Qualidade de dados dentro do contexto de *Big Data*: Uma revisão global

Magnon P. R. Souza[1], **Cleber Zanchettin**[1], **Divanilson R. S. Campelo**[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Recife – PE – Brazil

{mprs,cz,dcampelo}@cin.ufpe.br

Abstract. *The rapid expansion of Big Data has elevated the importance of data quality, posing unique challenges and opportunities. This work aims to undertake a review of data quality within the context of Big Data. It covers Big Data characteristics, its lifecycle and the intricacies of data quality. A literature review serves as methodology, offering an up-to-date structured review of the state-of-the-art in data quality for Big Data applications. The findings from this research aim to assist in the understanding of what is in the core of data quality for these systems, for better value extraction and therefore better decision-making processes.*

Resumo. *A rápida expansão do Big Data elevou a importância da qualidade dos dados, colocando desafios e oportunidades únicas. Este trabalho tem por objetivo efetuar uma análise da qualidade dos dados no contexto dos Big Data. O mesmo abrange as características do Big Data, o seu ciclo de vida e as complexidades da qualidade dos dados. Sua metodologia é composta por uma revisão de literatura, oferecendo uma visão atualizada e estruturada do estado da arte da qualidade dos dados para aplicações de Big Data. Os resultados desta investigação visam ajudar a compreender o que está no cerne da qualidade dos dados para estes sistemas, para uma melhor extração de valor e, por conseguinte, melhores processos de tomada de decisões.*

1. Introdução

Em um mundo constantemente bombardeado por novos dados, uma nova commoditie surge trazendo um impacto tão alto que chega a ser visto como mais valioso que petróleo [The Economist 2017]. Isso se mostra tanto uma realidade que 79% das empresas acreditam que vão falir se não utilizarem *Big Data* em seu modelo de negócios [Kothapalli 2023]. O contexto de *Big Data*, com sua vastidão em volume, velocidade e variedade, ressalta ainda mais a necessidade de garantir não apenas a quantidade, mas principalmente a qualidade desses dados.

Geralmente, os dados não estão prontos para serem processados assim que são coletados [Taleb et al. 2021]. Eles preciso passar por etapas de limpeza de dados (por exemplo, redução de dimensionalidade, tratamentos para dados ausentes, conversão de tipo de dado) e adequação ao contexto de negócio. Essa cadeia de processos é feita de forma que o valor agregado nessa imensidão de informação possa ser melhor aproveitado, mitigando o efeito *"garbage in, garbage out"*.

Ainda que este tópico seja de extrema relevância e venha sendo discutido há muitos anos dentro do escopo de bases de dados tradicionais, sua discussão no âmbito do *Big Data* ainda é precária. Isso acontece pois, os processos para esta realidade são mais complexos de implementar, visto que eles demandam mais tempo, poder computacional e, consequentemente, despesa. Entretanto, faz-se fundamental que esse debate seja expandido em todas esferas que tenham dados como uma fundação primária no âmbito operacional, tático e de tomada de decisão. Isso permitirá não apenas uma melhoria na entrega de valor, como também uma otimização na cadeia de processamento (melhoria de métricas de um modelo de *deep learning*, por exemplo).

Este trabalho visa fazer uma jornada analítica de revisão sobre a literatura de qualidade de dados dentro do âmbito de *Big Data* e trazer uma visão sobre o estado da arte e realçar os direcionamentos de pesquisa que tem se destacado nos últimos anos. O trabalho visa responder a perguntas como:

1. *Quais principais características dos sistemas de Big Data hoje em dia?*
2. *Que dimensões de qualidade de dados tem sido mais mencionadas atualmente?*
3. *Que tópicos tem ascendido na discussão sobre qualidade de dados em Big Data nos últimos anos?*

O resto desse trabalho é organizado da seguinte forma. A seção 2 trata sobre *Big Data* de forma geral. Em seguida, o tema qualidade de dados é discutido na seção 3. Seguido pela metodologia, na seção 4, e a discussão sobre os resultados da revisão na seção 5. Por fim, a seção 6 conclui o trabalho.

2. *Big Data*

Muito mais que uma *buzzword* das últimas décadas, *Big Data* é um fenômeno universal. Seja na esfera acadêmica ou nas mais diversas indústrias, o grande volume de dados gerado diariamente tem desempenhado um papel fundamental nos processos de análise e tomada de decisão. Com a dominação dos *smartphones* e das redes sociais, a ascensão do *IoT* e o uso disseminado de sensores nos mais diferentes dispositivos, fazem com que seja produzida uma quantidade extraordinária de dados em um curto intervalo de tempo. Foi estimado por [DOMO 2022] que a quantidade de dados produzida em 2022 foi de 97 *zettabytes* e projeta-se que esse número cresça para 181 *zettabytes* em 2025.

Dito isto, é uma falsa interpretação pensar que *Big Data* se resume apenas a um grande volume de dados. Na verdade, o significado subjacente ao termo "*Big Data*"vai muito além da magnitude do volume de dados, ele tem evoluído de maneira notável ao longo dos anos. Ao explorar suas múltiplas facetas, nota-se não apenas uma dimensão quantitativa, mas também elementos qualitativos que moldam sua natureza. Adentrando esse cenário, surge a necessidade de explorar em detalhes essas características peculiares da definição do que é *Big Data*, bem como entender como a qualidade dos dados desempenha um papel crucial nesse contexto.

2.1. Conceito de *Big Data*

Apesar da popularidade recente, o termo *Big Data* teve seu nascimento antes do que se imagina. O primeiro registro deste surgiu em torno de 1980 por Charles Telly e a sua definição como entendemos hoje em dia foi primeiramente elaborada por Tim O'Reilly em 2005 [Abdallah 2019].

Desde então, a definição de *Big Data* vem sendo revisitada e expandida por pesquisadores de forma a incluir as transformações que vão sendo agregadas ao contexto, desde características e atributos até ferramentas e arquiteturas. Como foi apontado por [Taleb et al. 2021], a IBM descreveu *Big Data* como caracterizado por altos volumes, alta velocidade e grande variedade de atributos que requerem formas inovadoras e economicamente viáveis de processamento de informações, a fim de aprimorar a geração de *insights* e auxiliar na tomada de decisão.

O termo *Big Data* envelopa os cenários em que estruturas tradicionais de bancos de dados sequenciais, e não sequenciais, não se aplicam. Isso se dá não apenas pelo volume de dados, como também pela natureza dos dados. Isso faz com que esses dados precisem passar por diferentes processos de tratamento, os quais estão intimamente ligados ao dado em si, às ferramentas disponíveis, ao poder computacional disponível e ao modelo de negócio em questão. Todas essas variáveis, dentre outras, por si só fazendo do *Big Data* algo extremamente dinâmico [Taleb et al. 2018a] e contextual, fazendo necessário um estudo mais aprofundado sobre suas características particulares, de forma que possamos entender os seus desafios inerentes na tarefa de extrair valor dos dados.

2.2. Características de *Big Data*

Uma das primeiras tentativas de construir uma caracterização dos componentes de *Big Data* foi elaborada em 2001 por Laney. Nesse momento ele elencou três atributos que ficaram conhecidos como os 3 V's de dados: volume, velocidade e variedade [Laney 2001].

Ao longo dos últimos 22 anos, novos termos foram sendo agregados a essa lista inciada por Laney. Posteriormente essa lista foi estendida para 4 V's: volume, velocidade, variedade e veracidade; Em seguida, foi proposto um modelo com 5 V's, adicionando "valor"; Um novo estudo, então, apresentou uma estrutura com 7 V's [Uddin et al. 2014]; Após isso houveram novos modelos propostos, em que uma levantava 10 V's e outro 14 V's; e, por conseguinte, mais uma extensão foi proposta com mais 3 novas características, totalizando 17 V's [Arockia Panimalar.S 2017].

Algumas das características mais descritas nos trabalhos recentes são: volume, velocidade, variedade, veracidade, valor, vitalidade, viscosidade, visualização e vulnerabilidade [Taleb et al. 2021]. A tabela 1 descreve essas características de *Big Data*.

Essa crescente lista se dá pelo perfil dinâmico e evolutivo do *Big Data*. Conforme a geração de dados vai aumentando, bem como a sua aplicabilidade em múltiplos âmbitos como negócios, academia, governo e saúde, por exemplo, passa-se a identificar novas características atreladas ao estado da arte do *Big Data* naquele dado momento do tempo.

Um exemplo desse processo dinâmico e evolutivo é a "vulnerabilidade". Ao passo que muitos dados são gerados e compartilhados via redes sociais, potencializou o problema de segurança quanto a vazamentos de dados pessoais, acendendo um alerta maior para o desenvolvedores de aplicações de *Big Data* nessa característica.

2.3. Ciclo de vida de *Big Data*

Os dados que entram num pipeline de *Big Data* passam por algumas etapas até o ponto final de geração de *insights*. Esse fluxo é conhecido como o ciclo de vida de *Big Data* ou cadeia de valor de *Big Data*. Ao abstrairmos detalhes de implementação e detalhes

Tabela 1. Características de *Big Data*

Volume	Refere-se ao tamanho dos dados sendo gerados, de todos tipos de fontes incluindo texto, áudio, vídeo, redes sociais, previsão do tempo etc.
Velocidade	A velocidade com que os dados são gerados é extremamente alta.
Variedade	Dados aparecem de formatos muito distintos: áudio, vídeo, imagem, texto, geolocalização etc.
Veracidade	O quanto este dado é confiável? refere-se aos possíveis viéses, pertubações e anormalidades nos dados.
Validade	Semelhante ao conceito de veracidade, se trata do quanto o dado está correto e acurado para o objetivo de uso final.
Vitalidade	Também conhecida como volatilidade, trata-se dos dados armazenados e por quanto tempo eles serão úteis.
Viscosidade	A diferença de tempo entre o momento que o evento aconteceu e quando ele está sendo utilizado.
Visualização	Processo de representação visual do valor abstrato que o dado carrega.
Vulnerabilidade	*Big Data* traz novas preocupações sobre segurança das informações produzidas.
Valor	Se trata do uso do dado para o propósito final da aplicação. É o objetivo do processamento de *Big Data*: extrair utilidade dos dados coletados.

de arquitetura (elementos particulares para cada contexto), é possível observar algumas etapas ou processos em comum. Ilustrados na figura 1, são eles:

- **Geração de dados:** Essa é a fase de entrada [Taleb et al. 2019] das fontes de dados, onde ocorre a criação dos dados. Fontes distintas podem criar dados de formas distintas. Eles podem vir de sensores de clima, de GPS, *posts* em redes sociais, índices do mercado financeiro, aparelhos de vigilância, etc.
- **Aquisição de dados:** Consiste essencialmente em três subprocessos: a coleta de dados, a transmissão de dados e o pré-processamento de dados [Hu et al. 2014]. Esses subprocessos podem ser descritos da seguinte forma:
  - **Coleta de dados:** Essa etapa refere-se à tecnologia aplicada para a extração dos dados brutos das fontes que produzem esses dados.
  - **Transmissão de dados:** Após a coleta, faz-se necessário uma ferramenta para transmitir esses dados de forma rápida e eficiente para os seus armazenamentos apropriados.
  - **Pré-processamento de dados:** Enfim, com os dados já coletados e armazenados, eles precisam passar por uma higienização. Devido às suas características, os dados de *Big Data* podem conter informações faltantes, duplicadas, erradas ou irrelevantes pro contexto da aplicação. Isso acaba por onerar não só o valor agregado do dado, como também o custo de armazenamento e operação. Portanto os dados precisam passar por uma série de operações que tratem esses possíveis problemas. As formas de fazer esse tratamento podem ser as mais diversas. Desde mudanças opera-

cionais, como aplicação de regras definidas por especialistas do domínio ou uso de algoritmos de aprendizado de máquina.

- **Armazenamento de dados:** Essa etapa diz respeito ao armazenamento de *dataset* de larga escala. Existem duas óticas para esta etapa: Infraestrutura de hardware, que deve ser distribuída e elástica para suportar fluxo intenso e instantâneo, e gestão de dados que devem ser geridos de forma a possuir uma boa interface de comunicação e consulta com sistemas externos.
- **Processamento e análise de dados:** Utiliza métodos de análise ou ferramentas para inspecionar, transformar e modelar os dados para extrair valor. Essa extração de valor culmina em *insight*. Um exemplo disso pode ser a mudança de uma regra de negócio, por causa do *output* de um modelo de *machine learning*. De forma semelhante ao pré-processamento de dados, aqui podem ser utilizados conhecimentos de especialistas no domínio, bem como métodos estatísticos e algoritmos de aprendizado de máquina, porém com o objetivo de extrair valor final dos dados e transformar o dado em *insight*.

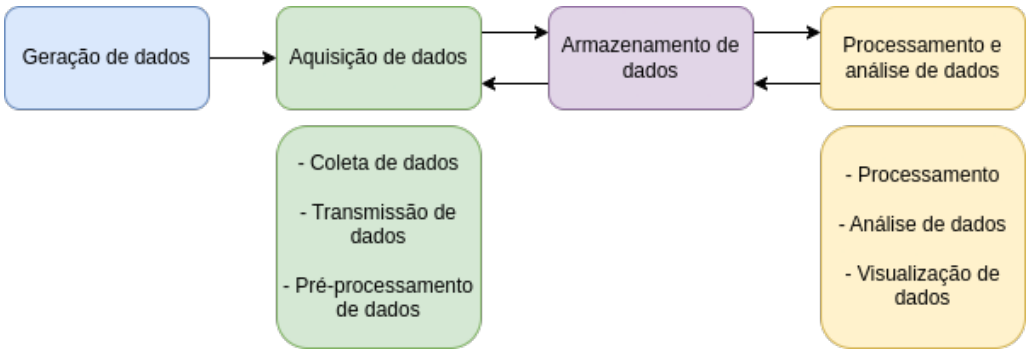

Figura 1. Ciclo de vida de *Big Data*

3. Qualidade de dados dentro do contexto de *Big Data*

Definir qualidade de dados não é uma tarefa simples [Oliveira et al. 2005]. Para construir esse conceito, precisa-se entender que ele é intrinsecamente dependente do contexto e varia significativamente com base no domínio, na aplicação e nos requisitos analíticos específicos. O que pode ser considerado dados de alta qualidade em um contexto pode não atender aos padrões ou necessidades de outro cenário. Por exemplo, dados em tempo real podem ser essenciais para aplicações como negociações financeiras ou sistemas de resposta a emergências. Nesse contexto, características como velocidade são mais cruciais para o sistema de *Big Data*. Por outro lado em pesquisas acadêmicas, dados históricos podem ser mais valiosos. Aqui, a questão do volume e variedade podem ser mais críticas.

3.1. Dimensões de qualidade de dados

Para se aprofundar no que entende-se por qualidade de dados, é preciso compreender melhor os elementos que a compõem. Esses elementos são chamados de **dimensões de qualidade de dados**. As dimensões de qualidade de dados (DQD) fornecem uma interface para mensurar, quantificar e gerir a qualidade dos dados.

Ao analisar a literatura, encontram-se diversas dimensões de qualidade de dados, mas, de acordo com [Wand and Wang 1996] e [Wang and Strong 1996], separam-as em

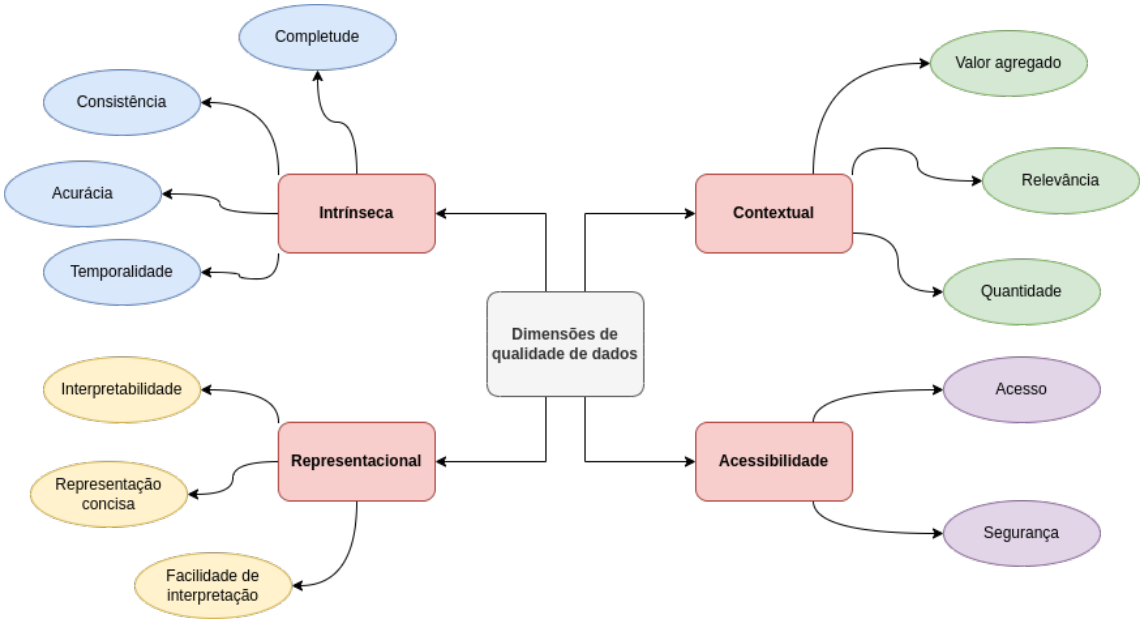

Figura 2. Representação de dimensões de qualidade de dados

intrínseca, contextual, representacional e acessibilidade. A imagem 2 traz uma ilustração de cada categoria.

- **Intrínseca:** Remete àquelas qualidades próprias à natureza dos dados.
- **Contextual:** Denota os requerimentos considerados dentro do contexto da aplicação.
- **Representacional:** Simboliza os atributos relacionados à representação dos próprios dados diante da tarefa em questão a ser trabalhada.
- **Acessibilidade:** Trata-se de questões atreladas à disponibilidade e ao acesso dos dados.

Dentro de uma aplicação, as dimensões de qualidade de dados são utilizadas para estabelecer métricas e mensurar a qualidade dos dados dentro do contexto da base de dados e do modelo de negócio. Em outras palavras, devem ser relevantes para o problema identificado. As métricas devem determinar se os valores de um dado atributo dos dados respeitam o parâmetro estabelecido para a respectiva dimensão de qualidade [Taleb et al. 2021]. Esses valores podem ser definidos através de estimativa estatística, uso de algoritmos de aprendizado de máquina ou de conhecimento de especialista. [Ehrlinger and Wöß 2022] traz uma discussão sobre algumas dimensões e apresenta fórmulas para calcular suas métricas.

3.2. Pipeline de qualidade de dados em *Big Data*

Tomando como ponto de partida, o ciclo de vida descrito na seção 2.3, é possível analisar cada etapa do ciclo de vida sob a ótica de qualidade de dados da seguinte maneira:

1. **Geração de dados:** Nessa etapa é preciso definir como e quais dados serão gerados. Neste momento, existe um *processo de decisão* sobre a estrutura dos dados que mais adiante serão coletados.

2. **Coleta de dados:** Esta etapa diz respeito a como, quando e como os dados serão coletados e tratados. Nesta etapa, faz-se necessário *construir uma definição* desses elementos bem como condições de verificação dos dados.
3. **Transmissão de dados:** A distribuição de dados nessa fase está atrelada a infraestrutura. Aqui a qualidade está condicionada às limitações do *hardware estabelecido pelos desenvolvedores* e da rede disponível e é expressa em termos de perda de dados e erros de transmissão.
4. **Pré-processamento de dados:** Aqui a tratativa é diretamente sobre os dados. Eles passam por processos de limpeza. São exemplos de técnicas utilizadas: padronização, cálculo de agregação, adequação de tipo de dado, tratamento de valores ausentes e normalização. Essas modificações são feitas afim de permitir que os dados possam entregar mais valor por meio de uma melhora dos resultados do processamento. Um exemplo disso pode ser descrito com uma melhora de resultados de um classificador e com esses resultados, possibilitar melhores resultados nos negócios.
5. **Armazenamento de dados:** Esta fase está relacionada, novamente, à infraestrutura do sistema. As questões de qualidade atreladas a esta etapa são de caráter de *saúde do hardware de armazenamento escolhido pelos desenvolvedores*. Soluções como replicação, redundância e backup de dados em múltiplas máquinas se fazem presentes nesse momento.
6. **Processamento de dados:** Neste momento, a qualidade é afetada *tanto pelos processos aplicados como também pela qualidade dos próprios dados*. Entre a aplicação de heurísticas e algoritmos de aprendizado de máquina, o estado que os dados são recebidos nesta etapa e o fluxo pelo qual eles passarão na presente etapa são determinantes da qualidade dos dados, impactando não apenas o resultado, como também o consumo de recursos do processamento.
7. **Análise e visualização de dados:** Nestas etapas finais, entra o trabalho do analista de dados de forma mais proeminente. Aqui ocorre a *manipulação dos dados finais, resultantes de toda cadeia, e aplicação de ferramentas de visualização*, como *dashboards* por exemplo, para a transformação dos dados em conhecimento e enfim a extração de valor.

Pode-se perceber que, tratar qualidade de dados no ciclo de vida de *Big Data* não é um trabalho pontual. Os problemas atrelados a questões de qualidade podem se manifestar em cada fase da cadeia. Eles vão surgir quando os requisitos de qualidade não forem atingidos pelos dados reais. Esses problemas ocorrem devido a diversos fatores a depender do nível que ocorrerem:

- **Fonte de dados:** incerteza, inconsistência, diferenças entre múltiplas fontes e particularidades do domínio.
- **Geração de dados:** *input* humano, leitura de sensores, redes sociais, dados não estruturados e dados faltantes.
- **Processos:** coleta e transmissão.

Dentro desse cenário, as estratégias de intervenção para abordar os problemas de qualidade podem ser categorizadas em duas estratégias de acordo com [Sidi et al. 2012]:

- **Relacionados ao processo:** O trabalho visa atacar o processo pelo qual o dado é gerado, de forma que o aumento da qualidade de dados seja promovido pela melhora na cadeia de produção do dado.

- **Relacionados ao dado:** Trata-se da modificação direta dos dados. Nessa estratégia, existe a aplicação de técnicas sobre os atributos ou sobre os dados propriamente.

A figura 3 ilustra onde as questões qualidade de dados pode ser aplicadas em cada etapa do ciclo de vida de *Big Data*. Nela percebe-se que a estratégia de tratamento de problemas de qualidade de dados relacionados a dados são aplicados essencialmente na fase de pré-processamento. São técnicas comuns dessa estratégia: normalização, *data cleasing* e filtragem de dados. Já a estratégia relacionada ao processo, pode ser aplicada em toda cadeia do ciclo de vida de *Big Data*.

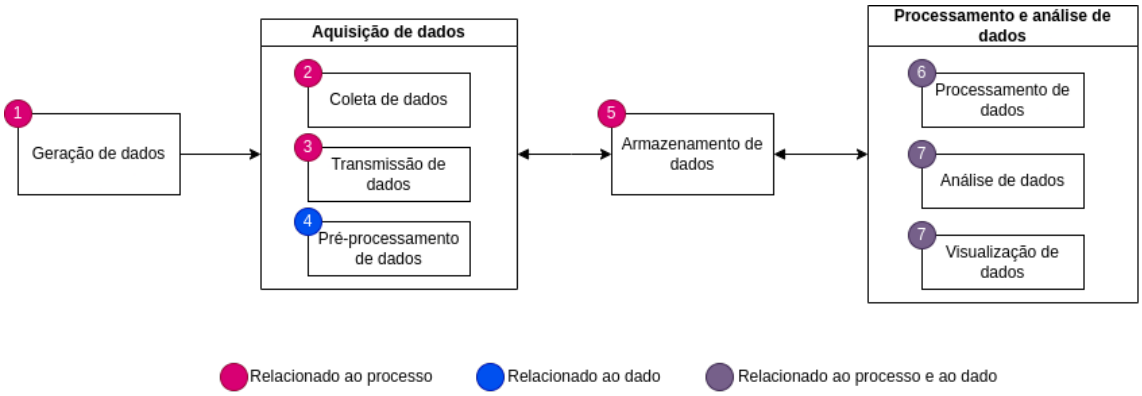

Figura 3. Problemas de qualidade dentro do ciclo de vida

4. Metodologia

Neste estudo, foi conduzida uma metodologia de revisão sistemática de literatura focada na análise e construção de *frameworks* de qualidade de dados voltados para sistemas de *Big Data*. A diretriz seguida foi inspirada nas orientações traçadas pelos estudos de Creswell & Creswell (2017) e por Kitchenham (2004). O fluxo da revisão está ilustrado na figura 4.

Para a pesquisa inicial, foi feita uma busca por termos mais genéricos, como ”*Data quality*”e ”*Big Data*”em fontes não necessariamente acadêmicas, como blogs técnicos por exemplo. Depois de uma primeira familiarização com o tema, foi feita uma nova busca, porém em fontes acadêmicas, onde pudesse encontrar resultados de artigos científicos e *papers*, a fim de um aprofundamento maior na teoria e definições formais. Para essa nova busca foram usados termos como *”data quality Big Data”*, *”Big Data quality”* e *”data quality framework for Big Data systems”*. Com os resultados dessa nova busca, foi feita uma seleção inicial a partir da leitura dos títulos, seguida de uma filtragem a partir da leitura dos resumos (*abstracts*) e por fim, a remoção de possíveis resultados duplicados, visto que um mesmo trabalho pode ser obtido de fontes distintas (por exemplo, *Research Gate* e *IEEE*).

Os resultados dessa seleção distribuídos por fonte estão resumidos na tabela 2. Foram inicialmente analisados 850 trabalhos como resultados da busca. Desses analisados, foi feita a leitura dos títulos e então 83 trabalhos foram selecionados. Por fim, foi feita uma filtragem, consistindo na leitura dos resumos (*abstracts*) e remoção de resultados duplicados, concluindo com 34 trabalhos filtrados.

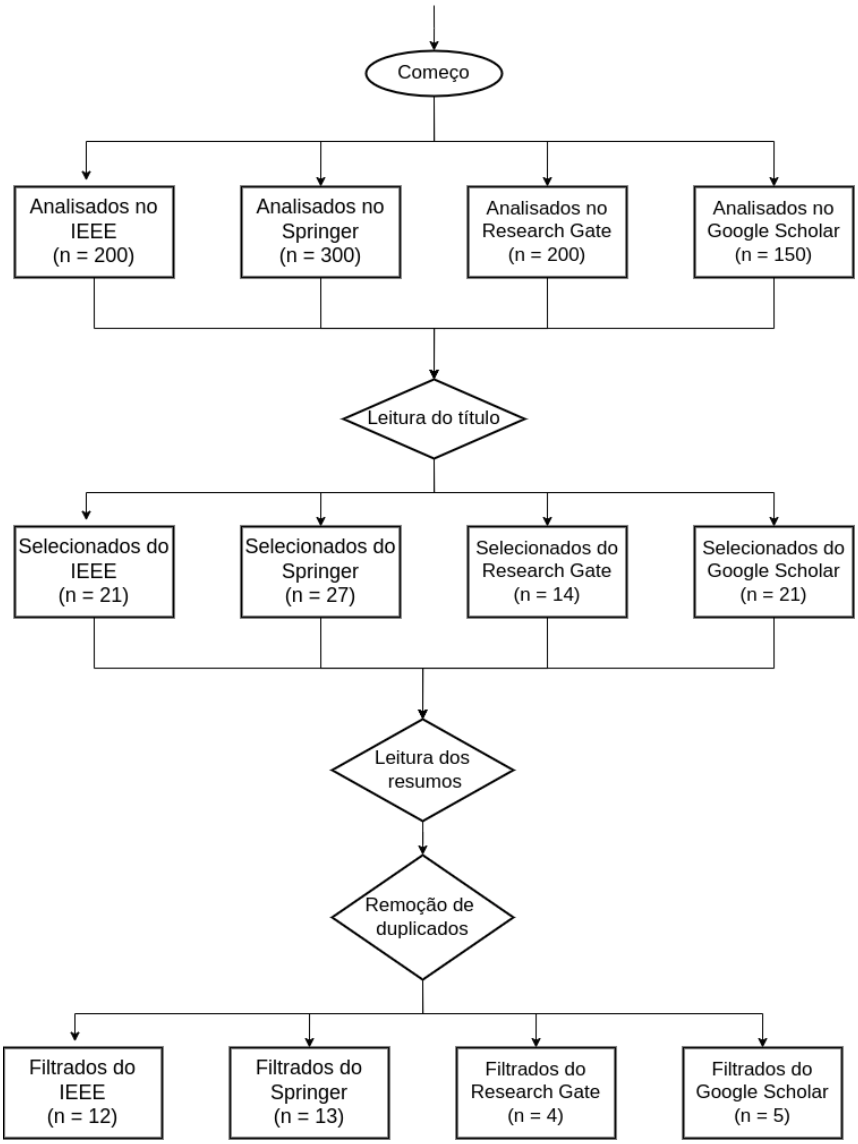

Figura 4. Fluxo do processo da revisão

4.1. Escopo da revisão

A pesquisa e filtragem foram feitas seguindo os seguintes parâmetros:

- Tratam de qualidade de dados especificamente dentro do contexto de *Big Data*;
- Propõem melhoras para os problemas encontrados nas dimensões de qualidade de dados pontuadas pelo trabalho analisado;
- O texto encontra-se disponível em plataformas digitais como:
  - IEEE Xplore (https://ieeexplore.org/);
  - Research Gate (https://www.researchgate.net/);
  - Google Scholar (https://scholar.google.com/);
  - SpringerLink (https://link.springer.com/);
- Estão disponíveis para acesso via acesso institucional ou foram disponibilizados publicamente;
- Foram publicados em inglês;

Tabela 2. Resultados da seleção e filtragem

	Analisado	Selecionado	Filtrado
IEEE	200	21	12
Springer	300	27	13
Research Gate	200	14	4
Google Scholar	150	21	5
Total	***850***	***83***	***34***

- Textos publicados entre os anos 2018 e 2023.

As plataformas de busca foram escolhidas devido à sua relevância, popularidade e credibilidade na esfera acadêmica. *IEEE Xplore* e *Springer* são plataformas de grande destaque, que ofertam uma vasta coleção de trabalhos bem avaliados, em especial na área de tecnologia. Em complemento às anteriores, *Research Gate* e *Google Scholar* foram utilizadas de forma a complementar os resultados da busca com outras fontes.

O intervalo de pesquisa selecionado, entre 2018 e 2023, se deu ao fato de relevância por tempo. Escolher os últimos 5 anos foi pensado de forma a selecionar os tópicos mais recentes discutidos e como está o estado da arte na área de qualidade de dados para *Big Data*.

Aliados a esses parâmetros, a revisão se propõe também a responder as seguintes questões:

- Quais as principais características dos sistemas de *Big Data* hoje em dia?
- Que dimensões são mais recorrentes nos estudos sobre *Big Data*?
- Quais técnicas são comumente utilizadas para tratar os problemas de qualidade de dados para essas dimensões encontradas?
- O quanto o contexto influencia na análise de qualidade de dados para os sistemas de *Big Data*?

5. Considerações sobre a revisão

Após a abordagem descrita na seção acima, nesta seção os principais resultados do processo de análise dos trabalhos filtrados ao fim foram detalhados. As informações apresentadas na tabela 3 visão oferecer uma visão abrangente de atualizada do panorama atual dos estudos sobre o domínio de qualidade de dados no contexto de *Big Data*.

Tabela 3. Contribuições dos trabalhos analisados na revisão

Referência	Contribuições
[Abdallah 2019]	- Divide dimensões de qualidade de dados em quatro perspectivas: perspectiva dos dados, da gestão, do processamento e serviços e do usuário.
[Taleb et al. 2019]	- Considera 9 V's como características do *Big Data*.

Continuação na próxima página

Tabela 3 – (Continuação)

Referência	Contribuições
	- *Data profiling*: processo de definição de um perfil pros dados nas diferentes etapas do pipeline. - Propõe uso de amostragem. - Ataca majoritariamente dados estruturados. - Trabalhos futuros: atacar regras de *data profiling* para dados não-estruturados (que representam cerca de 80% dos dados em *Big Data*).
[Taleb et al. 2018b]	- Considera 10 V's como características do *Big Data*. - Propõe avaliação de qualidade para dados não-estruturados. - Descreve tipos de dados não-estruturados, bem como estratégias usadas para tratamento de cada tipo. - A qualidade dos dados não-estruturados está atrelada à estrutura que os dados serão postos pós-processamento: Quanto mais estruturável, maior a qualidade e mais fácil a metrificação. - Levanta 5 etapas para extração de valor em dados não-estruturados, dentre elas, amostragem de dados.
[Fadlallah et al. 2023a]	- Define problemas de qualidade de dados em três categorias: a nível de gestão, a nível de processamento e a nível de abstração e padronização. - Escassez de soluções para *Big Data*, aplicações geralmente só comentam sobre limpeza de dados. - Propõe uso de *data profiling*. - Segue padrão da ISO/IEC 25012. - Sugere criação de repositórios de domínio de conhecimento.
[Fadlallah et al. 2023b]	- Propõe uma metodologia que considera contexto. - Faz levantamento de soluções apontando pontos positivos e negativos, bem como o tipo de dados apropriados pra cada uma. - Sugere criação de repositórios de domínio de conhecimento.
[Wahyudi et al. 2018c]	- Considera 11 V's como características do *Big Data*. - Sugere que pesquisas devem focar mais na interpretabilidade dos dados.

Continuação na próxima página

Tabela 3 – (Continuação)

Referência	Contribuições
	- Estudos de casos com empresa de telecomunicação e de manufatura.
[Taleb et al. 2018a]	- Considera 10 V's como características do *Big Data*. - Define 4 categorias de dimensões de qualidade de dados: intrínseca, contextual, representacional, acessibilidade. - Propõe um *framekwork*, em que cada etapa gera um relatório capaz de propor melhoras para diferentes etapas do ciclo de vida.
[Salih et al. 2019]	- Considera 5 V's como características do *Big Data*. - Propõe uso de amostragem de dados e data profiling.
[Onyeabor and Ta'a 2019]	- Divide dimensões de qualidade de dados em duas: intrínseca e contextual. - Dimensões relacionadas a processo: latência, tempo de resposta, *throughput*, capacidade e escalabilidade.
[Arolfo and Vaisman 2018]	- Considera 4 V's como características do *Big Data*. - Aponta que modelos de gestão de dados devem levar em consideração tempo, espaço, contexto histórico, contexto e perfil de usuário. - Dimensões de qualidade de dados apontadas: acurácia, completude, consistência, temporalidade, confiança, redundância, utilidade, acessibilidade... - Classifica *Big Data* de acordo com a fonte dos dados: humano, mediado por processos, gerada a partir máquinas. - Estabelece métricas para quantificação das dimensões de qualidade de dados. - Traz uma implementação utilizando Apache Kafka e Apache Zookeeper. - Dados utilizados da API do Twitter.
[Elouataoui et al. 2022b]	- Considera 7 V's como características do *Big Data*. - Define 4 novas métricas de qualidade de dados: integridade, acessibilidade, facilidade de manipulação e segurança. - Define dimensões em 5 categorias diferentes: confiabilidade, disponibilidade, usabilidade, validade e pertinência.

Continuação na próxima página

Tabela 3 – (Continuação)

Referência	Contribuições
	- Define fórmulas para as métricas das dimensões. - Traz uma implementação utilizando Apache Spark e Great Expectations. - Propõe um modelo com pesos em métricas com base no modelo de negócio.
[Desai 2018]	- Faz levantamento de ferramentas para trabalhar com *Big Data*. - Propõe definição algébrica e matricial para cálculo de métricas de qualidade.
[Elouataoui et al. 2022a]	- Considera 14 V's como características do *Big Data*. - Agrupa dimensões em 4 categorias: confiabilidade, disponibilidade, usabilidade e relevância. - Pré-processamento e processamento são as etapas que mais receberam destaque na literatura. - A maioria das pesquisas foca em dimensões em comum com dados tradicionais, como acurácia, completude e consistência. - *Frameworks* atuais utilizam amostragem de dados, *data profiling* e inteligência artificial, mas apenas para definir as regras de qualidade. - Sugere que as técnicas acima poderiam ser aplicadas para identificar também potenciais anomalias nos dados. - Contexto é extremamente relevante na definição do *framework* e deveria ser considerado nas pesquisas.
[Wahyudi et al. 2018a]	- Considera 11 V's como características do *Big Data*. - Aplicação dentro de um contexto específico: bancário e financeiro. - 4 categorias de dimensões de qualidade de dados: intrínseca, representacional, acessibilidade e contextual. - Constrói um mapeamento entre dimensões de qualidade e características de *Big Data* aplicadas ao contexto em estudo.
[Ehrlinger and Wöß 2022]	- Apresenta fórmulas matemáticas para cálculo de métricas de dimensões de qualidade de dados.

Continuação na próxima página

Tabela 3 – (Continuação)

Referência	Contribuições
	- Cataloga ferramentas de gestão e avaliação de qualidade de dados não associadas a um domínio específico.
[Ridzuan et al. 2022]	- Considera 5 V's como características do *Big Data*. - Constrói um levantamento dos desafios de *Big Data* associados a cada característica.
[Ghasemaghaei and Calic 2019]	- Segmentação de dimensões de qualidade de dados em: intrínseca, contextual, representacional e acessibilidade. - Levantamento de testes estatísticos para avaliar o impacto de qualidade de *Big Data* na tomada de decisão.
[Han and Jochum 2020]	- Uso de aprendizado de máquina para controle de qualidade de dados. - Aplicação dentro de contexto específico: dados geográficos.
[Hongxun et al. 2018]	- Aplicação dentro de contexto específico: sistema elétrico. - Uso de 6 dimensões de qualidade de dados: redundância, integridade, acurácia, consistência, temporalidade e inteligência.
[Hossen et al. 2020]	- Apontamento da importância de qualidade de dados na esfera do *e-commerce*. - Levantamento de ferramentas para manutenção de qualidade de dados.
[Juneja and Das 2019]	- Aplicação dentro de contexto específico: monitoramento de tempo e clima. - Considera 5 V's como características do *Big Data*. - Classificação de dimensões de qualidade de dados em: completude, temporalidade, conformidade, unicidade, integridade, consistência, acurácia.
[Loetpipatwanich and Vichitthamaros 2020]	- Ferramenta: Pacote Python para qualidade de dados. - Uso de três dimensões de qualidade de dados: completude, integridade e consistência.
[Molinari and Nollo 2020]	- Considera 5 V's como características do *Big Data*. - Aplicação dentro de contexto específico: medicina e plano de saúde. - Veracidade é o mais crucial.

Continuação na próxima página

Tabela 3 – (Continuação)

Referência	Contribuições
	- Ressalta a importância de um identificador único para os registros.
[Merino et al. 2020]	- Propõe 4 métodos de avaliação de qualidade em *Big Data*: *embedded*, *parallel*, *in-line* e *independent*. - A escolha do método será dada de acordo com a natureza da aplicação. - Caso de estudo: *smart cities*. - Exemplo de aplicação: predição de passageiros pegando ônibus.
[Kumar et al. 2019]	- Aplicação dentro de contexto específico: dados atmosféricos e climáticos.
[Shanmugam et al. 2023]	- Apresenta dimensões de qualidade de dados em 5 categorias: disponibilidade, usabilidade, confiabilidade, relevância e qualidade de apresentação.
[Taleb et al. 2021]	- Divide dimensões de qualidade de dados em 4 categorias: intrínseco, contextual, representacional e acessibilidade. - Aponta problemas de qualidade de dados a nível de dados e a nível de processos dentro do ciclo de vida de *Big Data*. - Apresenta fórmulas matemáticas para calcular métricas. - Propões o uso de *data profiling* e amostragem de dados.
[Kothapalli 2023]	- Levantamento de ferramentas para gestão de pipeline de qualidade de *Big Data*: Apache NiFi, Apache Hadoop, Apache Storm, Talend, Trifacta, DataRobot e Apache Zeppelin. - Comenta sobre uso de aprendizado de máquina na detecção de *outliers*, reconhecimento de padrões e predição. - Comenta desafios para diferentes indústrias como: saúde, finanças, vendas, manufaturação, governo. - Aponta algoritmos e técnicas usadas no trabalho de qualidade de dados como: análise estatística, inteligência artificial, *data profiling*, *data cleaning*, *data matching*, enriquecimento de dados, padronização de dados, integração de dados, histórico de dados e governança de dados.

Continuação na próxima página

Tabela 3 – (Continuação)

Referência	Contribuições
	- Levantamento de técnicas para manutenção de qualidade de dados em *Big Data*: automatizado, *crowdsourcing*, tecnologias semânticas (ontologias e *linked data*), *blockchain* e governança de dados.
[Wahyudi et al. 2018b]	- Divide dimensões de qualidade de dados em 4 categorias: intrínseco, contextual, representacional e acessibilidade. - Estudo de caso com empresa de telecomunicação. - Definição do pipeline de *Big Data* como: descoberta, acesso, exploração, análise e, por fim gestão, que faz parte de todo ciclo. - Propõe um modelo de reconhecimento de padrão de processos para identificar falhas em qualidade de dados.
[Wong and Wong 2020]	- Aplicação dentro de contexto específico: sistema bancário. - Uso de redes neurais profundas para calcular peso de data noise.
[Wook et al. 2021]	- Considera 17 V's como características do *Big Data*. - Divisão de dimensões de qualidade em 4 categorias: intrínseca, contextual, representacional e acessibilidade. - Acessibilidade (facilidade de manipulação) pode influenciar fortemente na aplicação final. - Usa métricas estatísticas para avaliar suas hipóteses de impacto de dimensões de qualidade sobre características de *Big Data*. - Aponta segurança e privacidade como caminhos futuros relevantes de pesquisa.
[Zhang 2020]	- Propõe um modelo definido matematicamente para governança de dados baseado em rastreamento de dados num *loop* fechado de validação, rastreio e revisão. - Modelo se baseia em três entidades principais: *data owner*, *data provider* e *data user*.
[Gyulgyulyan et al. 2019]	- Considera 7 V's como características do *Big Data*. - Propõe um modelo não para corrigir problemas de qualidade em *Big Data*, mas para alertar sobre esses problemas.

Continuação na próxima página

Tabela 3 – (Continuação)

Referência	Contribuições
	- Comenta dimensões de qualidade de dados tradicionais como: consistência, unicidade, acurácia, completude e temporalidade. - Traz também outras dimensões específicas do contexto de *Big Data*: sincronização, interpretabilidade e confiabilidade.
[Montero et al. 2021]	- Revisa literatura dos modelos de qualidade em *Big Data* propostos entre 2010 e 2020. - Houve um aumento considerável de trabalhos no tópico a partir de 2014, sendo 67% dos trabalhos selecionados publicados entre 2018 e 2020. - Quase 75% foi construída independente de contexto, ou seja, pode ser aplicado para qualquer base. - As dimensões de qualidade de dados mais comuns foram: completude, acurácia, consistência e temporalidade.

Ao tomar os estudos avaliados na revisão acima, é possível levantar algumas conclusões sobre o estado atual das pesquisas sobre qualidade de dados em *Big Data*. São elas:

- Não existe um consenso sobre a quantidade de características de *Big Data* a serem consideradas, identifica-se uma unanimidade quanto aos 3 V's levantados por [Laney 2001]. Em seguida, existem características mais citadas: *valor, validade, vulnerabilidade, visualização, variabilidade, veracidade*.
- Nos trabalhos avaliados, existem duas classificações para dimensões de qualidade de dados mais recorrentes. Uma parte dos trabalhos as divide em: intrínsecas, contextuais, representacionais e acessibilidade; já outros propõem a seguinte classificação: confiabilidade, disponibilidade, usabilidade, validade e pertinência. Alguns trabalhos trazem também variações dessas duas classificações, trazendo menos categorias.
- Das dimensões levantadas nos trabalhos, algumas das mais citadas são: *integridade, acurácia, unicidade, consistência, segurança, acesso, valor adicionado, completude* e *temporalidade*.
- Dentre os *frameworks* de qualidade de dados propostos, as técnicas que mais se destacaram foram amostragem de dados, *data profiling* bem como o uso de algoritmos de *machine learning*. O propósito da primeira é reduzir o custo do processamento de dados para identificação de problemas de qualidade. Já a segunda visa a construção de um perfil de qualidade dos dados em cada etapa do ciclo de vida de forma que seja usado como referência do estado atual dos dados e de que melhorias eles precisam, se for o caso. Por fim, o último pretende fazer uso de inteligência artificial para identificar mais facilmente padrões e *outliers*.
- Uma pequena parcela dos trabalhos foi construído dentro de um contexto específico (14%), já os outros foram construídos para ser independentes de contexto.

Por outro lado, foi levantada a importância do contexto na definição dos elementos de qualidade de dados relevantes, visto que cada contexto pode trazer elementos com maior importância que outros.

- Poucos trabalhos citam as ferramentas que utilizaram para a construção e manipulação dos seus *pipelines* de *Big Data*, mas as principais citadas foram: *Apache NiFi, Apache Hadoop, Apache Storm, Talend, Trifacta, DataRobot, Apache Zeppelin, Apache Kafka, Apache Zookeeper, DataCleaner* e *Great Expectations*.
- Por fim, os trabalhos trazem alguns avanços que merecem destaque. Um ponto relevante em bastante ascensão foi uma crescente preocupação sobre segurança e privacidade. Outro destaque foi uma atenção maior para o destinatário final, ou seja, uma preocupação com a interpretabilidade e visualização pelo usuário

6. Conclusão

Big Data atraiu a atenção dos pesquisadores e da indústria devido a todo seu potencial, hoje amplamente explorado. Mas esse potencial não chega a ter um bom aproveitamento se os dados forem de baixa qualidade. Este trabalho revisou os fundamentos básicos do que é *Big Data*, bem como do que se entende por qualidade de dados, construindo uma representação de um *pipeline* (ou ciclo de vida) para *Big Data* e seus problemas. Foi feita uma revisão de 34 trabalhos extraídos de 4 fontes distintas. Em trabalhos futuros, baseado nos resultados da revisão, sugere-se algumas vias de pesquisa:

1. Estudo sobre qualidade de dados para *Big Data* dentro de contextos específicos, bem como sua implementação.
2. Impacto da segurança e privacidade, em particular pós legislações como LGPD e GDPR, por exemplo, na qualidade dos sistemas de Big Data
3. Uso de técnicas de *machine learning* não apenas para etapas de processamento, mas também para a detecção de *outliers* e seu impacto na qualidade de dados.

Referências

Abdallah, M. (2019). Big data quality challenges. In *2019 International Conference on Big Data and Computational Intelligence (ICBDCI)*, pages 1–3. IEEE.

Arockia Panimalar.S, Varnekha Shree.S, V. K. (2017). The 17 v's of big data. In *International Research Journal of Engineering and Technology (IRJET)*.

Arolfo, F. and Vaisman, A. (2018). Data quality in a big data context. In *European Conference on Advances in Databases and Information Systems*, pages 159–172. Springer.

Desai, K. Y. (2018). Big data quality modeling and validation.

DOMO (2022). Data never sleeps 10.0. `https://www.domo.com/data-never-sleeps/`. Acesso em 2023-08-13.

Ehrlinger, L. and Wöß, W. (2022). A survey of data quality measurement and monitoring tools. *Frontiers in big data*, 5:850611.

Elouataoui, W., Alaoui, I. E., and Gahi, Y. (2022a). Data quality in the era of big data: a global review. *Big Data Intelligence for Smart Applications*, pages 1–25.

Elouataoui, W., El Alaoui, I., El Mendili, S., and Gahi, Y. (2022b). An advanced big data quality framework based on weighted metrics. *Big Data and Cognitive Computing*, 6(4):153.

Fadlallah, H., Kilany, R., Dhayne, H., El Haddad, R., Haque, R., Taher, Y., and Jaber, A. (2023a). Bigqa: Declarative big data quality assessment. *ACM Journal of Data and Information Quality*, 15(3):1–30.

Fadlallah, H., Kilany, R., Dhayne, H., El Haddad, R., Haque, R., Taher, Y., and Jaber, A. (2023b). Context-aware big data quality assessment: a scoping review. *ACM Journal of Data and Information Quality*, 15(3):1–33.

Ghasemaghaei, M. and Calic, G. (2019). Can big data improve firm decision quality? the role of data quality and data diagnosticity. *Decision Support Systems*, 120:38–49.

Gyulgyulyan, E., Aligon, J., Ravat, F., and Astsatryan, H. (2019). Data quality alerting model for big data analytics. In *New Trends in Databases and Information Systems: ADBIS 2019 Short Papers, Workshops BBIGAP, QAUCA, SemBDM, SIMPDA, M2P, MADEISD, and Doctoral Consortium, Bled, Slovenia, September 8–11, 2019, Proceedings 23*, pages 489–500. Springer.

Han, W. and Jochum, M. (2020). A machine learning approach for data quality control of earth observation data management system. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 3101–3103. IEEE.

Hongxun, T., Honggang, W., Kun, Z., Mingtai, S., Haosong, L., Zhongping, X., Taifeng, K., Jin, L., and Yaqi, C. (2018). Data quality assessment for on-line monitoring and measuring system of power quality based on big data and data provenance theory. In *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 248–252. IEEE.

Hossen, M. I., Goh, M., Hossen, A., and Rahman, M. A. (2020). A study on the aspects of quality of big data on online business and recent tools and trends towards cleaning dirty data. In *2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, pages 209–213. IEEE.

Hu, H., Wen, Y., Chua, T.-S., and Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE access*, 2:652–687.

Juneja, A. and Das, N. N. (2019). Big data quality framework: Pre-processing data in weather monitoring application. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 559–563. IEEE.

Kothapalli, M. (2023). The challenges of data quality and data quality assessment in the big data.

Kumar, J., Crow, M. C., Devarakonda, R., Giansiracusa, M., Guntupally, K., Olatt, J. V., Price, Z., Shanafield, H. A., and Singh, A. (2019). Provenance–aware workflow for data quality management and improvement for large continuous scientific data streams. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 3260–3266. IEEE.

Laney, D. (2001). 3d data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6.

Loetpipatwanich, S. and Vichitthamaros, P. (2020). Sakdas: a python package for data profiling and data quality auditing. In *2020 1st International Conference on Big Data Analytics and Practices (IBDAP)*, pages 1–4. IEEE.

Merino, J., Xie, X., Parlikad, A. K., Lewis, I., and McFarlane, D. (2020). Impact of data quality in real-time big data systems.

Molinari, A. and Nollo, G. (2020). The quality concerns in health care big data. In *2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON)*, pages 302–305. IEEE.

Montero, O., Crespo, Y., and Piatini, M. (2021). Big data quality models: a systematic mapping study. In *Quality of Information and Communications Technology: 14th International Conference, QUATIC 2021, Algarve, Portugal, September 8–11, 2021, Proceedings 14*, pages 416–430. Springer.

Oliveira, P., Rodrigues, F., and Henriques, P. R. (2005). A formal definition of data quality problems. In *ICIQ*.

Onyeabor, G. A. and Ta'a, A. (2019). A model for addressing quality issues in big data. In *Recent Trends in Data Science and Soft Computing: Proceedings of the 3rd International Conference of Reliable Information and Communication Technology (IRICT 2018)*, pages 65–73. Springer.

Ridzuan, F., Wan Zainon, W. M. N., and Zairul, M. (2022). A thematic review on data quality challenges and dimension in the era of big data. In *Proceedings of the 12th National Technical Seminar on Unmanned System Technology 2020: NUSYS'20*, pages 725–737. Springer.

Salih, F. I., Ismail, S. A., Hamed, M. M., Mohd Yusop, O., Azmi, A., and Mohd Azmi, N. F. (2019). Data quality issues in big data: a review. In *Recent Trends in Data Science and Soft Computing: Proceedings of the 3rd International Conference of Reliable Information and Communication Technology (IRICT 2018)*, pages 105–116. Springer.

Shanmugam, D., Dhilipan, J., Prabhu, T., Sivasankari, A., and Vignesh, A. (2023). The management of data quality assessment in big data presents a complex challenge, accompanied by various issues related to data quality. *Research Highlights in Mathematics and Computer Science Vol. 8*, pages 78–91.

Sidi, F., Panahy, P. H. S., 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. IEEE.

Taleb, I., Serhani, M. A., Bouhaddioui, C., and Dssouli, R. (2021). Big data quality framework: a holistic approach to continuous quality management. *Journal of Big Data*, 8(1):1–41.

Taleb, I., Serhani, M. A., and Dssouli, R. (2018a). Big data quality: A survey. In *2018 IEEE International Congress on Big Data (BigData Congress)*, pages 166–173. IEEE.

Taleb, I., Serhani, M. A., and Dssouli, R. (2018b). Big data quality assessment model for unstructured data. In *2018 International Conference on Innovations in Information Technology (IIT)*, pages 69–74. IEEE.

Taleb, I., Serhani, M. A., and Dssouli, R. (2019). Big data quality: a data quality profiling model. In *World Congress on Services*, pages 61–77. Springer.

The Economist (2017). The world's most valuable resource is no longer oil, but data.

Uddin, M. F., Gupta, N., et al. (2014). Seven v's of big data understanding big data to extract value. In *Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education*, pages 1–5. IEEE.

Wahyudi, A., Farhani, A., and Janssen, M. (2018a). Relating big data and data quality in financial service organizations. In *Challenges and Opportunities in the Digital Era: 17th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2018, Kuwait City, Kuwait, October 30–November 1, 2018, Proceedings 17*, pages 504–519. Springer.

Wahyudi, A., Kuk, G., and Janssen, M. (2018b). A process pattern model for tackling and improving big data quality. *Information Systems Frontiers*, 20:457–469.

Wahyudi, A., Pekkola, S., and Janssen, M. (2018c). Representational quality challenges of big data: Insights from comparative case studies. In *Challenges and Opportunities in the Digital Era: 17th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2018, Kuwait City, Kuwait, October 30–November 1, 2018, Proceedings 17*, pages 520–538. Springer.

Wand, Y. and Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33.

Wong, K. Y. and Wong, R. K. (2020). Big data quality prediction on banking applications. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 791–792. IEEE.

Wook, M., Hasbullah, N. A., Zainudin, N. M., Jabar, Z. Z. A., Ramli, S., Razali, N. A. M., and Yusop, N. M. M. (2021). Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modelling. *Journal of Big Data*, 8:1–15.

Zhang, G. (2020). A data traceability method to improve data quality in a big data environment. In *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*, pages 290–294. IEEE.