ReGeo ISSN: 2177-3246

GOVERNANÇA E QUALIDADE DE DADOS NA ORIGEM: UM FRAMEWORK PREVENTIVO PARA SÉRIES TEMPORAIS INDUSTRIAIS ANTES DO ARMAZENAMENTO EM DATA LAKES

GOVERNANCE AND DATA QUALITY AT THE SOURCE: A PREVENTIVE FRAMEWORK FOR INDUSTRIAL TIME SERIES BEFORE STORAGE IN DATA LAKES

GOBERNANZA Y CALIDAD DE DATOS EN LA ORIGEN: UN FRAMEWORK PREVENTIVO PARA SERIES TEMPORALES INDUSTRIALES ANTES DEL ALMACENAMIENTO EN DATA LAKES

https://doi.org/10.56238/revgeov17n10-162

Data de submissão: 05/09/2026 **Data de publicação:** 20/10/2026

Anna Hellen Andrade Moura
Bacharel em Engenharia de Software
E-mail: annahellen40@gmail.com
Lattes: http://lattes.cnpq.br/7592942564785724

RESUMO

A digitalização industrial multiplicou sensores, controladores e historiadores que produzem séries temporais em volume crescente, e consolidou o data lake como destino padrão desses fluxos. A prática dominante, contudo, transfere a qualidade para depois do armazenamento: ingere-se tudo e corrige-se a jusante, quando a análise falha. Este artigo propõe um framework preventivo de governança e qualidade de dados aplicado na origem, entre o sensor e o lake, específico para séries temporais industriais. A investigação, de natureza teórico-propositiva com base em pesquisa bibliográfica estruturada, articulou os fundamentos consolidados da qualidade de dados (as dimensões clássicas de acurácia, completude, consistência, atualidade, validade e unicidade), os corpos normativos (DAMA-DMBOK, série ISO 8000, ISO/IEC 38505 e os padrões industriais ISA-95 e OPC UA), o modelo de alocação de decisões de governança e a literatura revisada por pares sobre validação de dados em sistemas de aprendizado de máquina e qualidade de dados de sensores. A contribuição organiza-se em três peças. Primeira, o framework de cinco data quality gates encadeados (G1, aquisição instrumentada; G2, validação na borda; G3, contextualização por modelo de ativos; G4, contrato de dados versionado; G5, ingestão governada com quarentena), no qual as leituras suspeitas são marcadas e preservadas, nunca descartadas silenciosamente, e cada gate tem controles, papéis e métricas próprios. Segunda, a matriz de interceptação, que vincula cada defeito típico das séries industriais (deriva de sensor, lacunas de amostragem, carimbos de tempo inconsistentes, valores congelados, picos espúrios e heterogeneidade semântica) ao gate capaz de interceptá-lo com o contexto disponível. Terceira, a análise do impacto a jusante na manutenção preditiva, mostrando como cada defeito não interceptado se converte em falso positivo, falso negativo ou padrão artificial nos modelos que justificam o próprio lake. Conclui-se que a qualidade tratada como propriedade de projeto da tubulação de dados, e não como remediação analítica, converte o data lake de repositório de riscos em ativo confiável, condição para análises avançadas e aplicações de inteligência artificial industrial.

Palavras-chave: Data Quality Gates. Manutenção Preditiva. Edge Computing.

ABSTRACT

Industrial digitalization has multiplied sensors, controllers and historians that produce time series in growing volume, and has consolidated the data lake as the standard destination for these flows. The

dominant practice, however, transfers quality to after storage: everything is ingested and corrected downstream, when analysis fails. This article proposes a preventive data governance and quality framework applied at the source, between the sensor and the lake, specific to industrial time series. The investigation, of a theoretical-propositional nature based on structured bibliographic research, articulated the consolidated foundations of data quality (the classic dimensions of accuracy, completeness, consistency, timeliness, validity and uniqueness), the normative bodies (DAMA-DMBOK, the ISO 8000 series, ISO/IEC 38505 and the industrial standards ISA-95 and OPC UA), the governance decision-allocation model, and the peer-reviewed literature on data validation in machine learning systems and sensor data quality. The contribution is organized in three parts. First, the framework of five chained data quality gates (G1, instrumented acquisition; G2, edge validation; G3, contextualization by asset model; G4, versioned data contract; G5, governed ingestion with quarantine), in which suspicious readings are flagged and preserved, never silently discarded, and each gate has its own controls, roles and metrics. Second, the interception matrix, which links each typical defect of industrial series (sensor drift, sampling gaps, inconsistent timestamps, frozen values, spurious spikes and semantic heterogeneity) to the gate capable of intercepting it with the available context. Third, the analysis of the downstream impact on predictive maintenance, showing how each unintercepted defect becomes a false positive, a false negative or an artificial pattern in the very models that justify the lake. It is concluded that quality treated as a design property of the data pipeline, rather than as analytical remediation, converts the data lake from a repository of risks into a reliable asset, a condition for advanced analytics and industrial artificial intelligence applications.

Keywords: Data Quality Gates. Predictive Maintenance. Edge Computing.

RESUMEN

La digitalización industrial ha multiplicado sensores, controladores e historiadores que producen series temporales en volumen creciente, y consolidó el data lake como destino estándar de estos flujos. La práctica dominante, sin embargo, transfiere la calidad para después del almacenamiento: se ingiere todo y se corrige aguas abajo, cuando el análisis falla. Este artículo propone un marco preventivo de gobernanza y calidad de datos aplicado en el origen, entre el sensor y el lake, específico para series temporales industriales. La investigación, de naturaleza teórico-propositiva con base en investigación bibliográfica estructurada, articuló los fundamentos consolidados de la calidad de datos (las dimensiones clásicas de exactitud, completitud, consistencia, actualidad, validez y unicidad), los cuerpos normativos (DAMA-DMBOK, serie ISO 8000, ISO/IEC 38505 y los estándares industriales ISA-95 y OPC UA), el modelo de asignación de decisiones de gobernanza y la literatura revisada por pares sobre validación de datos en sistemas de aprendizaje automático y calidad de datos de sensores. La contribución se organiza en tres componentes. Primero, el marco de cinco data quality gates encadenados (G1, adquisición instrumentada; G2, validación en el borde; G3, contextualización por modelo de activos; G4, contrato de datos versionado; G5, ingestión gobernada con cuarentena), en el cual las lecturas sospechosas se marcan y se preservan, nunca se descartan silenciosamente, y cada gate tiene controles, roles y métricas propios. Segundo, la matriz de interceptación, que vincula cada defecto típico de las series industriales (deriva de sensor, brechas de muestreo, marcas de tiempo inconsistentes, valores congelados, picos espurios y heterogeneidad semántica) al gate capaz de interceptarlo con el contexto disponible. Tercero, el análisis del impacto aguas abajo en el mantenimiento predictivo, mostrando cómo cada defecto no interceptado se convierte en falso positivo, falso negativo o patrón artificial en los modelos que justifican el propio lake. Se concluye que la calidad tratada como propiedad de diseño de la tubería de datos, y no como remediación analítica, convierte el data lake de repositorio de riesgos en activo confiable, condición para análisis avanzados y aplicaciones de inteligencia artificial industrial.

ISSN: 2177-3246

Palabras clave: Data Quality Gates. Mantenimiento Predictivo. Edge Computing.

1 INTRODUÇÃO

A fábrica contemporânea é, antes de tudo, uma produtora de séries temporais. Cada sensor de temperatura, pressão, vibração ou vazão, cada controlador lógico e cada sistema supervisório emite leituras carimbadas no tempo, e a arquitetura de dados da indústria 4.0 convergiu para um padrão: coletar esses fluxos e depositá-los em data lakes, repositórios de baixo custo que armazenam dados brutos em escala para uso analítico posterior. A escala do fenômeno é bem documentada: em projeção publicada em 2018, em estudo da IDC patrocinado pela Seagate, a datasfera global, então estimada em 33 zettabytes, alcançaria 175 zettabytes em 2025, com parcela crescente gerada na borda industrial e por dispositivos conectados (Reinsel; Gantz; Rydning, 2018). O lake tornou-se o ponto de encontro entre a operação e a análise, alimentando desde relatórios de eficiência até modelos de manutenção preditiva e aplicações de inteligência artificial.

A prática dominante de povoá-lo, contudo, carrega um vício de projeto. O paradigma do esquema na leitura, que dispensa validação prévia em nome da flexibilidade, degenerou com frequência no que a literatura apelidou de data swamp, o pântano de dados: um acervo volumoso, sem metadados confiáveis, sem linhagem e sem garantias de qualidade, no qual cada análise recomeça o trabalho de descobrir o que os números significam e se merecem confiança (Hai; Geisler; Quix, 2016; Armbrust *et al.*, 2021). O custo dessa transferência da qualidade para a jusante é mensurável. A Gartner estima que dados de má qualidade custam, em média, US$ 12,9 milhões por ano a cada organização (Gartner, 2021). Para a economia norte-americana, o custo agregado foi calculado em US$ 3,1 trilhões anuais (Redman, 2016). E as pesquisas com profissionais convergem: cientistas de dados dedicam cerca de 60% do tempo à limpeza e organização de dados, fração que se aproxima de 80% quando somadas a coleta e a preparação, precisamente o trabalho que a validação na origem evitaria repetir a cada projeto (CrowdFlower, 2016; Press, 2016). O princípio em jogo é o mais antigo da computação: garbage in, garbage out, entra lixo, sai lixo. Nos sistemas de aprendizado de máquina, a literatura deu a esse princípio forma técnica precisa: a maior parte do débito oculto desses sistemas não mora no código do modelo, mora nas dependências de dados, e defeitos de dados propagam-se em cascata, compondo danos difíceis de rastrear exatamente porque entram cedo e silenciosamente (Sculley *et al.*, 2015; Sambasivan *et al.*, 2021). Este artigo toma esse princípio como fio condutor: se o lixo não pode sair, é porque não deve entrar sem ser, no mínimo, identificado na entrada.

Nas séries temporais industriais, o problema tem fisionomia própria. Os defeitos não são os de um cadastro, são os da física e da infraestrutura: sensores derivam com o desgaste e a falta de calibração, redes perdem pacotes e abrem lacunas de amostragem, relógios dessincronizados produzem carimbos de tempo fora de ordem, instrumentos travados repetem o mesmo valor por horas, picos

espúrios nascem de interferência elétrica, e a mesma grandeza chega ao lake em unidades, faixas e convenções distintas conforme o equipamento e o fornecedor, taxonomia que a revisão sistemática da qualidade de dados de sensores consolidou (Teh; Kempa-Liehr; Wang, 2020). Corrigir esses defeitos meses depois, sem o contexto do processo, torna-se tarefa muito mais custosa e menos confiável, porque a informação necessária para julgar uma leitura (o estado do ativo, a ordem de manutenção, a batelada em curso) existe na origem e se dissipa com o tempo.

A pergunta de pesquisa que orienta este estudo é dupla: quais fundamentos de governança e de qualidade de dados se aplicam à especificidade das séries temporais industriais, e como organizá-los em um framework preventivo que atue entre o sensor e o data lake, interceptando os defeitos antes do armazenamento? A justificativa opera em três planos. No plano científico, a literatura de governança consolidou modelos de decisão e dimensões de qualidade, mas sua tradução para o regime de alta frequência, contexto físico e tempo real da indústria permanece dispersa; a produção recente sobre governança corporativa e interoperabilidade de dados confirma a demanda por arquiteturas de referência aplicadas (Souza, 2022; Madureira *et al.*, 2026), e o modelo de integração entre operação e sistemas corporativos da ISA-95 oferece a âncora industrial para essa tradução (IEC, 2013). No plano econômico, os custos citados mostram que a remediação a jusante é o modo mais caro de obter qualidade. No plano tecnológico, as aplicações que justificam o lake (gêmeos digitais, manutenção preditiva, otimização de processo) são exatamente as mais sensíveis a dados ruins, na lógica do garbage in, garbage out que a seção 3.5 desenvolve. O objetivo geral é propor um framework preventivo de governança e qualidade de dados na origem para séries temporais industriais. São objetivos específicos: caracterizar os defeitos típicos dessas séries e seu custo a jusante; sistematizar os fundamentos normativos e conceituais aplicáveis; estruturar o framework em gates, com controles, papéis e métricas; analisar o impacto dos defeitos não interceptados sobre a manutenção preditiva; e discutir implantação, limites e relação com as arquiteturas de lake existentes.

2 METODOLOGIA

O estudo é um ensaio teórico-propositivo, de natureza qualitativa, apoiado em pesquisa bibliográfica e documental estruturada, cujo percurso de levantamento seguiu as etapas da revisão integrativa, da definição da questão norteadora à busca sistematizada, à seleção por critérios explícitos e à síntese crítica das fontes (Souza; Silva; Carvalho, 2010). A classificação como ensaio propositivo é deliberada: o produto do estudo não é um balanço exaustivo da literatura, e sim um framework derivado analiticamente de fontes selecionadas por pertinência, e as decisões de desenho são do autor, fundamentadas nas fontes, mas não redutíveis a elas.

As buscas foram realizadas em setembro de 2026, nas bases Scopus, Web of Science, IEEE Xplore, SciELO, Google Scholar e DOAJ, e nos repositórios dos corpos normativos e técnicos pertinentes, além dos anais das conferências de sistemas de dados e aprendizado de máquina (SIGMOD, VLDB, NeurIPS, MLSys, CHI). Os descritores utilizados, em português e inglês, foram: governança de dados; qualidade de dados; séries temporais; dados industriais; data lake; internet das coisas industrial; contrato de dados; data governance; data quality; sensor data quality; data validation; machine learning; industrial time series; edge computing; data contract; lakehouse. As combinações empregaram os operadores AND e OR, como em data quality AND time series AND industrial, e data validation AND machine learning.

O corpus compreendeu os fundamentos conceituais clássicos da qualidade de dados e da governança, com destaque para as dimensões de qualidade orientadas ao consumidor do dado e para o modelo de alocação de domínios de decisão; os corpos normativos DAMA-DMBOK, série ISO 8000 e ISO/IEC 38505, além dos padrões industriais ISA-95 (IEC 62264), para o modelo de ativos e a integração operação-corporação, e OPC UA (IEC 62541), para a interoperabilidade semântica; a literatura revisada por pares sobre qualidade de dados de sensores, validação de dados em escala e em sistemas de aprendizado de máquina, débito técnico e cascatas de dados; a literatura de arquiteturas de dados analíticos, incluída a proposta do lakehouse, a organização em zonas e a documentação técnica da arquitetura em camadas bronze, prata e ouro; as evidências quantitativas sobre custos da má qualidade e alocação do tempo de cientistas de dados; e artigos revisados por pares sobre governança de dados em ambientes corporativos e interoperabilidade. O recorte temporal privilegiou publicações de 1996 a 2026, admitidos os clássicos fundadores. Foram incluídos artigos revisados por pares, normas técnicas publicadas, documentação técnica de referência e literatura cinzenta de instituições reconhecidas (relatórios de consultorias e institutos de pesquisa de mercado), citada preferencialmente pela fonte primária e sinalizada como tal; foram excluídos materiais promocionais sem lastro verificável. Essa escolha tem custo declarado: parte das evidências econômicas provém de fontes com interesse comercial no tema (o estudo da IDC sobre a datasfera foi patrocinado pela Seagate; a pesquisa de alocação de tempo é da CrowdFlower, fornecedora de serviços de preparação de dados), limitação reconhecida na seção 5. A síntese organizou os achados em cinco eixos: defeitos e custos, fundamentos, desenho do framework, impacto a jusante e implantação. O framework proposto é derivado analiticamente dessas fontes, sem estudo de caso empírico próprio, delimitação declarada e retomada nas limitações; nenhum quantitativo de desempenho do framework é afirmado ou estimado. O corpus final citado compreende 23 obras, selecionadas por aderência direta aos cinco eixos; por se tratar de

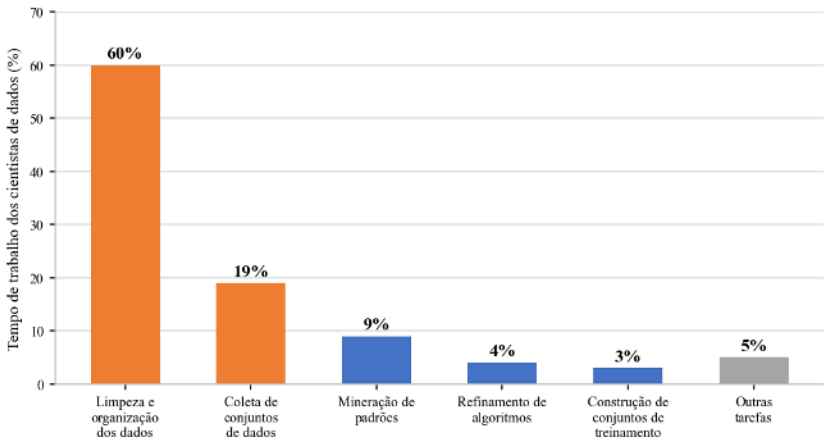
ensaio propositivo, e não de revisão sistemática, não se reporta funil censitário de triagem, e a seleção não pretende exaustividade, limite assumido do desenho.

3 DEFEITOS, FUNDAMENTOS E O DESENHO DO FRAMEWORK

3.1 SÉRIES TEMPORAIS INDUSTRIAIS E O CUSTO DA QUALIDADE ADIADA

O ponto de partida é a assimetria entre o crescimento dos dados e o da capacidade de tratá-los. O Gráfico 1 apresenta a distribuição clássica do tempo de trabalho dos cientistas de dados apurada em pesquisa com profissionais: 60% consumidos em limpeza e organização e 19% na coleta de conjuntos de dados, restando menos de um quinto para a mineração de padrões, o refinamento de algoritmos e a construção de conjuntos de treinamento que constituem sua função-fim (CrowdFlower, 2016; Press, 2016). Essa alocação é o retrato operacional da qualidade adiada: cada projeto analítico paga novamente, em horas especializadas, o débito de validação que a ingestão não pagou uma única vez, o mesmo padrão que a literatura de aprendizado de máquina descreve como cascata de dados, defeitos que entram cedo, compõem-se em silêncio e explodem tarde, quando o custo de correção é máximo (Sambasivan *et al.*, 2021). Os custos agregados acompanham: US$ 12,9 milhões anuais por organização na estimativa da Gartner e US$ 3,1 trilhões anuais para a economia norte-americana no cálculo popularizado pela Harvard Business Review (Gartner, 2021; Redman, 2016). E a escala do insumo só cresce: a datasfera global foi projetada, em 2018, para mais que quintuplicar até 2025, com a borda industrial entre os vetores de expansão, o que torna a triagem manual a jusante progressivamente inviável (Reinsel; Gantz; Rydning, 2018).

Gráfico 1. Distribuição do tempo de trabalho dos cientistas de dados por atividade (%)



Fonte: elaboração própria com base em CrowdFlower (2016) e Press (2016), a partir de pesquisa com profissionais de ciência de dados.

Nas séries industriais, o débito de qualidade tem taxonomia própria, que o framework precisa endereçar defeito a defeito. A deriva de sensor degrada a acurácia de forma lenta e silenciosa, invisível a validações pontuais. As lacunas de amostragem, causadas por falhas de rede ou de coleta, ferem a completude e distorcem agregações. Os carimbos de tempo inconsistentes, por relógios dessincronizados ou fusos mal configurados, quebram a ordenação que define uma série temporal. Os valores congelados, típicos de instrumento travado, simulam estabilidade onde há falha. Os picos espúrios de origem elétrica contaminam estatísticas e disparam alarmes falsos. E a heterogeneidade semântica, a mesma grandeza em unidades, faixas e nomes distintos, mina a consistência entre ativos e plantas (Teh; Kempa-Liehr; Wang, 2020). Esses defeitos tornam-se muito mais custosos e menos confiáveis de detectar meses depois, longe do contexto do processo; todos são interceptáveis na origem, onde o estado do ativo e o conhecimento da operação ainda estão disponíveis.

3.2 FUNDAMENTOS: DIMENSÕES DE QUALIDADE, CORPOS NORMATIVOS E ALOCAÇÃO DE DECISÕES

O framework não parte do zero conceitual. A pesquisa fundadora sobre qualidade de dados estabeleceu que qualidade é adequação ao uso na perspectiva do consumidor do dado, desdobrada em dimensões mensuráveis, das quais acurácia, completude, consistência, atualidade, validade e unicidade formam o núcleo operacional (Wang; Strong, 1996). Os corpos normativos deram a essas dimensões institucionalidade: o DAMA-DMBOK organiza a gestão de dados em áreas de conhecimento que incluem qualidade, metadados, arquitetura e governança como função integradora (DAMA International, 2017); a série ISO 8000 fornece o modelo de referência de processos para a gestão da qualidade de dados, permitindo avaliar maturidade e certificar práticas (ISO, 2016); e a ISO/IEC 38505 estende à governança de dados os princípios de governança corporativa de TI, vinculando o tema à direção da organização, e não apenas à área técnica (ISO/IEC, 2017). No plano industrial, a ISA-95 (IEC 62264) fornece o modelo de referência que o framework pressupõe: a hierarquia de equipamentos e a integração entre o chão de fábrica e os sistemas corporativos que permitem ancorar cada série temporal a um ativo definido (IEC, 2013). No plano organizacional, o modelo clássico de alocação de decisões distingue os domínios decisórios da governança (princípios, qualidade, metadados, acesso e ciclo de vida) e exige definir, para cada um, quem decide e quem responde, na figura de papéis como proprietários e curadores de dados (Khatri; Brown, 2010). A literatura recente confirma a atualidade desse arcabouço nos ambientes corporativos e nos desafios de interoperabilidade de grandes ecossistemas de dados, servindo aqui como apoio contextual (Souza, 2022; Madureira *et al.*, 2026). O

Quadro 1 traduz as dimensões clássicas para a fisionomia das séries industriais, com os defeitos típicos que as ameaçam.

Quadro 1. Dimensões de qualidade de dados e sua manifestação em séries temporais industriais

Dimensão	Definição operacional	Defeitos típicos na série industrial
Acurácia	Correspondência entre o valor registrado e a grandeza física real	Deriva de sensor, descalibração e viés de instalação
Completude	Presença das leituras esperadas na frequência pactuada	Lacunas de amostragem por falha de rede, coleta ou buffer
Consistência	Coerência semântica entre fontes, ativos e períodos	Unidades, faixas e nomenclaturas divergentes entre equipamentos
Atualidade	Disponibilidade do dado dentro da latência exigida pelo uso	Atrasos de transmissão e reprocessamentos tardios
Validade	Aderência a domínios, faixas e regras físicas do processo	Picos espúrios, valores fora de faixa, violações de limites físicos e valores congelados de instrumento travado
Unicidade e ordenação	Ausência de duplicatas e ordem temporal íntegra	Carimbos de tempo dessincronizados, duplicados ou fora de ordem

Fonte: elaboração própria com base em Wang e Strong (1996), DAMA International (2017), ISO (2016) e Teh, Kempa-Liehr e Wang (2020).

3.3 DO LAKE AO PÂNTANO: POR QUE A CORREÇÃO A JUSANTE FALHA

A arquitetura analítica reconheceu o problema e evoluiu na direção certa, mas parou no meio do caminho. A proposta do lakehouse disciplinou o depósito, e a organização do lake em zonas de refinamento progressivo, bronze para o dado bruto, prata para o dado limpo e conformado, ouro para o dado agregado de consumo, deu à qualidade um lugar na topologia, no desenho que a documentação técnica consagrou como arquitetura medalhão (Armbrust *et al.*, 2021; Databricks, 2026). O limite dessas arquiteturas é que operam depois da ingestão: a zona bronze institucionaliza a entrada do defeito, e a transformação bronze-prata reexecuta, em lote e sem contexto de processo, a validação que a origem poderia ter feito com contexto e em tempo real. Para dados transacionais, esse arranjo é frequentemente suficiente. Para séries industriais, três razões o tornam insuficiente. Primeira, a irrecuperabilidade do contexto: julgar se um pico de vibração é evento real ou interferência exige saber o estado operacional do ativo naquele instante, informação que os sistemas de origem têm e o lake não. Segunda, o acúmulo: em fluxos de alta frequência, defeitos entram mais depressa do que rotinas de limpeza os removem, e o pântano cresce (Hai; Geisler; Quix, 2016). Terceira, a multiplicação do retrabalho: sem validação única na origem, cada consumidor da série refaz a própria limpeza, com critérios divergentes que produzem resultados inconsistentes a partir do mesmo dado bruto, exatamente a alocação de esforço que o Gráfico 1 documenta (CrowdFlower, 2016; Armbrust *et al.*, 2021). O

Quadro 2 sistematiza a comparação entre os três pontos possíveis de validação e sustenta o argumento central: a borda é o único ponto que combina contexto máximo, custo mínimo por defeito e reversibilidade plena.

Quadro 2. Três pontos de validação comparados: borda, ingestão e pós-armazenamento

Critério	Validação na borda (antes da ingestão)	Validação na ingestão (entrada do lake)	Correção pós-armazenamento (bronze-prata)
Contexto disponível	Máximo: estado do ativo, calibração, condição da rede e da operação em tempo real	Parcial: contrato e esquema, sem o estado do processo	Mínimo: apenas o dado e os metadados que sobreviveram
Latência da detecção	Segundos: defeito detectado no instante em que nasce	Minutos: detectado na admissão	Semanas a meses: detectado quando a análise falha
Custo por defeito tratado	Menor: tratamento único, automatizado, por fluxo	Médio: rejeição e reprocessamento por lote	Maior: horas de especialista por projeto, repetidas a cada consumidor
Reversibilidade	Plena: a origem ainda pode reparar o instrumento, a rede ou a configuração	Parcial: o defeito já ocorreu; evita-se apenas sua entrada	Baixa: o contexto dissipou-se; a correção é inferência
Risco de perda de dados	Baixo, se as leituras forem marcadas e preservadas, e não descartadas	Médio: rejeições em lote podem suprimir eventos reais	Baixo para o acervo, alto para a confiabilidade do que se usa

Fonte: elaboração própria com base em Teh, Kempa-Liehr e Wang (2020), Breck *et al.* (2019), Schelter *et al.* (2018) e Armbrust *et al.* (2021).

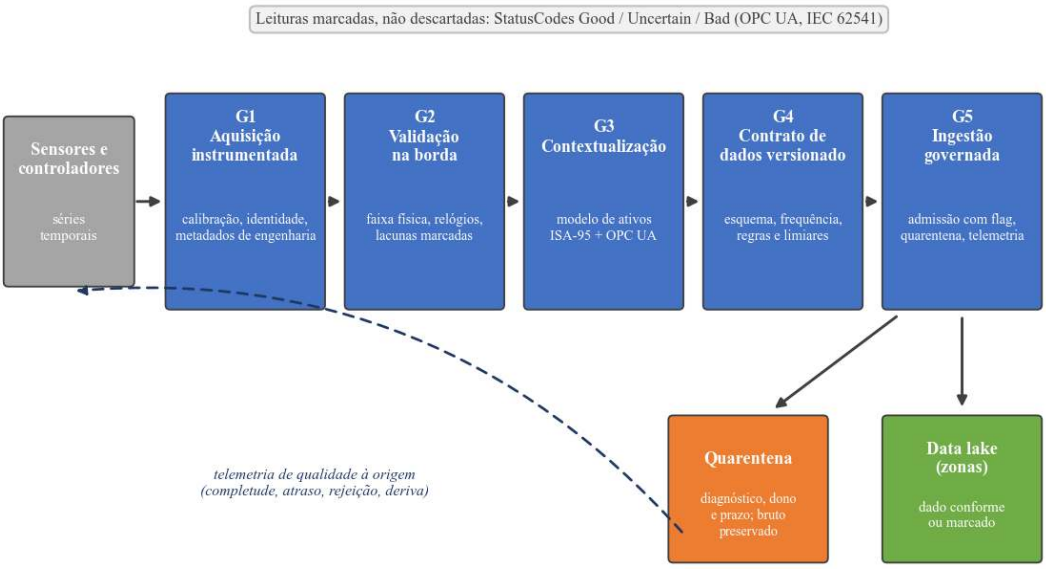
3.4 O FRAMEWORK PREVENTIVO: CINCO DATA QUALITY GATES

A proposta central deste artigo desloca a fronteira da qualidade para antes do armazenamento, organizando-a em cinco data quality gates encadeados, G1 a G5, que a Figura 1 situa entre o sensor e o lake e o Quadro 3 detalha com seus controles, papéis e métricas. O gate G1 é a aquisição instrumentada: o dado nasce qualificado quando o sensor tem calibração vigente, identidade única e metadados de engenharia (faixa, unidade, incerteza, localização no ativo) registrados desde a instalação. O gate G2 é a validação na borda: no gateway ou no dispositivo de edge computing, cada leitura passa por verificações sintáticas e de faixa física, os relógios são disciplinados por sincronização de tempo e as falhas de rede são absorvidas por buffers com marcação explícita de lacuna, de modo que a ausência vire informação, e não silêncio. O gate G3 é a contextualização: as séries são vinculadas ao modelo de ativos da planta, estruturado segundo a hierarquia da ISA-95 e com semântica padronizada pela modelagem de informação do OPC UA, o que resolve na origem a heterogeneidade de unidades e nomes e carrega o contexto operacional que a análise futura exigirá (IEC, 2013; 2020). O gate G4 é o contrato de dados: um artefato versionado, pactuado entre a operação produtora e os consumidores analíticos, que declara esquema, frequência, latência, regras de qualidade e limiares de

marcação, convertendo expectativas difusas em interface verificável, com proprietário e curador designados na lógica da alocação de decisões, no mesmo espírito da validação contínua de dados que a engenharia de aprendizado de máquina adotou em produção (Khatri; Brown, 2010; Breck *et al.*, 2019). O gate G5 é a ingestão governada: a admissão verifica cada fluxo contra o contrato, com a verificação automatizada de restrições de qualidade em escala que a literatura de sistemas demonstrou viável (Schelter *et al.*, 2018), e devolve à origem a telemetria de qualidade (completude, atraso, taxa de marcação, deriva estimada), fechando o laço de melhoria contínua no padrão do modelo de processos da ISO 8000 (ISO, 2016).

Uma decisão de desenho merece destaque, porque corrige o risco mais sério de qualquer gate: o descarte indevido. Em manutenção preditiva, um pico aparentemente espúrio pode ser o primeiro sinal de uma falha real, exatamente o evento que o modelo existe para capturar. Por isso, nos gates G2 e G5, as leituras que violam regras não são descartadas: são marcadas e preservadas. O dado bruto segue para o acervo com sua marca de qualidade, a leitura suspeita vai à zona de quarentena com diagnóstico, dono e prazo, e o consumidor decide, com a marca à vista, o que usar. O OPC UA oferece esse mecanismo nativamente, pelos StatusCodes que acompanham cada valor (Good, Uncertain e Bad), de modo que a marcação de qualidade viaja com a própria leitura desde a borda, sem artefatos paralelos (IEC, 2020). Descartar é decisão de consumo, tomada a jusante e reversível; marcar é decisão de qualidade, tomada na origem e definitiva.

Figura 1. O framework preventivo: do sensor ao data lake pelos cinco data quality gates



Fonte: elaboração própria.

Quadro 3. Os cinco data quality gates: controles, papéis e métricas

Gate	Controles principais	Papel responsável	Métricas
G1. Aquisição instrumentada	Calibração vigente, identidade única do ponto, metadados de engenharia desde a instalação	Engenharia de instrumentação	Vigência de calibração e cobertura de metadados
G2. Validação na borda	Checagens sintáticas e de faixa física, sincronização de relógios, buffer com marcação de lacunas, StatusCodes de qualidade	Engenharia de automação	Taxa de marcação na borda e atraso de transmissão
G3. Contextualização	Vínculo ao modelo de ativos (ISA-95), semântica OPC UA, unidades padronizadas e linhagem	Curadoria de dados industriais	Cobertura do modelo de ativos e consistência semântica
G4. Contrato de dados	Esquema, frequência, latência e regras de qualidade versionados e pactuados	Proprietário do dado e consumidores	Conformidade ao contrato e mudanças versionadas
G5. Ingestão governada	Admissão com verificação automatizada, marcação preservando o bruto, quarentena com diagnóstico e telemetria de qualidade à origem	Plataforma de dados	Completude, quarentena resolvida e tempo de detecção de anomalia de dados

Fonte: elaboração própria com base em Khatri e Brown (2010), DAMA International (2017), ISO (2016), IEC (2013; 2020), Breck *et al.* (2019) e Schelter *et al.* (2018).

Um exemplo ilustrativo torna o encadeamento concreto. Considere o sinal de vibração de uma bomba centrífuga crítica. Em G1, o acelerômetro entra em operação com calibração registrada, identidade única e metadados de faixa e incerteza; sem isso, nenhuma leitura futura é julgável. Em G2, o gateway valida cada amostra contra a faixa física do instrumento, sincroniza o carimbo de tempo e, quando a rede oscila, registra a lacuna como lacuna, em vez de silenciá-la; um pico abrupto recebe StatusCode Uncertain, mas segue adiante. Em G3, o sinal é vinculado ao ativo bomba P-101 no modelo ISA-95, com a unidade padronizada e o contexto operacional (em carga, em partida, em manutenção) anexado pela semântica OPC UA. Em G4, o contrato declara que a série chega a 10 kHz, com latência máxima pactuada e regra de marcação para variação nula prolongada, o sintoma do instrumento travado. Em G5, a admissão verifica o contrato, grava o fluxo com suas marcas, envia o trecho do pico Uncertain à quarentena para diagnóstico e devolve à engenharia a telemetria do período. Semanas depois, quando o modelo de manutenção preditiva consumir essa série, o pico estará lá, marcado: se o diagnóstico da quarentena confirmou interferência elétrica, o modelo o ignora com fundamento; se confirmou evento mecânico real, ele é exatamente o sinal precoce que evita a quebra. Em nenhum dos dois cenários o dado sumiu ou entrou travestido de normal, e essa é a diferença entre o gate e o filtro.

A eficácia do desenho está no pareamento entre cada defeito típico e o gate que o intercepta, sistematizado no Quadro 4, a contribuição mais operacional deste artigo. A lógica é a da prevenção posicionada: cada gate captura a classe de defeito para a qual detém o contexto necessário, e o que

escapa de um gate encontra o seguinte. A deriva de sensor, invisível para regras de faixa, é combatida pela disciplina de calibração de G1 e monitorada pela telemetria de longo prazo de G5. As lacunas e os atrasos são absorvidos e explicitados em G2, onde a rede ainda pode ser diagnosticada. A heterogeneidade semântica morre em G3, único gate em que o modelo de ativos é conhecido. As violações de esquema e de regra pactuada são barradas pelo contrato de G4, e a quarentena de G5 garante que nenhum defeito residual entre no acervo confiável sem marca, diagnóstico e destino.

Quadro 4. Matriz de interceptação: defeitos típicos e gates preventivos correspondentes

Defeito típico	Gate e controle que o intercepta	Consequência evitada a jusante
Deriva de sensor e descalibração	G1 (calibração vigente) e G5 (telemetria de deriva)	Modelos treinados sobre medições enviesadas
Lacunas de amostragem	G2 (buffer com marcação explícita de lacuna)	Agregações distorcidas e interpolações silenciosas
Carimbos de tempo inconsistentes	G2 (sincronização de relógios e validação de ordenação)	Séries desordenadas e correlações espúrias entre sinais
Valores congelados e picos espúrios	G2 (regras de variação e faixa física, com marcação) e G5 (quarentena diagnóstica)	Falsa estabilidade e alarmes analíticos indevidos
Unidades e nomenclaturas divergentes	G3 (modelo de ativos ISA-95 e semântica OPC UA)	Comparações inválidas entre ativos e plantas
Mudanças não comunicadas na fonte	G4 (contrato de dados versionado)	Quebra silenciosa de painéis e modelos em produção

Fonte: elaboração própria com base nas fontes citadas nas seções 3.2 e 3.4.

3.5 O IMPACTO A JUSANTE: O QUE OS DEFEITOS FAZEM COM A MANUTENÇÃO PREDITIVA

A manutenção preditiva é o caso de uso que melhor expõe por que o garbage in, garbage out não é uma metáfora, e sim um mecanismo. Os modelos que estimam a saúde de um ativo aprendem a associar padrões do sinal a estados do equipamento, e cada defeito de dado não interceptado entra nessa associação como se fosse física. A deriva de sensor é o exemplo mais traiçoeiro: o aumento lento e monotônico da leitura de vibração ou temperatura causado pela descalibração é, para o modelo, indistinguível da degradação real do ativo, e produz o falso positivo de falha, com parada desnecessária, troca de componente são e erosão da confiança da operação no sistema. O valor congelado opera no sentido inverso: o instrumento travado alimenta o modelo com estabilidade fictícia enquanto o ativo degrada, produzindo o falso negativo que culmina na quebra não anunciada, exatamente o evento que o investimento em manutenção preditiva existia para evitar. As lacunas interpoladas silenciosamente na limpeza a jusante criam padrões artificiais no conjunto de treinamento, suavidades e tendências que nunca existiram no processo e que o modelo aprende como

comportamento normal. E os carimbos de tempo desalinhados entre sinais quebram as correlações usadas como variáveis preditoras: a relação entre a corrente do motor e a vibração do mancal, deslocada em segundos, transforma uma assinatura diagnóstica em ruído (Teh; Kempa-Liehr; Wang, 2020). A literatura de sistemas de aprendizado de máquina nomeia a raiz do fenômeno: as dependências de dados são o maior componente do débito técnico oculto desses sistemas, mais instáveis e menos testadas que o código (Sculley *et al.*, 2015), e os danos compostos das cascatas de dados são relatados pelos próprios praticantes como a causa dominante de retrabalho e fracasso em aplicações de alto risco (Sambasivan *et al.*, 2021). A validação contínua contra expectativas declaradas, resposta da engenharia de produção a esse quadro, é precisamente o que os gates G4 e G5 trazem para dentro da tubulação industrial, antes do armazenamento (Breck *et al.*, 2019; Schelter *et al.*, 2018). Em síntese: cada linha da matriz do Quadro 4 corresponde, na manutenção preditiva, a um modo de falha analítica específico, e interceptar o defeito no gate certo é a forma mais barata de evitar o alarme falso, a quebra surda e o modelo que aprendeu o que nunca aconteceu.

4 RESULTADOS E DISCUSSÃO

O primeiro resultado da análise é a reformulação do problema que o framework materializa: qualidade de dados industriais não é uma etapa analítica, é uma propriedade de engenharia da tubulação que liga o sensor ao lake, e deve ser projetada com a mesma disciplina com que se projetam a instrumentação e a rede. Essa reformulação inverte a economia do esforço documentada no Gráfico 1: em vez de cada consumidor pagar repetidamente, em horas de cientista de dados, o custo variável da limpeza, a organização paga uma vez o custo fixo da validação posicionada na origem, e o paga onde ele é menor, porque o contexto necessário para julgar o dado (calibração, estado do ativo, condição da rede) existe apenas ali, como o Quadro 2 sistematiza (CrowdFlower, 2016; Wang; Strong, 1996). Os corpos normativos dão a essa inversão o suporte institucional: a ISO 8000 fornece o modelo de processos para operá-la como melhoria contínua, e a ISO/IEC 38505 a eleva ao plano da governança, com a direção da organização respondendo pelo dado como ativo (ISO, 2016; ISO/IEC, 2017).

O segundo resultado é a complementaridade, e não a rivalidade, entre o framework e as arquiteturas de lake existentes. A organização em zonas e o lakehouse resolvem o problema do refinamento dentro do repositório, e o framework resolve o problema da admissão: com os cinco gates operantes, a zona bronze deixa de ser o depósito do defeito e passa a ser o registro auditável de um fluxo já contratado e marcado, a transformação para a prata torna-se conformação leve em vez de arqueologia, e a quarentena substitui a limpeza heroica por um processo com dono, prazo e diagnóstico (Armbrust *et al.*, 2021; Databricks, 2026). O contrato de dados é a peça que faz a junção: ao declarar

esquema, frequência e regras verificáveis, ele dá ao lake um critério de admissão e dá à origem uma especificação de saída, criando a interface estável que os modelos de alocação de decisões preconizam, com proprietários e curadores respondendo por domínios definidos, na linha do que a literatura de governança corporativa de dados descreve como controles mínimos, agora especializados para o regime temporal da indústria e equipados com a verificação automatizada que a engenharia de dados em escala demonstrou (Khatri; Brown, 2010; Souza, 2022; Schelter *et al.*, 2018).

O terceiro resultado ordena as condições de implantação e os limites. A implantação recomendada é incremental e orientada por valor: começa pelos fluxos que alimentam as aplicações mais sensíveis (manutenção preditiva, balanços de processo, relatórios regulatórios), instrumenta primeiro os gates G2 e G4, validação na borda e contrato, que, por hipótese derivada da análise do Quadro 2, tendem a oferecer a maior interceptação por unidade de esforço, e expande o modelo de ativos de G3 conforme a cobertura avança. Os limites são reais e devem ser declarados. O framework não elimina a necessidade de limpeza a jusante para o legado já armazenado, não substitui a manutenção da instrumentação, da qual G1 depende, e impõe custos de latência e processamento na borda que precisam ser dimensionados para cada classe de fluxo. Há também um limite organizacional: contratos de dados exigem que a operação assuma a produção de dados como responsabilidade formal, mudança cultural que a experiência de governança em grandes ecossistemas mostra ser o verdadeiro caminho crítico, mais do que a tecnologia (Madureira *et al.*, 2026). O retorno esperado, porém, tem as ordens de grandeza a seu favor: contra o custo recorrente de US$ 12,9 milhões anuais que a má qualidade impõe em média a cada organização, a prevenção posicionada disputa exatamente a fração evitável desse desperdício, e o faz uma única vez por fluxo, e não uma vez por projeto (Gartner, 2021; Redman, 2016).

5 LIMITAÇÕES E LACUNAS CIENTÍFICAS

Os limites deste estudo decorrem de seu desenho propositivo. O framework foi derivado analiticamente da literatura e dos corpos normativos, sem validação empírica própria em planta industrial, de modo que nenhuma métrica de desempenho (taxa de interceptação, redução de retrabalho ou retorno financeiro) é afirmada; a priorização dos gates G2 e G4 na implantação é hipótese analítica, não resultado medido. Os valores econômicos citados dimensionam o problema geral da má qualidade de dados e não constituem estimativa dos ganhos do framework; além disso, parte dessas evidências provém de literatura cinzenta com interesse comercial no tema, o estudo da datasfera da IDC foi patrocinado pela Seagate e a pesquisa de alocação de tempo é da CrowdFlower, fornecedora de serviços de preparação de dados, razão pela qual foram citadas pela fonte primária, sinalizadas como

estimativas institucionais e usadas apenas para dimensionar ordem de grandeza, nunca como evidência de eficácia. Não há estatística brasileira oficial equivalente sobre custos de qualidade de dados industriais, ausência que é, em si, uma lacuna. A proposta prioriza séries temporais de processo e não cobre integralmente dados de eventos discretos, imagens industriais e registros de manutenção textuais, que possuem defeitos e controles próprios. Entre as lacunas científicas abertas estão: a validação empírica do framework em estudos de caso multissetoriais, com medição de taxas de interceptação por gate e de custos evitados; o desenvolvimento de métricas padronizadas de deriva e de qualidade de carimbo de tempo para telemetria de longo prazo; a investigação de contratos de dados para fluxos de alta frequência, incluindo sua verificação automatizada na borda; o estudo dos arranjos organizacionais que sustentam a curadoria de dados industriais na fronteira entre operação e tecnologia da informação; e a mensuração, em contexto brasileiro, dos custos da má qualidade de dados na indústria, hoje inferidos de estimativas internacionais.

6 CONCLUSÃO

As duas perguntas postas na introdução encontraram resposta. Quanto aos fundamentos: as dimensões clássicas de qualidade (acurácia, completude, consistência, atualidade, validade e unicidade) traduzem-se para as séries temporais industriais em uma taxonomia própria de defeitos (deriva de sensor, lacunas de amostragem, carimbos de tempo inconsistentes, valores congelados, picos espúrios e heterogeneidade semântica), e os corpos normativos DAMA-DMBOK, ISO 8000 e ISO/IEC 38505, articulados ao modelo de alocação de decisões, à hierarquia de ativos da ISA-95 e à semântica do OPC UA, fornecem o arcabouço institucional para tratá-los como responsabilidade de governança, e não como acidente analítico. Quanto ao desenho: o framework proposto organiza essa responsabilidade em cinco data quality gates encadeados entre o sensor e o lake (G1, aquisição instrumentada; G2, validação na borda; G3, contextualização; G4, contrato de dados; G5, ingestão governada), nos quais a leitura suspeita é marcada e preservada, nunca descartada em silêncio, e tem na matriz de interceptação do Quadro 4 a sua contribuição mais operacional: o vínculo explícito entre cada defeito, o gate que o intercepta com contexto e o modo de falha analítica que deixa de alcançar a manutenção preditiva. A lição de fundo é a que a indústria aprendeu há um século e a engenharia de dados reaprende agora: qualidade se constrói no processo, não se inspeciona no produto final, e o dado, como qualquer produto, deve sair da linha conforme à especificação ou sair marcado, com diagnóstico e destino.

A agenda que o estudo deixa pode ser formulada como hipóteses testáveis. H1: em plantas comparáveis, fluxos operados com os gates G2 e G4 apresentam taxa de defeitos residuais no acervo

significativamente menor que fluxos ingeridos diretamente, com a diferença concentrada em lacunas e violações de contrato. H2: a fração do tempo de preparação de dados nos projetos analíticos decresce com a cobertura dos gates, aproximando a alocação de esforço documentada no Gráfico 1 da sua imagem invertida. H3: em modelos de manutenção preditiva, a marcação de qualidade na origem reduz simultaneamente falsos positivos atribuíveis a deriva e falsos negativos atribuíveis a valores congelados, em comparação com a limpeza a jusante do mesmo dado. H4: o custo total de qualidade por fluxo, somadas prevenção e remediação, é menor no regime preventivo a partir de um limiar de reuso analítico da série, limiar que a pesquisa empírica deve estimar. Submeter essas hipóteses a estudos de caso multissetoriais, com métricas padronizadas de interceptação por gate, é o passo que transforma a proposta deste artigo em conhecimento medido, e o chão de fábrica da confiança, em linha de produção auditável.

REFERÊNCIAS

ARMBRUST, M.; GHODSI, A.; XIN, R.; ZAHARIA, M. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In: CONFERENCE ON INNOVATIVE DATA SYSTEMS RESEARCH (CIDR), 11., 2021. Proceedings [...]. Chaminade: CIDR, 2021. Disponível em: https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf. Acesso em: 15 out. 2026.

BRECK, Eric; POLYZOTIS, Neoklis; ROY, Sudip; WHANG, Steven Euijong; ZINKEVICH, Martin. Data validation for machine learning. In: Proceedings of Machine Learning and Systems (MLSys 2019), v. 1, p. 334-347, 2019. Disponível em: https://proceedings.mlsys.org/paper_files/paper/2019/hash/928f1160e52192e3e0017fb63ab65391-Abstract.html. Acesso em: 15 out. 2026.

CROWDFLOWER. 2016 Data Science Report. San Francisco: CrowdFlower, 2016. Disponível em: https://visit.figure-eight.com/data-science-report. Acesso em: 15 out. 2026 (via cópia do Internet Archive da página oficial).

DAMA INTERNATIONAL. DAMA-DMBOK: Data Management Body of Knowledge. 2. ed. Basking Ridge: Technics Publications, 2017.

DATABRICKS. What is the medallion lakehouse architecture?. Databricks Documentation, 2026. Disponível em: https://docs.databricks.com/aws/en/lakehouse/medallion. Acesso em: 15 out. 2026.

GARTNER. How to improve your data quality. Stamford: Gartner, 2021. Disponível em: https://www.gartner.com/smarterwithgartner/how-to-improve-your-data-quality. Acesso em: 15 out. 2026.

HAI, R.; GEISLER, S.; QUIX, C. Constance: an intelligent data lake system. In: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA (SIGMOD), 2016, San Francisco. Proceedings [...]. New York: ACM, 2016. p. 2097-2100. DOI: https://doi.org/10.1145/2882903.2899389.

IEC. IEC 62264-1: Enterprise-control system integration - Part 1: Models and terminology. 2. ed. Geneva: International Electrotechnical Commission, 2013. Disponível em: https://webstore.iec.ch/en/publication/6675. Acesso em: 15 out. 2026.

IEC. IEC 62541: OPC Unified Architecture. Geneva: International Electrotechnical Commission, 2020.

ISO. ISO 8000-61:2016. Data quality - Part 61: Data quality management: process reference model. Geneva: International Organization for Standardization, 2016.

ISO/IEC. ISO/IEC 38505-1:2017. Information technology - Governance of IT - Governance of data - Part 1: Application of ISO/IEC 38500 to the governance of data. Geneva: ISO/IEC, 2017.

KHATRI, V.; BROWN, C. V. Designing data governance. Communications of the ACM, New York, v. 53, n. 1, p. 148-152, 2010. DOI: https://doi.org/10.1145/1629175.1629210.

MADUREIRA, A. S.; RIBEIRO, A. V. M.; CASTRO, D. C.; RIBEIRO, G. da S.; SILVA, L. de F. B.; MILHOMEM, M. J. C. de S.; GUIMARÃES, V. M. A governança de dados em saúde e os desafios da

interoperabilidade nas políticas públicas: o caso do SUS Digital. Aracê, São José dos Pinhais, v. 8, n. 3, e12574, 2026. DOI: https://doi.org/10.56238/arev8n3-089.

PRESS, G. Cleaning big data: most time-consuming, least enjoyable data science task, survey says. Forbes, Jersey City, 23 mar. 2016. Disponível em: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/. Acesso em: 15 out. 2026.

REDMAN, T. C. Bad data costs the U.S. $3 trillion per year. Harvard Business Review, Boston, 22 set. 2016. Disponível em: https://hbr.org/2016/09/bad-data-costs-the-us-3-trillion-per-year. Acesso em: 15 out. 2026.

REINSEL, D.; GANTZ, J.; RYDNING, J. The digitization of the world: from edge to core. IDC White Paper US44413318, patrocinado pela Seagate. Framingham: International Data Corporation, 2018. Disponível em: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. Acesso em: 15 out. 2026.

SAMBASIVAN, Nithya; KAPANIA, Shivani; HIGHFILL, Hannah; AKRONG, Diana; PARITOSH, Praveen; AROYO, Lora M. "Everyone wants to do the model work, not the data work": data cascades in high-stakes AI. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York: ACM, 2021. p. 1-15. DOI: https://doi.org/10.1145/3411764.3445518.

SCHELTER, Sebastian; LANGE, Dustin; SCHMIDT, Philipp; CELIKEL, Meltem; BIESSMANN, Felix; GRAFBERGER, Andreas. Automating large-scale data quality verification. Proceedings of the VLDB Endowment, v. 11, n. 12, p. 1781-1794, 2018. DOI: https://doi.org/10.14778/3229863.3229867.

SCULLEY, D.; HOLT, Gary; GOLOVIN, Daniel; DAVYDOV, Eugene; PHILLIPS, Todd; EBNER, Dietmar; CHAUDHARY, Vinay; YOUNG, Michael; CRESPO, Jean-François; DENNISON, Dan. Hidden technical debt in machine learning systems. In: Advances in Neural Information Processing Systems 28 (NIPS 2015). Montreal: Curran Associates, 2015. Disponível em: https://papers.nips.cc/paper_files/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html. Acesso em: 15 out. 2026.

SOUZA, E. dos S. Governança de dados em ambientes corporativos: modelos, papéis e controles mínimos. Lumen et Virtus, São José dos Pinhais, v. 12, n. 30, e12892, 2022. DOI: https://doi.org/10.56238/leved.esp.v12n30-008.

SOUZA, M. T. de; SILVA, M. D. da; CARVALHO, R. de. Revisão integrativa: o que é e como fazer. Einstein (São Paulo), São Paulo, v. 8, n. 1, p. 102-106, 2010. DOI: https://doi.org/10.1590/s1679-45082010rw1134.

TEH, H. Y.; KEMPA-LIEHR, A. W.; WANG, K. I-K. Sensor data quality: a systematic review. Journal of Big Data, v. 7, n. 1, 11, 2020. DOI: https://doi.org/10.1186/s40537-020-0285-1.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems, Abingdon, v. 12, n. 4, p. 5-33, 1996. DOI: https://doi.org/10.1080/07421222.1996.11518099.