., and Holanda, M. (2020). Qualiosm: Melhorando a qualidade dos dados na ferramenta de mapeamento colaborativo openstreetmap. In *SBBD*, pages 77–82. SBC.

Pipino, L. L. et al. (2002). Data quality assessment. *Commun. ACM*, 45(4):211 – 218.

Pushkarev, V. et al. (2010). An overview of open source data quality tools. In *IKE*, pages 370–376. CSREA Press.

Scannapieco, M. and Catarci, T. (2002). Data quality under a computer science perspective. *Journal of The ACM - JACM*, 2:1–12.

Sessions, V. and Valtorta, M. (2006). The effects of data quality on machine learning algorithms. In *ICIQ*, pages 485–498. MIT.

Zöllner, F. et al. (2016). An open source software for analysis of dynamic contrast enhanced magnetic resonance images: Ummperfusion revisited. *BMC Med Imaging*, 16(7):1–13.