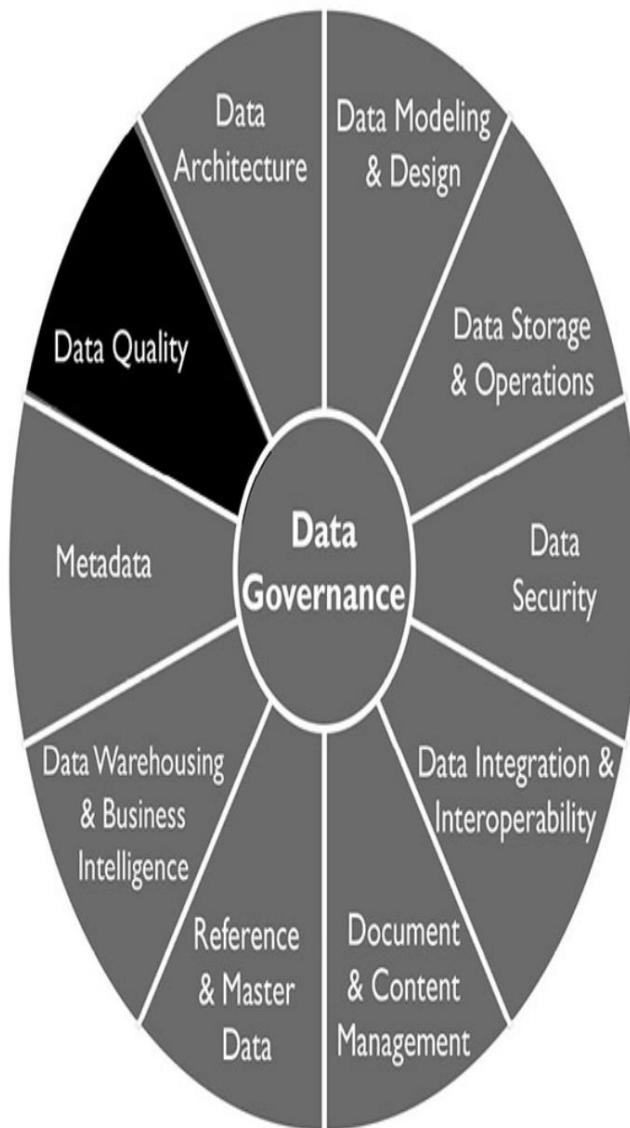


CAPÍTULO 13

Qualidade de dados



DAMA-DMBOK2 Data Management Framework

Copyright © 2017 by DAMA International

1. Introdução

A gestão eficaz de dados envolve um conjunto de processos complexos e inter-relacionados

processos que permitem que uma organização use seus dados para atingir objetivos estratégicos. O gerenciamento de dados inclui a capacidade de projetar dados para aplicativos, armazená-los e acessá-los com segurança, compartilhá-los adequadamente, aprender com eles e garantir que atendam às necessidades comerciais. Uma suposição subjacente às afirmações sobre o valor dos dados é que os dados em si são confiáveis e confiáveis. Em outras palavras, que são de alta qualidade.

No entanto, muitos fatores podem minar essa suposição ao contribuir para dados de baixa qualidade: Falta de entendimento sobre os efeitos de dados de baixa qualidade no sucesso organizacional, planejamento ruim, design de sistema "isolado", processos de desenvolvimento inconsistentes, documentação incompleta, falta de padrões ou falta de governança. Muitas organizações falham em definir o que torna os dados adequados para propósito.

Todas as disciplinas de gerenciamento de dados contribuem para a qualidade dos dados, e dados de alta qualidade que dão suporte à organização devem ser o objetivo de todas as disciplinas de gerenciamento de dados. Como decisões ou ações desinformadas por qualquer pessoa que interaja com dados podem resultar em dados de baixa qualidade, produzir dados de alta qualidade requer comprometimento e coordenação interfuncionais. Organizações e equipes devem estar cientes disso e devem planejar dados de alta qualidade, executando processos e projetos de maneiras que levem em conta o risco relacionado a condições inesperadas ou inaceitáveis nos dados.

Como nenhuma organização tem processos de negócios perfeitos, processos técnicos perfeitos ou práticas perfeitas de gerenciamento de dados, todas as organizações enfrentam problemas relacionados à qualidade de seus dados.

Organizações que gerenciam formalmente a qualidade dos dados têm menos problemas do que aquelas que deixam a qualidade dos dados ao acaso.

O gerenciamento formal da qualidade de dados é semelhante ao gerenciamento contínuo da qualidade para outros produtos. Ele inclui o gerenciamento de dados durante seu ciclo de vida, definindo padrões, incorporando qualidade aos processos que criam, transformam e armazenam dados e medindo dados em relação aos padrões. O gerenciamento de dados nesse nível geralmente requer uma equipe de programa de Qualidade de Dados. A equipe do programa de Qualidade de Dados é responsável por envolver profissionais de gerenciamento de dados comerciais e técnicos e conduzir o trabalho de aplicação de técnicas de gerenciamento de qualidade para

dados para garantir que os dados sejam adequados para consumo para uma variedade de propósitos. A equipe provavelmente estará envolvida com uma série de projetos por meio dos quais eles podem estabelecer processos e melhores práticas enquanto abordam problemas de dados de alta prioridade.

Como gerenciar a qualidade dos dados envolve gerenciar o ciclo de vida dos dados, um programa de Qualidade de Dados também terá responsabilidades operacionais relacionadas ao uso de dados. Por exemplo, relatar os níveis de qualidade dos dados e se envolver na análise, quantificação e priorização de problemas de dados. A equipe também é responsável por trabalhar com aqueles que precisam de dados para fazer seus trabalhos para garantir que os dados atendam às suas necessidades e trabalhar com aqueles que criam, atualizam ou excluem dados no curso de seus trabalhos para garantir que estejam lidando adequadamente com os dados. A qualidade dos dados depende de todos que interagem com os dados, não apenas dos profissionais de gerenciamento de dados.

Como é o caso com a Governança de Dados e com o gerenciamento de dados como um todo, o Gerenciamento de Qualidade de Dados é um programa, não um projeto. Ele incluirá trabalho de projeto e manutenção, juntamente com um compromisso com comunicações e treinamento. Mais importante, o sucesso a longo prazo do programa de melhoria da qualidade de dados depende de fazer com que uma organização mude sua cultura e adote uma mentalidade de qualidade. Conforme declarado no *Manifesto de Dados do Líder*: uma mudança fundamental e duradoura requer liderança comprometida e envolvimento de pessoas em todos os níveis de uma organização. Pessoas que usam dados para fazer seu trabalho — o que na maioria das organizações é uma porcentagem muito grande de funcionários — precisam impulsionar a mudança. Uma das mudanças mais críticas para se concentrar é como suas organizações gerenciam e melhoram a qualidade de seus dados.⁷¹

Figura 91 Diagrama de Contexto:

Qualidade de dados

Data Quality Management

Definition: The planning, implementation, and control of activities that apply quality management techniques to data, in order to assure it is fit for consumption and meets the needs of data consumers.

Goals:

1. Develop a governed approach to make data fit for purpose based on data consumers' requirements.
2. Define standards, requirements, and specifications for data quality controls as part of the data lifecycle.
3. Define and implement processes to measure, monitor, and report on data quality levels.
4. Identify and advocate for opportunities to improve the quality of data, through process and system improvements.



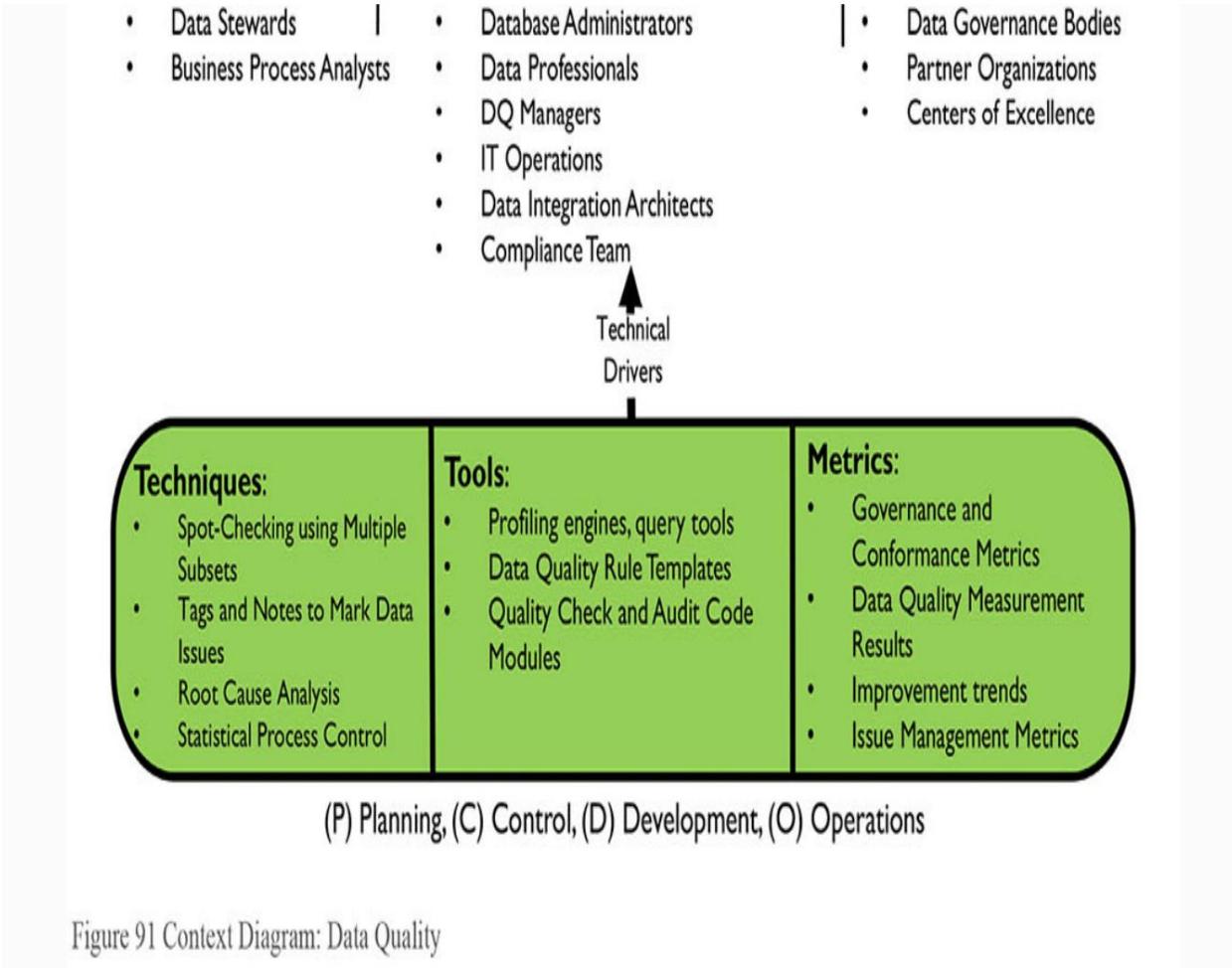


Figure 91 Context Diagram: Data Quality

1.1 Drivers de negócios

Os motivadores de negócios para o estabelecimento de uma Qualidade de Dados formal
O programa de gestão inclui:

- Aumentar o valor dos dados organizacionais e as oportunidades de usá-los
- Redução de riscos e custos associados a dados de baixa qualidade
- Melhorar a eficiência e a produtividade organizacional
- Proteger e melhorar a reputação da organização

As organizações que querem obter valor de seus dados reconhecem que dados de alta qualidade são mais valiosos do que dados de baixa qualidade. Má qualidade

os dados são carregados de risco ([veja o Capítulo 1](#)). Eles podem prejudicar a reputação de uma organização, resultando em multas, perda de receita, perda de clientes e exposição negativa na mídia. Os requisitos regulatórios geralmente exigem dados de alta qualidade. Além disso, muitos custos diretos estão associados a dados de baixa qualidade. Por exemplo,

- Incapacidade de faturar corretamente
- Aumento de chamadas de atendimento ao cliente e diminuição da capacidade de resolvê-las
- Perda de receita devido a oportunidades de negócios perdidas
- Atraso na integração durante fusões e aquisições
- Maior exposição à fraude
- Perda devido a más decisões empresariais motivadas por dados incorretos
- Perda de negócios devido à falta de boa reputação de crédito

Ainda assim, dados de alta qualidade não são um fim em si mesmos. São um meio para o sucesso organizacional. Dados confiáveis não apenas mitigam riscos e reduzem custos, mas também melhoram a eficiência. Os funcionários podem responder perguntas de forma mais rápida e consistente quando trabalham com dados confiáveis. Eles gastam menos tempo tentando descobrir se os dados estão corretos e mais tempo usando os dados para obter insights, tomar decisões e atender clientes.

1.2 Metas e princípios Os programas de qualidade de dados concentram-se nestas metas gerais:

- Desenvolver uma abordagem governada para tornar os dados adequados à finalidade com base nos requisitos dos consumidores de dados
- Definir padrões e especificações para controles de qualidade de dados como parte do ciclo de vida dos dados
- Definir e implementar processos para medir, monitorar e relatar os níveis de qualidade dos dados

- Identificar e defender oportunidades para melhorar a qualidade dos dados, por meio de mudanças em processos e sistemas e se envolver em atividades que melhorem mensuravelmente a qualidade dos dados com base nos requisitos do consumidor de dados

Os programas de qualidade de dados devem ser orientados pelos seguintes princípios:

- **Criticidade:** Um programa de qualidade de dados deve se concentrar nos dados mais críticos para a empresa e seus clientes.
As prioridades de melhoria devem basear-se na criticidade dos dados e no nível de risco caso os dados não sejam corretos.
- **Gerenciamento do ciclo de vida:** A qualidade dos dados deve ser gerenciada em todo o ciclo de vida dos dados, desde a criação ou aquisição até o descarte. Isso inclui gerenciar dados conforme eles se movem dentro e entre sistemas (ou seja, cada elo na cadeia de dados deve garantir que a saída de dados seja de alta qualidade).
- **Prevenção:** O foco de um programa de Qualidade de Dados deve ser prevenir erros de dados e condições que reduzam a usabilidade dos dados; ele não deve se concentrar apenas na correção de registros.
- **Remediação da causa raiz:** melhorar a qualidade dos dados vai além da correção de erros. Problemas com a qualidade dos dados devem ser compreendidos e tratados em suas causas raiz, em vez de apenas seus sintomas. Como essas causas geralmente estão relacionadas ao design do processo ou do sistema, melhorar a qualidade dos dados geralmente requer mudanças nos processos e nos sistemas que os suportam.
- **Governança:** As atividades de governança de dados devem dar suporte ao desenvolvimento de dados de alta qualidade, e as atividades do programa de qualidade de dados devem dar suporte e sustentar um ambiente de dados governado.
- **Orientado por padrões:** todas as partes interessadas no ciclo de vida dos dados

ter requisitos de qualidade de dados. Na medida do possível, esses requisitos devem ser definidos na forma de padrões e expectativas mensuráveis contra os quais a qualidade dos dados pode ser medida.

- **Medição objetiva e transparência:** os níveis de qualidade dos dados precisam ser medidos de forma objetiva e consistente. As medições e a metodologia de medição devem ser compartilhadas com as partes interessadas, pois elas são os árbitros da qualidade.
- **Incorporado em processos de negócios:** Os proprietários de processos de negócios são responsáveis pela qualidade dos dados produzidos por meio de seus processos. Eles devem impor padrões de qualidade de dados em seus processos.
- **Aplicado sistematicamente:** os proprietários do sistema devem aplicar sistematicamente os requisitos de qualidade de dados.
- **Conectado aos níveis de serviço:** relatórios de qualidade de dados e gerenciamento de problemas devem ser incorporados aos Acordos de Nível de Serviço (ANS).

1.3 Conceitos Essenciais

1.3.1 Qualidade de

dados O termo *qualidade de dados* se refere tanto às características associadas a dados de alta qualidade quanto aos processos usados para medir ou melhorar a qualidade dos dados. Esses usos duplos podem ser confusos, então ajuda separá-los e esclarecer o que constitui dados de alta qualidade.⁷² Os dados são de alta qualidade na medida em que atendem às expectativas e necessidades dos consumidores de dados. Ou seja, se os dados são adequados para os propósitos aos quais eles querem aplicá-los. São de baixa qualidade se não são adequados para esses propósitos. A qualidade dos dados depende, portanto, do contexto e das necessidades do consumidor de dados.

Um dos desafios na gestão da qualidade dos dados é que as expectativas relacionadas à qualidade nem sempre são conhecidas. Os clientes podem

não os articulem. Frequentemente, as pessoas que gerenciam dados nem perguntam sobre esses requisitos. No entanto, se os dados devem ser confiáveis e confiáveis, os profissionais de gerenciamento de dados precisam entender melhor os requisitos de qualidade de seus clientes e como medi-los. Isso precisa ser uma discussão contínua, pois os requisitos mudam ao longo do tempo conforme as necessidades do negócio e as forças externas evoluem.

1.3.2 Dados Críticos

A maioria das organizações tem muitos dados, e nem todos são de igual importância. Um princípio do Data Quality Management é concentrar esforços de melhoria em dados que são mais importantes para a organização e seus clientes. Fazer isso dá ao programa escopo e foco e permite que ele tenha um impacto direto e mensurável nas necessidades do negócio.

Embora os drivers específicos para criticidade sejam diferentes por setor, há características comuns entre as organizações. Os dados podem ser avaliados com base em se são exigidos por:

- Relatórios regulatórios
- Relatórios financeiros
- Política empresarial
- Operações em andamento
- Estratégia empresarial, especialmente esforços de diferenciação competitiva

Master Data é crítico por definição. Conjuntos de dados ou elementos de dados individuais podem ser avaliados quanto à criticidade com base nos processos que os consomem, na natureza dos relatórios em que aparecem ou no risco financeiro, regulatório ou de reputação para a organização se algo desse errado com os dados.⁷³

1.3.3 Dimensões de qualidade de dados

Uma dimensão de qualidade de dados é um recurso ou característica mensurável de

dados. O termo *dimensão* é usado para fazer a conexão com dimensões na medição de objetos físicos (por exemplo, comprimento, largura, altura). Dimensões de qualidade de dados fornecem um vocabulário para definir requisitos de qualidade de dados. A partir daí, elas podem ser usadas para definir resultados da avaliação inicial da qualidade de dados, bem como medições contínuas. Para medir a qualidade dos dados, uma organização precisa estabelecer características que sejam importantes para os processos de negócios (que valham a pena medir) e mensuráveis. Dimensões fornecem uma base para regras mensuráveis, que por si só devem ser diretamente conectadas a riscos potenciais em processos críticos.

Por exemplo, se os dados no campo de endereço de e-mail do cliente estiverem incompletos, não poderemos enviar informações do produto para nossos clientes por e-mail e perderemos vendas potenciais. Portanto, mediremos a porcentagem de clientes para os quais temos endereços de e-mail utilizáveis e melhoraremos nossos processos até que tenhamos um endereço de e-mail utilizável para pelo menos 98% de nossos clientes.

Muitos pensadores importantes em qualidade de dados publicaram conjuntos de dimensões.⁷⁴ Os três mais influentes são descritos aqui porque fornecem insights sobre como pensar sobre o que significa ter dados de alta qualidade, bem como sobre como a qualidade dos dados pode ser medida.

A estrutura Strong-Wang (1996) foca nas percepções dos consumidores de dados sobre os dados. Ela descreve 15 dimensões em quatro categorias gerais de qualidade de dados:

- DQ intrínseco
 - Precisão
 - Objetividade
 - Credibilidade
 - Reputação
- DQ contextual

- Valor agregado
- Relevância
- Pontualidade
- Completude
- Quantidade adequada de dados
- DQ representativo
 - Interpretabilidade
 - Facilidade de compreensão
 - Consistência representacional
 - Representação concisa
- Acessibilidade DQ
 - Acessibilidade
 - Segurança de acesso

Em *Data Quality for the Information Age* (1996), Thomas Redman formulou um conjunto de dimensões de qualidade de dados enraizadas na estrutura de dados.⁷⁵ Redman define um item de dados como um “triplo representável”: um valor do domínio de um atributo dentro de uma entidade. As dimensões podem ser associadas a qualquer uma das partes componentes dos dados – o modelo (entidades e atributos), bem como os valores. Redman inclui a dimensão de representação, que ele define como um conjunto de regras para registrar itens de dados. Dentro dessas três categorias gerais (modelo de dados, valores de dados, representação), ele descreve mais de duas dúzias de dimensões. Elas incluem o seguinte: Modelo de Dados:

- Conteúdo:
 - Relevância dos dados
 - A capacidade de obter os valores
 - Clareza das definições
- Nível de detalhe:
 - Granularidade de atributo
 - Precisão dos domínios de atributos
- Composição:
 - Naturalidade: A ideia de que cada atributo deve ter uma contrapartida simples no mundo real e que cada atributo deve ter relação com um único fato sobre a entidade
 - Identificabilidade: Cada entidade deve ser distingível de todas as outras entidades
 - Homogeneidade
 - Redundância mínima necessária
- Consistência:
 - Consistência semântica dos componentes do modelo
 - Consistência estrutural de atributos em todos os tipos de entidade
- Reação à mudança:

- Robustez
- Flexibilidade

Valores de dados:

- Precisão
- Completude
- Moeda
- Consistência

Representação:

- Adequação
- Interpretabilidade
- Portabilidade
- Precisão do formato
- Flexibilidade de formato
- Capacidade de representar valores nulos
- Uso eficiente do armazenamento
- Instâncias físicas de dados de acordo com seus formatos

Redman reconhece que a consistência de entidades, valores e representação pode ser entendida em termos de restrições. Diferentes tipos de consistência estão sujeitos a diferentes tipos de restrições.

Em *Improving Data Warehouse and Business Information Quality* (1999), Larry English apresenta um conjunto abrangente de dimensões divididas em duas grandes categorias: inherente e pragmática.⁷⁶ As características inherentes são independentes do uso dos dados. Características pragmáticas

estão associados à apresentação de dados e são dinâmicos; seu valor (qualidade) pode mudar dependendo dos usos dos dados.

- Características de qualidade **inherentes**

- Conformidade definicional
- Completude de valores
- Validade ou conformidade com a regra de negócios
- Precisão para uma fonte substituta
- Precisão com a realidade
- Precisão
- Não duplicação
- Equivalência de dados redundantes ou distribuídos
- Concorrência de dados redundantes ou distribuídos

- Características de qualidade **pragmáticas**

- Acessibilidade
- Pontualidade
- Clareza contextual
- Usabilidade
- Integridade de derivação
- Correção ou completude dos fatos

Em 2013, a DAMA UK produziu um white paper descrevendo seis dimensões principais da qualidade dos dados:

- **Completude:** A proporção de dados armazenados em relação ao potencial de 100%.

- **Exclusividade:** Nenhuma instância de entidade (coisa) será registrada mais de uma vez com base em como essa coisa é identificada.
- **Oportunidade:** O grau em que os dados representam a realidade a partir do momento necessário.
- **Validade:** Os dados são válidos se estiverem em conformidade com a sintaxe (formato, tipo, intervalo) de sua definição.
- **Precisão:** O grau em que os dados descrevem corretamente o objeto ou evento do "mundo real" que está sendo descrito.
- **Consistência:** Ausência de diferença ao comparar duas ou mais representações de uma coisa em relação a uma definição.

O white paper da DAMA UK também descreve outras características que têm impacto na qualidade. Embora o white paper não nomeie essas dimensões, elas funcionam de maneira semelhante ao DQ contextual e representacional de Strong e Wang e às características pragmáticas do inglês.

- **Usabilidade:** Os dados são compreensíveis, simples, relevantes, acessíveis, fáceis de manter e têm o nível certo de precisão?
- **Problemas de tempo** (além da pontualidade em si): é estável e ainda assim responde a solicitações de mudança legítimas?
- **Flexibilidade:** Os dados são comparáveis e compatíveis com outros dados? Eles têm agrupamentos e classificações úteis? Eles podem ser reaproveitados? Eles são fáceis de manipular?
- **Confiança:** Os processos de Governança de Dados, Proteção de Dados e Segurança de Dados estão em vigor? Qual é a reputação dos dados e eles são verificados ou verificáveis?
- **Valor:** Existe um bom caso de custo/benefício para os dados? Eles estão sendo usados de forma otimizada? Eles colocam em risco a segurança ou a privacidade das pessoas, ou as responsabilidades legais da empresa? Eles apoiam ou contradizem a imagem corporativa ou a identidade corporativa?

mensagem?

Embora não haja um único conjunto acordado de dimensões de qualidade de dados, essas formulações contêm ideias comuns. As dimensões incluem algumas características que podem ser medidas objetivamente (integridade, validade, conformidade de formato) e outras que dependem fortemente do contexto ou da interpretação subjetiva (usabilidade, confiabilidade, reputação). Quaisquer que sejam os nomes usados, as dimensões focam se há dados suficientes (integridade), se estão corretos (precisão, validade), quão bem eles se encaixam (consistência, integridade, exclusividade), se estão atualizados (oportunidade), acessíveis, utilizáveis e seguros.

A Tabela 29 contém definições de um conjunto de dimensões de qualidade de dados, sobre as quais há acordo geral, e descreve abordagens para medi-las.

Tabela 29 Dimensões comuns da qualidade dos dados

Table 29 Common Dimensions of Data Quality

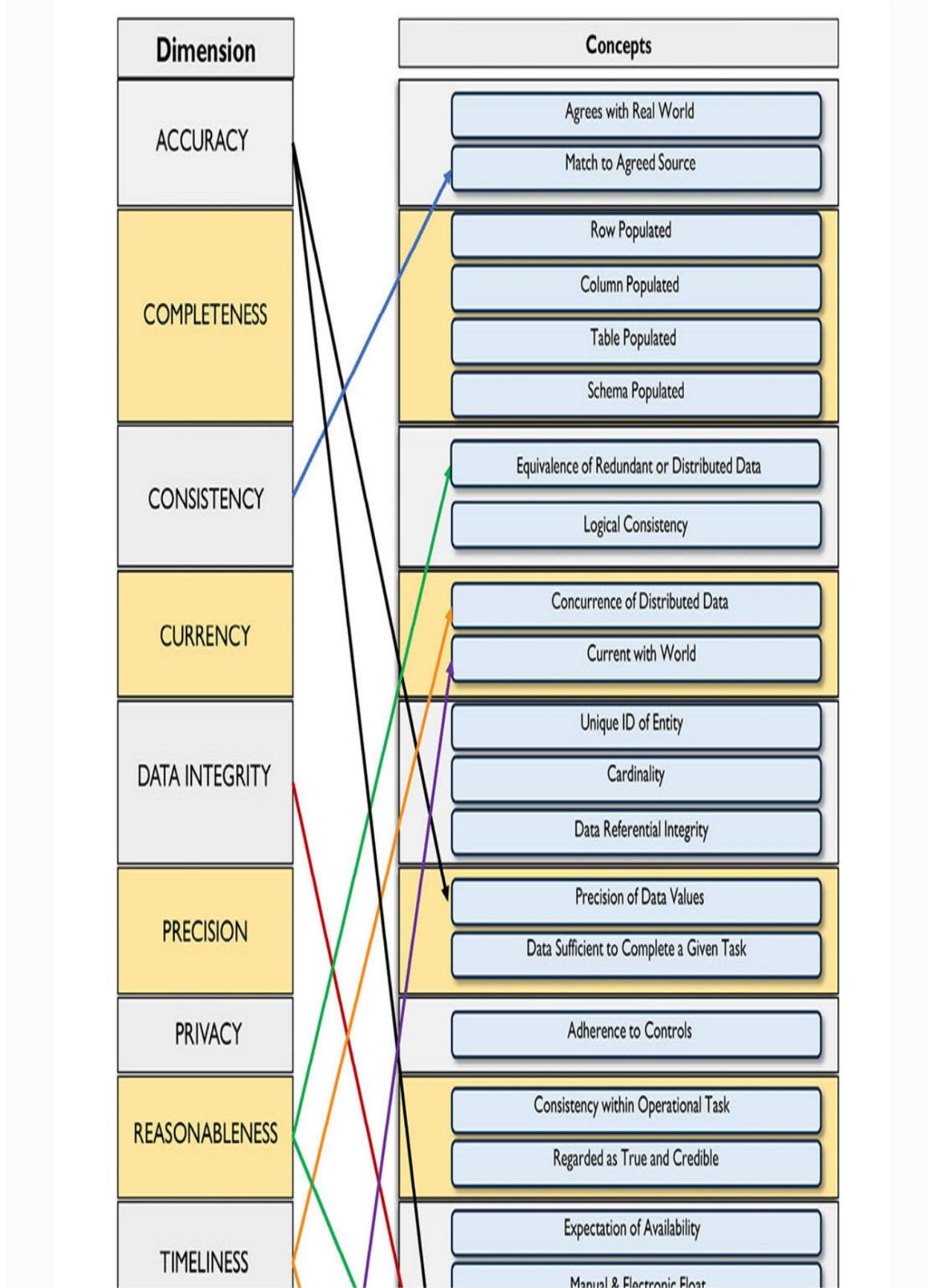
Dimension of Quality	Description
Accuracy	Accuracy refers to the degree that data correctly represents ‘real-life’ entities. Accuracy is difficult to measure, unless an organization can reproduce data collection or manually confirm accuracy of records. Most measures of accuracy rely on comparison to a data source that has been verified as accurate, such as a system of record or data from a reliable source (e.g., Dun and Bradstreet Reference Data).
Completeness	Completeness refers to whether all required data is present. Completeness can be measured at the data set, record, or column level. Does the data set contain all the records expected? Are records populated correctly? (Records with different statuses may have different expectations for completeness.) Are columns/attributes populated to the level expected? (Some columns are mandatory. Optional columns are populated only under specific conditions.) Assign completeness rules to a data set with varying levels of constraint: Mandatory attributes that require a value, data elements with conditional and optional values, and inapplicable attribute values. Data set level measurements may require comparison to a source of record or may be based on historical levels of population.
Consistency	Consistency can refer to ensuring that data values are consistently represented within a data set and between data sets, and consistently associated across data sets. It can also refer to the size and composition of data sets between systems or across time. Consistency may be defined between one set of attribute values and another attribute set within the same record (record-level consistency), between one set of attribute values and another attribute set in different records (cross-record consistency), or between one set of attribute values and the same attribute set within the same record at different points in time (temporal consistency). Consistency can also be used to refer to consistency of format. Take care not to confuse consistency with accuracy or correctness. Characteristics that are expected to be consistent within and across data sets can be used as the basis for standardizing data. Data Standardization refers to the conditioning of input data to ensure that data meets rules for content and format. Standardizing data enables more effective matching and facilitates consistent output. Encapsulate consistency constraints as a set of rules that specify consistent relationships between values of attributes, either across a record or message, or along all values of a single attribute (such as a range or list of valid values). For example, one might expect that the number of transactions each day does not exceed 105% of the running average number of transactions for the previous 30 days.

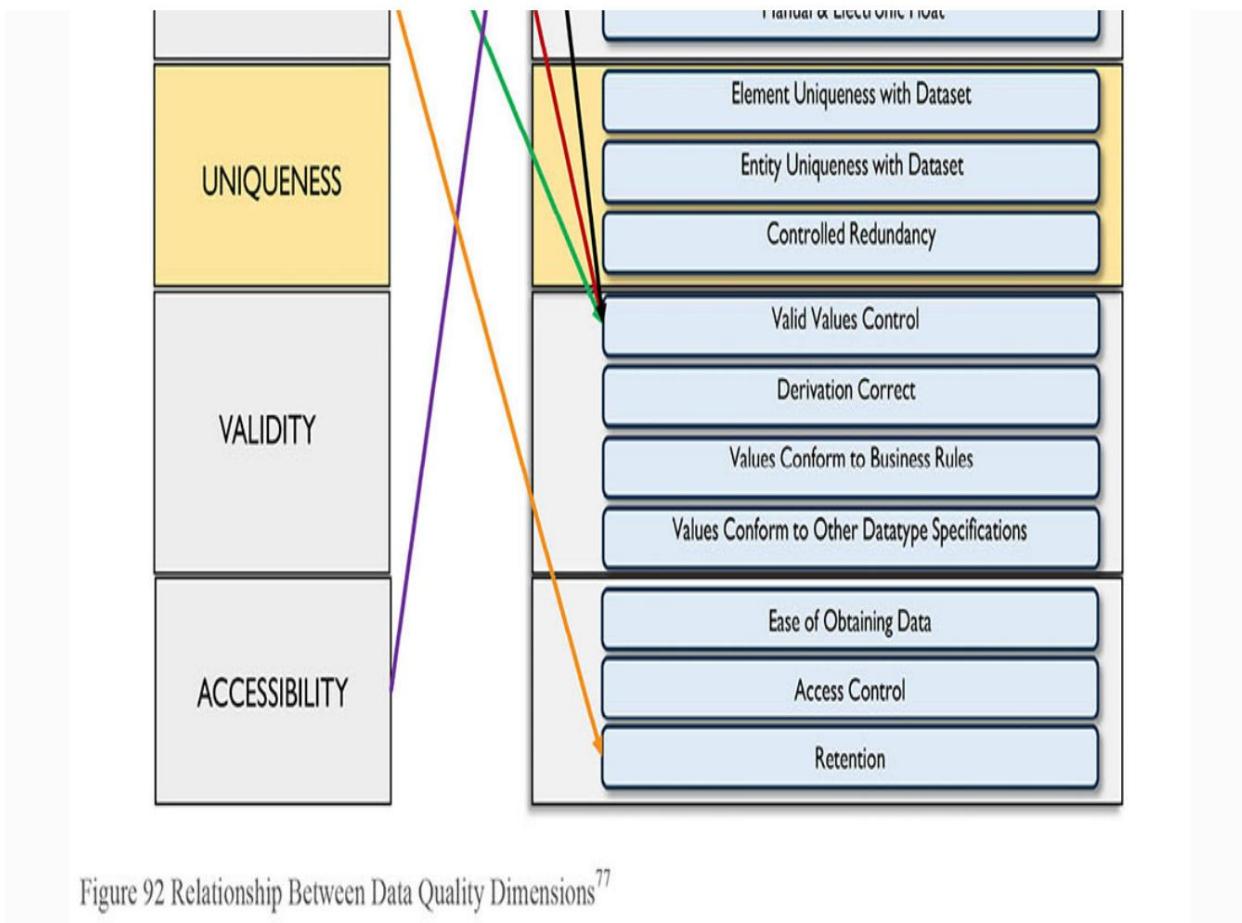
Integrity	Data Integrity (or Coherence) includes ideas associated with completeness, accuracy, and consistency. In data, integrity usually refers to either referential integrity (consistency between data objects via a reference key contained in both objects) or internal consistency within a data set such that there are no holes or missing parts. Data sets without integrity are seen as corrupted, or have data loss. Data sets without <i>referential</i> integrity have ‘orphans’ – invalid reference keys, or ‘duplicates’ – identical rows which may negatively affect aggregation functions. The level of orphan records can be measured as a raw count or as a percentage of the data set.
Reasonability	Reasonability asks whether a data pattern meets expectations. For example, whether a distribution of sales across a geographic area makes sense based on what is known about the customers in that area. Measurement of reasonability can take different forms. For example, reasonability may be based on comparison to benchmark data, or past instances of a similar data set (e.g., sales from the previous quarter). Some ideas about reasonability may be perceived as subjective. If this is the case, work with data consumers to articulate the basis of their expectations of data to formulate objective comparisons. Once benchmark measurements of reasonability are established, these can be used to objectively compare new instances of the same data set in order to detect change. (See Section 4.5.)
Timeliness	The concept of data Timeliness refers to several characteristics of data. Measures of timeliness need to be understood in terms of expected volatility – how frequently data is likely to change and for what reasons. Data currency is the measure of whether data values are the most up-to-date version of the information. Relatively static data, for example some Reference Data values like country codes, may remain current for a long period. Volatile data remains current for a short period. Some data, for example, stock prices on financial web pages, will often be shown with an as-of-time, so that data consumers understand the risk that the data has changed since it was recorded. During the day, while the markets are open, such data will be updated frequently. Once markets close, the data will remain unchanged, but will still be current, since the market itself is inactive. Latency measures the time between when the data was created and when it was made available for use. For example, overnight batch processing can give a latency of 1 day at 8am for data entered into the system during the prior day, but only one hour for data generated during the batch processing. (See Chapter 8.)
Uniqueness / Deduplication	Uniqueness states that no entity exists more than once within the data set. Asserting uniqueness of the entities within a data set implies that a key value relates to each unique entity, and only that specific entity, within the data set. Measure uniqueness by testing against key structure. (See Chapter 5.)
Validity	Validity refers to whether data values are consistent with a defined domain of values. A domain of values may be a defined set of valid values (such as in a reference table), a range of values, or value that can be determined via rules. The data type, format, and precision of expected values

must be accounted for in defining the domain. Data may also only be valid for a specific length of time, for example data that is generated from RFID (radio frequency ID) or some scientific data sets. Validate data by comparing it to domain constraints. Keep in mind that data may be valid (i.e., it may meet domain requirements) and still not be accurate or correctly associated with particular records.

A Figura 92 alinha as dimensões de qualidade de dados e os conceitos associados a essas dimensões. As setas indicam sobreposições significativas entre os conceitos e também demonstram que não há acordo sobre um conjunto específico. Por exemplo, a dimensão de precisão está associada a 'concorda com o mundo real' e 'corresponde à fonte acordada' e também aos conceitos associados à validade, como 'derivação correta'

Figura 92 Relação entre dimensões de qualidade de dados⁷⁷



Figure 92 Relationship Between Data Quality Dimensions⁷⁷

1.3.4 Qualidade de Dados e Metadados

Os metadados são essenciais para gerenciar a qualidade dos dados. A qualidade dos dados é baseada em quão bem eles atendem aos requisitos dos consumidores de dados. Os metadados definem o que os dados representam. Ter um processo robusto pelo qual os dados são definidos dá suporte à capacidade de uma organização de formalizar e documentar os padrões e requisitos pelos quais a qualidade dos dados pode ser medida. A qualidade dos dados é sobre atender às expectativas. Os metadados são um meio primário de esclarecer as expectativas.

Metadados bem gerenciados também podem dar suporte ao esforço de melhorar a qualidade dos dados. Um repositório de Metadados pode abrigar resultados de medições de qualidade de dados para que sejam compartilhados por toda a organização e a equipe de Qualidade de Dados possa trabalhar em direção ao consenso sobre prioridades e drivers para melhoria. ([Consulte o Capítulo 12.](#))

1.3.5 Qualidade de Dados Padrão ISO

ISO 8000, o padrão internacional para qualidade de dados, está sendo desenvolvido para permitir a troca de dados complexos em um formato neutro de aplicação. Na introdução ao padrão, a ISO afirma: "A capacidade de criar, coletar, armazenar, manter, transferir, processar e apresentar dados para dar suporte a processos de negócios de forma oportuna e econômica requer tanto uma compreensão das características dos dados que determinam sua qualidade, quanto uma capacidade de medir, gerenciar e relatar a qualidade dos dados."

A ISO 8000 define características que podem ser testadas por qualquer organização na cadeia de fornecimento de dados para determinar objetivamente a conformidade dos dados com a ISO 8000.⁷⁸

A primeira parte publicada da ISO 8000 (parte 110, publicada em 2008) focou na sintaxe, codificação semântica e conformidade com a especificação de dados do Master Data. Outras partes projetadas para o padrão incluem a parte 100 - Introdução, parte 120 - Proveniência, parte 130 - Precisão e parte 140 - Completude.⁷⁹ A ISO define dados de qualidade como

"dados portáteis que atendem aos requisitos declarados".⁸⁰ O padrão de qualidade de dados está relacionado ao trabalho geral da ISO sobre portabilidade e preservação de dados. Os dados são considerados "portáveis" se puderem ser separados de um aplicativo de software. Os dados que só podem ser usados ou lidos usando um aplicativo de software licenciado específico estão sujeitos aos termos da licença de software. Uma organização pode não ser capaz de usar os dados que criou, a menos que esses dados possam ser separados do software que foi usado para criá-los.

Para atender aos requisitos declarados, é necessário que esses requisitos sejam definidos de forma clara e inequívoca. A ISO 8000 é suportada pela ISO 22745, um padrão para definir e trocar Dados Mestres. A ISO 22745 define como as declarações de requisitos de dados devem ser construídas, fornece exemplos em XML e define um formato para a troca de dados codificados.⁸¹ A ISO 22745 cria dados portáveis ao rotular os dados usando um Open Technical Dictionary compatível com a ISO 22745, como o ECCMA Open Technical Dictionary (eOTD).

A intenção da ISO 8000 é ajudar as organizações a definir o que é e o que é

não são dados de qualidade, habilite-os a pedir dados de qualidade usando convenções padrão e verifique se receberam dados de qualidade usando esses mesmos padrões. Quando os padrões são seguidos, os requisitos podem ser confirmados por meio de um programa de computador.

O modelo de referência do processo de gestão da qualidade da informação e dos dados ISO 8000 - Parte 61 está em desenvolvimento.⁸² Esta norma descreverá a estrutura e a organização da gestão da qualidade dos dados, incluindo:

- Planejamento de qualidade de dados
- Controle de qualidade de dados
- Garantia de qualidade de dados
- Melhoria da qualidade dos dados

1.3.6 Ciclo de vida de melhoria da qualidade de dados

A maioria das abordagens para melhorar a qualidade de dados é baseada nas técnicas de melhoria da qualidade na fabricação de produtos físicos.⁸³ Neste paradigma, os dados são entendidos como o produto de um conjunto de processos. Em sua forma mais simples, um processo é definido como uma série de etapas que transforma entradas em saídas. Um processo que cria dados pode consistir em uma etapa (coleta de dados) ou muitas etapas: coleta de dados, integração em um data warehouse, agregação em um data mart, etc. Em qualquer etapa, os dados podem ser afetados negativamente. Eles podem ser coletados incorretamente, descartados ou duplicados entre sistemas, alinhados ou agregados incorretamente, etc. Melhorar a qualidade dos dados requer a capacidade de avaliar o relacionamento entre entradas e saídas, a fim de garantir que as entradas atendam aos requisitos do processo e que as saídas estejam em conformidade com as expectativas. Como as saídas de um processo se tornam entradas para outros processos, os requisitos devem ser definidos ao longo de toda a cadeia de dados.

Uma abordagem geral para melhoria da qualidade de dados, mostrada na Figura 93, é uma versão do ciclo Shewhart/Deming.⁸⁴ Baseado no método científico, o ciclo Shewhart/Deming é um modelo de solução de problemas conhecido como 'planejar-fazer-verificar-agir'. A melhoria vem por meio de um processo definido

conjunto de etapas. A condição dos dados deve ser medida em relação aos padrões e, se não atender aos padrões, a(s) causa(s) raiz da discrepância dos padrões deve(m) ser identificada(s) e remediada(s). As causas raiz podem ser encontradas em qualquer uma das etapas do processo, técnicas ou não técnicas. Uma vez remediados, os dados devem ser monitorados para garantir que continuem a atender aos requisitos.

Figura 93 O gráfico de Shewhart

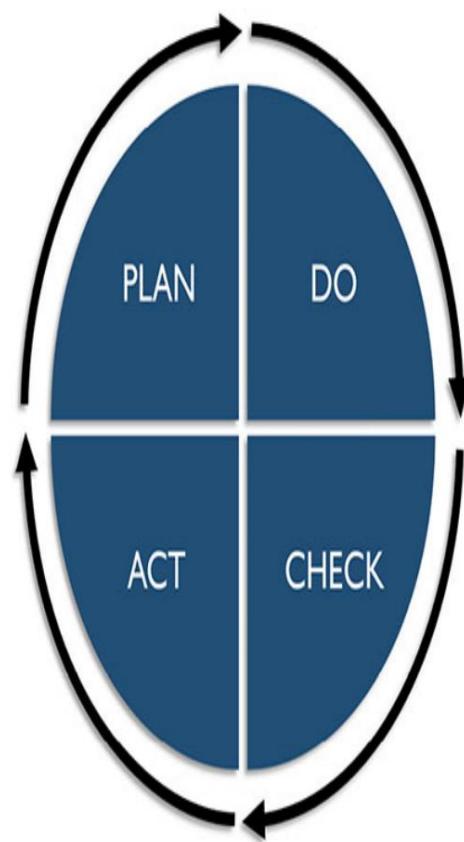


Figure 93 The Shewhart Chart

Para um determinado conjunto de dados, um ciclo de Gerenciamento de Qualidade de Dados começa identificando os dados que não atendem aos requisitos dos consumidores de dados e os problemas de dados que são obstáculos para a obtenção dos objetivos de negócios. Os dados precisam ser avaliados em relação às principais dimensões de qualidade e aos requisitos de negócios conhecidos. As causas raiz dos problemas precisarão ser identificadas para que as partes interessadas possam entender os custos de

remediação e os riscos de não remediar os problemas. Este trabalho é frequentemente feito em conjunto com Data Stewards e outras partes interessadas.

No estágio *de Plano*, a equipe de Qualidade de Dados avalia o escopo, o impacto e a prioridade de problemas conhecidos e avalia alternativas para resolvê-los. Este plano deve ser baseado em uma base sólida de análise das causas raiz dos problemas. A partir do conhecimento das causas e do impacto dos problemas, o custo/benefício pode ser compreendido, a prioridade pode ser determinada e um plano básico pode ser formulado para resolvê-los.

No estágio *Do*, a equipe DQ lidera esforços para abordar as causas raiz dos problemas e planejar o monitoramento contínuo dos dados. Para causas raiz que são baseadas em processos não técnicos, a equipe DQ pode trabalhar com os proprietários do processo para implementar mudanças. Para causas raiz que exigem mudanças técnicas, a equipe DQ deve trabalhar com equipes técnicas para garantir que os requisitos sejam implementados corretamente e que as mudanças técnicas não introduzam erros.

O estágio *Check* envolve o monitoramento ativo da qualidade dos dados, conforme medido em relação aos requisitos. Desde que os dados atendam aos limites definidos para qualidade, ações adicionais não são necessárias. Os processos serão considerados sob controle e atendendo aos requisitos de negócios. No entanto, se os dados caírem abaixo dos limites de qualidade aceitáveis, ações adicionais devem ser tomadas para elevá-los a níveis aceitáveis.

O estágio *Act* é para atividades que abordam e resolvem problemas emergentes de qualidade de dados. O ciclo reinicia, à medida que as causas dos problemas são avaliadas e soluções propostas. A melhoria contínua é alcançada ao iniciar um novo ciclo. Novos ciclos começam como:

- As medições existentes ficam abaixo dos limites
- Novos conjuntos de dados estão sob investigação
- Novos requisitos de qualidade de dados surgem para dados existentes conjuntos
- Mudanças nas regras, padrões ou expectativas de negócios

O custo de obter dados corretos na primeira vez é mais barato do que os custos de obter dados errados e corrigi-los mais tarde. Incorporar qualidade aos processos de gerenciamento de dados desde o início custa menos do que adaptá-los. Manter dados de alta qualidade durante todo o ciclo de vida dos dados é menos arriscado do que tentar melhorar a qualidade em um processo existente. Também cria um impacto muito menor na organização.

Estabelecer critérios para qualidade de dados no início de um processo ou construção de sistema é um sinal de uma Organização de Gerenciamento de Dados madura.

Fazer isso exige governança e disciplina, bem como colaboração interfuncional.

1.3.7 Tipos de Regras de Negócios de Qualidade

de Dados As regras de negócios descrevem como os negócios devem operar internamente, para serem bem-sucedidos e compatíveis com o mundo externo. As Regras de Negócios de Qualidade de Dados descrevem como os dados devem existir para serem úteis e utilizáveis dentro de uma organização. Essas regras podem ser alinhadas com dimensões de qualidade e usadas para descrever requisitos de qualidade de dados. Por exemplo, uma regra de negócios que todos os campos de código de estado devem estar em conformidade com as Abreviações de Estado dos EUA pode ser aplicada por listas de seleção de entrada de dados e pesquisas de integração de dados. O nível de registros válidos ou inválidos pode então ser medido.

Regras de negócios são comumente implementadas em software, ou usando modelos de documentos para entrada de dados. Alguns tipos comuns de regras de negócios simples são:

- **Conformidade definicional:** Confirme se o mesmo entendimento das definições de dados é implementado e usado corretamente em processos em toda a organização. A confirmação inclui acordo algorítmico em campos calculados, incluindo qualquer tempo, ou restrições locais, e regras de interdependência de rollup e status.
- **Presença de valor e integridade do registro:** regras que definem as condições sob as quais valores ausentes são aceitáveis ou inaceitáveis.

- **Conformidade de formato:** um ou mais padrões especificam valores atribuídos a um elemento de dados, como padrões para formatação de números de telefone.
- **Associação ao domínio de valor:** especifique que o valor atribuído a um elemento de dados seja incluído naqueles enumerados em um domínio de valor de dados definido, como Códigos Postais dos Estados Unidos de 2 caracteres para um campo ESTADO.
- **Conformidade de intervalo:** um valor atribuído a um elemento de dados deve estar dentro de um intervalo numérico, lexicográfico ou de tempo definido, como maior que 0 e menor que 100 para um intervalo numérico.
- **Conformidade de mapeamento:** Indica que o valor atribuído a um elemento de dados deve corresponder a um selecionado de um domínio de valor que mapeia para outro(s) domínio(s) de valor correspondente equivalente(s). O domínio de dados STATE novamente fornece um bom exemplo, já que os valores de State podem ser representados usando diferentes domínios de valor (códigos postais USPS, códigos FIPS de 2 dígitos, nomes completos), e esses tipos de regras validam que 'AL' e '01' mapeiam para 'Alabama'. **Regras de consistência:** Aserções condicionais que se referem à manutenção de um relacionamento entre dois (ou mais) atributos com base nos valores reais desses atributos. Por exemplo, validação de endereço onde os códigos postais correspondem a estados ou províncias específicos.
- **Verificação de precisão:** compare um valor de dados com um valor correspondente em um sistema de registro ou outra fonte verificada (por exemplo, dados de marketing adquiridos de um fornecedor) para verificar se os valores correspondem.
- **Verificação de exclusividade:** regras que especificam quais entidades devem ter uma representação única e se existe apenas um registro para cada objeto do mundo real representado.
- **Validação da tempestividade:** Regras que indicam as características associadas às expectativas de

acessibilidade e disponibilidade de dados.

Outros tipos de regras podem envolver funções de agregação aplicadas a conjuntos de instâncias de dados (consulte [a Seção 4.5](#)). Exemplos de verificações de agregação incluem:

- Valide a razoabilidade do número de registros em um arquivo.
Isso requer manter estatísticas ao longo do tempo para gerar tendências.
- Validar a razoabilidade de um valor médio calculado a partir de um conjunto de transações. Isso requer o estabelecimento de limites para comparação e pode ser baseado em estatísticas ao longo do tempo.
- Valide a variância esperada na contagem de transações em um período de tempo especificado. Isso requer manter estatísticas ao longo do tempo e usá-las para estabelecer limites.

1.3.8 Causas comuns de problemas de qualidade de

dados Problemas de qualidade de dados podem surgir em qualquer ponto do ciclo de vida dos dados, da criação ao descarte. Ao investigar as causas raiz, os analistas devem procurar culpados em potencial, como problemas com entrada de dados, processamento de dados, design do sistema e intervenção manual em processos automatizados. Muitos problemas terão múltiplas causas e fatores contribuintes (especialmente se as pessoas criaram maneiras de contorná-los).

Essas causas de problemas também implicam maneiras de preveni-los: por meio de melhorias no design da interface, testes de regras de qualidade de dados como parte do processamento, foco na qualidade de dados no design do sistema e controles rigorosos na intervenção manual em processos automatizados.

1.3.8.1 Problemas Causados pela Falta de Liderança Muitas

pessoas assumem que a maioria dos problemas de qualidade de dados são causados por erros de entrada de dados. Um entendimento mais sofisticado reconhece que lacunas ou execução deficiente de processos comerciais e técnicos causam muitos

mais problemas do que digitação errada. No entanto, o senso comum diz e a pesquisa indica que muitos problemas de qualidade de dados são causados por uma falta de comprometimento organizacional com dados de alta qualidade, o que por si só decorre de uma falta de liderança, na forma de governança e gestão.

Toda organização tem ativos de informação e dados que são valiosos para suas operações. De fato, as operações de toda organização dependem da capacidade de compartilhar informações. Apesar disso, poucas organizações gerenciam esses ativos com rigor. Na maioria das organizações, a disparidade de dados (diferenças na estrutura de dados, formato e uso de valores) é um problema maior do que apenas erros simples; pode ser um grande obstáculo à integração de dados. Uma das razões pelas quais os programas de administração de dados se concentram em definir termos e consolidar a linguagem em torno dos dados é porque esse é o ponto de partida para obter dados mais consistentes.

Muitos programas de governança e ativos de informação são movidos somente pela conformidade, em vez do valor potencial a ser derivado dos dados como um ativo. Uma falta de reconhecimento por parte da liderança significa uma falta de comprometimento dentro de uma organização para gerenciar dados como um ativo, incluindo gerenciar sua qualidade (Evans e Price, 2012). (Veja a [Figura 94](#).)

As barreiras à gestão eficaz da qualidade dos dados incluem:⁸⁵ ---

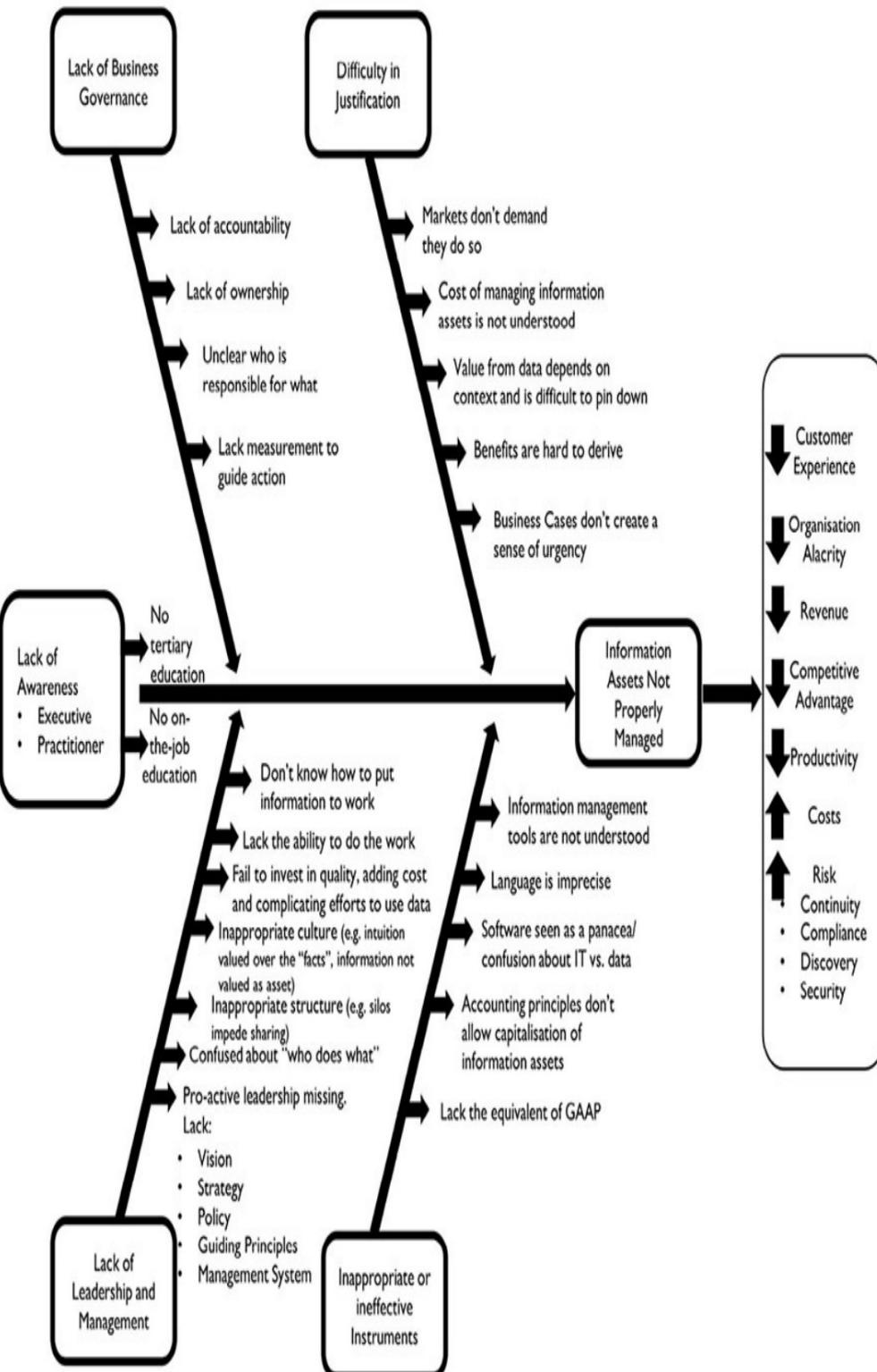
- Falta de conscientização por parte da liderança e da equipe
- Falta de governança empresarial
- Falta de liderança e gestão
- Dificuldade em justificar melhorias
- Instrumentos inadequados ou ineficazes para medir valor

Essas barreiras têm efeitos negativos na experiência do cliente, produtividade, moral, eficácia organizacional, receita e vantagem competitiva. Elas aumentam os custos de funcionamento da organização e também introduzem riscos. (Veja o [Capítulo 11](#).)

1.3.8.2 Problemas causados por processos de entrada de dados

- **Problemas de interface de entrada de dados:** Interfaces de entrada de dados mal projetadas podem contribuir para problemas de qualidade de dados. Se uma interface de entrada de dados não tiver edições ou controles para evitar que dados incorretos sejam colocados no sistema, os processadores de dados provavelmente tomarão atalhos, como pular campos não obrigatórios e deixar de atualizar campos padrão.
- **Posicionamento de entrada de lista:** até mesmo recursos simples de interfaces de entrada de dados, como a ordem dos valores em uma lista suspensa, podem contribuir para erros de entrada de dados.
- **Sobrecarga de campos:** Algumas organizações reutilizam campos ao longo do tempo para diferentes propósitos comerciais em vez de fazer alterações no modelo de dados e na interface do usuário. Essa prática resulta em preenchimento inconsistente e confuso dos campos.
- **Problemas de treinamento:** a falta de conhecimento do processo pode levar à entrada incorreta de dados, mesmo que controles e edições estejam em vigor. Se os processadores de dados não estiverem cientes do impacto de dados incorretos ou se forem incentivados pela velocidade, em vez da precisão, eles provavelmente farão escolhas com base em fatores diferentes da qualidade dos dados.

Figura 94 Barreiras à gestão da informação como um ativo empresarial⁸⁶



Barriers that slow/hinder/prevent
companies from managing their
information as a business asset

Most commonly observed root causes

Danette McGilvray / James Price / Tom Redman

October 2016

Work based on research by Dr. Nina Evans and James Price, see
"Barriers to the Effective Deployment of Information Assets" at
www.dataleaders.org

Figure 94 Barriers to Managing Information as a Business Asset⁸⁶

- **Alterações nos processos de negócios:** os processos de negócios mudam ao longo do tempo e, com essas alterações, novas regras de negócios e requisitos de qualidade de dados são introduzidos. No entanto, as mudanças nas regras de negócios nem sempre são incorporadas aos sistemas de forma oportuna ou abrangente. Erros de dados ocorrerão se uma interface não for atualizada para acomodar requisitos novos ou alterados. Além disso, os dados provavelmente serão impactados, a menos que as mudanças nas regras de negócios sejam propagadas por todo o sistema.
- **Execução inconsistente de processos de negócios:** Dados criados por meio de processos que são executados de forma inconsistente provavelmente serão inconsistentes. A execução inconsistente pode ser devido a problemas de treinamento ou documentação, bem como a requisitos de mudança.

1.3.8.3 Problemas causados por funções de processamento de dados

- **Suposições incorretas sobre fontes de dados:** Problemas de produção podem ocorrer devido a erros ou alterações, documentação inadequada ou obsoleta do sistema ou transferência inadequada de conhecimento (por exemplo, quando as PMEs saem sem

documentando seu conhecimento). Atividades de consolidação de sistemas, como aquelas associadas a fusões e aquisições, geralmente são baseadas em conhecimento limitado sobre o relacionamento entre sistemas. Quando vários sistemas de origem e feeds de dados precisam ser integrados, sempre há o risco de que detalhes sejam perdidos, especialmente com níveis variados de conhecimento de origem disponíveis e cronogramas apertados.

- **Regras de negócios obsoletas:** com o tempo, as regras de negócios mudam. Eles devem ser revisados e atualizados periodicamente. Se houver medição automatizada de regras, o processo técnico para medição de regras também deve ser atualizado. Se não for atualizado, problemas podem não ser identificados ou falsos positivos serão produzidos (ou ambos).
- **Estruturas de dados alteradas:** Os sistemas de origem podem alterar estruturas sem informar os consumidores downstream (humanos e do sistema) ou sem fornecer tempo suficiente para contabilizar as alterações. Isso pode resultar em valores inválidos ou outras condições que impedem a movimentação e o carregamento de dados, ou em alterações mais sutis que podem não ser detectadas imediatamente.

1.3.8.4 Problemas causados pelo design do sistema

- **Falha em impor integridade referencial:** A integridade referencial é necessária para garantir dados de alta qualidade em um nível de aplicativo ou sistema. Se a integridade referencial não for imposta ou se a validação for desligada (por exemplo, para melhorar os tempos de resposta), vários problemas de qualidade de dados podem surgir:
 - Dados duplicados que quebram regras de exclusividade
 - Linhas órfãs, que podem ser incluídas em algumas

relatórios e excluídos de outros, levando a valores múltiplos para o mesmo cálculo

- Incapacidade de atualização devido a requisitos de integridade referencial restaurados ou alterados
- Dados imprecisos devido à falta de dados sendo atribuídos valores padrão

- **Falha ao impor restrições de exclusividade:** várias cópias de instâncias de dados dentro de uma tabela ou arquivo devem conter instâncias exclusivas. Se houver verificações insuficientes para exclusividade de instâncias, ou se as restrições exclusivas forem desativadas no banco de dados para melhorar o desempenho, os resultados da agregação de dados podem ser exagerados.
- **Imprecisões e lacunas na codificação:** se o mapeamento ou layout dos dados estiver incorreto, ou se as regras para processamento dos dados não forem precisas, os dados processados terão problemas de qualidade, que vão desde cálculos incorretos até dados atribuídos ou vinculados a campos, chaves ou relacionamentos impróprios.
- **Imprecisões do modelo de dados:** se as suposições dentro do modelo de dados não forem suportadas pelos dados reais, haverá problemas de qualidade de dados que vão desde perda de dados devido a comprimentos de campo excedidos pelos dados reais até dados sendo atribuídos a IDs ou chaves impróprias.
- **Sobrecarga de campos:** a reutilização de campos ao longo do tempo para diferentes propósitos, em vez de alterar o modelo de dados ou o código, pode resultar em conjuntos confusos de valores, significado pouco claro e, potencialmente, problemas estruturais, como chaves atribuídas incorretamente.
- **Incompatibilidades de dados temporais:** na ausência de um dicionário de dados consolidado, vários sistemas podem implementar formatos de data ou horários diferentes, o que, por sua vez, leva à incompatibilidade de dados e à perda de dados quando os dados são armazenados.

a sincronização ocorre entre diferentes sistemas de origem.

- **Gerenciamento de dados mestres fraco:** o gerenciamento de dados mestres imaturo pode levar à escolha de fontes não confiáveis para dados, o que pode causar problemas de qualidade de dados que são muito difíceis de encontrar até que a suposição de que a fonte de dados é precisa seja refutada.
- **Duplicação de dados:** A duplicação desnecessária de dados é frequentemente resultado de um gerenciamento de dados ruim. Existem dois tipos principais de problemas de duplicação indesejáveis:
 - **Fonte Única – Várias Instâncias Locais:** Por exemplo, instâncias do mesmo cliente em várias tabelas (similares ou idênticas) no mesmo banco de dados. Saber qual instância é a mais precisa para uso pode ser difícil sem conhecimento específico do sistema.
 - **Várias fontes – Instância única:** instâncias de dados com várias fontes autoritativas ou sistemas de registro. Por exemplo, instâncias de clientes individuais vindas de vários sistemas de ponto de venda. Ao processar esses dados para uso, pode haver áreas de armazenamento temporário duplicadas. As regras de mesclagem determinam qual fonte tem prioridade sobre as outras ao processar em áreas de dados de produção permanentes.

1.3.8.5 Problemas Causados pela Correção de Problemas

Os patches de dados manuais são alterações feitas diretamente nos dados do banco de dados, não por meio das regras de negócios nas interfaces ou processamento do aplicativo. Esses são scripts ou comandos manuais geralmente criados às pressas e usados para "consertar" dados em uma emergência, como injeção intencional de dados ruins, falha de segurança, fraude interna ou

fonte externa de interrupção de negócios.

Como qualquer código não testado, eles têm um alto risco de causar mais erros por consequências não intencionais, alterando mais dados do que o necessário ou não propagando o patch para todos os dados históricos afetados pelo problema original. A maioria desses patches também altera os dados no local, em vez de preservar o estado anterior e adicionar linhas corrigidas.

Essas alterações geralmente NÃO são desfazíveis sem uma restauração completa do backup, pois há apenas o log do banco de dados para mostrar as alterações. Portanto, esses atalhos são fortemente desencorajados – eles são oportunidades para violações de segurança e interrupção de negócios por mais tempo do que uma correção adequada causaria. Todas as alterações devem passar por um processo de gerenciamento de alterações governado.

1.3.9 Criação de Perfil de

Dados Criação de Perfil de Dados é uma forma de análise de dados usada para inspecionar dados e avaliar a qualidade. A criação de perfil de dados usa técnicas estatísticas para descobrir a verdadeira estrutura, conteúdo e qualidade de uma coleção de dados (Olson, 2003). Um mecanismo de criação de perfil produz estatísticas que os analistas podem usar para identificar padrões no conteúdo e na estrutura dos dados. Por exemplo:

- **Contagens de nulos:** identifica a existência de nulos e permite a inspeção para saber se eles são permitidos ou não **Valor**
- **máximo/mínimo:** identifica valores atípicos, como negativos
- **Comprimento máximo/mínimo:** identifica valores atípicos ou inválidos para campos com requisitos de comprimento
- específicos **Distribuição de frequência** de valores para colunas individuais: permite a avaliação da razoabilidade (por exemplo, distribuição de códigos de país para transações, inspeção de valores que ocorrem com frequência ou pouca frequência, bem como a porcentagem de registros preenchidos com valores padrão)
- **Tipo e formato dos dados:** identifica o nível de não conformidade com os requisitos de formato, bem como a identificação de formatos inesperados (por exemplo, número de decimais, espaços incorporados, valores de amostra)

O perfil também inclui análise entre colunas, que pode identificar colunas sobrepostas ou duplicadas e expor dependências de valores incorporados. A análise entre tabelas explora conjuntos de valores sobrepostos e ajuda a identificar relacionamentos de chave estrangeira. A maioria das ferramentas de perfil de dados permite detalhar os dados analisados para investigação posterior.

Os resultados do mecanismo de criação de perfil devem ser avaliados por um analista para determinar se os dados estão em conformidade com as regras e outros requisitos. Um bom analista pode usar os resultados da criação de perfil para confirmar relacionamentos conhecidos e descobrir características e padrões ocultos dentro e entre conjuntos de dados, incluindo regras de negócios e restrições de validade. A criação de perfil geralmente é usada como parte da descoberta de dados para projetos (especialmente projetos de integração de dados; consulte o [Capítulo 8](#)) ou para avaliar o estado atual dos dados que são alvos de melhoria. Os resultados da criação de perfil de dados podem ser usados para identificar oportunidades de melhorar a qualidade dos dados e dos metadados (Olson, 2003; Maydanchik, 2007).

Embora a criação de perfil seja uma maneira eficaz de entender os dados, ela é apenas um primeiro passo para a melhoria da qualidade dos dados. Ela permite que as organizações identifiquem problemas potenciais. Resolver problemas requer outras formas de análise, incluindo análise de processos de negócios, análise de linhagem de dados e análise de dados mais profunda que pode ajudar a isolar as causas raiz dos problemas.

1.3.10 Qualidade de dados e processamento de dados

Embora o foco dos esforços de melhoria da qualidade de dados esteja frequentemente na prevenção de erros, a qualidade dos dados também pode ser melhorada por meio de algumas formas de processamento de dados. ([Consulte o Capítulo 8](#).)

1.3.10.1 Limpeza de Dados A

Limpeza ou *Depuração de Dados* transforma os dados para torná-los conformes aos padrões de dados e regras de domínio. A limpeza inclui detectar e corrigir erros de dados para levar a qualidade dos dados a um nível aceitável.

Custa dinheiro e introduz risco para remediar continuamente os dados por meio da limpeza. Idealmente, a necessidade de limpeza de dados deve diminuir ao longo do tempo, à medida que as causas raiz dos problemas de dados são resolvidas.

a necessidade de limpeza de dados pode ser abordada por:

- Implementação de controles para evitar erros de entrada de dados
- Corrigindo os dados no sistema de origem
- Melhorar os processos de negócios que criam os dados

Em algumas situações, pode ser necessário fazer correções contínuas, pois o reprocessamento dos dados em um sistema intermediário é mais barato do que qualquer outra alternativa.

1.3.10.2 Aprimoramento de dados

O aprimoramento ou enriquecimento de dados é o processo de adicionar atributos a um conjunto de dados para aumentar sua qualidade e usabilidade. Alguns aprimoramentos são obtidos pela integração de conjuntos de dados internos a uma organização. Dados externos também podem ser adquiridos para aprimorar dados organizacionais (consulte [o Capítulo 10](#)).

Exemplos de aprimoramento de dados incluem:

- **Carimbos de data/hora:** Uma maneira de melhorar os dados é documentar a hora e a data em que os itens de dados são criados, modificados ou aposentados, o que pode ajudar a rastrear eventos de dados históricos. Se forem detectados problemas com os dados, os carimbos de data/hora podem ser muito valiosos na análise da causa raiz, porque permitem que os analistas isolem o período de tempo do problema.
- **Dados de auditoria:** a auditoria pode documentar a linhagem de dados, o que é importante para o rastreamento histórico e também para a validação.
- **Vocabulários de referência:** terminologia, ontologias e glossários específicos de negócios melhoraram a compreensão e o controle, ao mesmo tempo em que trazem um contexto comercial personalizado.
- **Informações contextuais:** adicionar contexto, como localização, ambiente ou métodos de acesso, e marcar dados para revisão e análise.
- **Informações geográficas:** As informações geográficas podem ser aprimoradas por meio da padronização de endereços e geocodificação,

que inclui codificação regional, município, mapeamento de bairros, pares de latitude/longitude ou outros tipos de dados baseados em localização.

- **Informações demográficas:** Dados do cliente podem ser aprimorados por meio de informações demográficas, como idade, estado civil, gênero, renda ou codificação étnica. Dados de entidade empresarial podem ser associados à receita anual, número de funcionários, tamanho do espaço ocupado, etc.
- **Informações psicográficas:** dados usados para segmentar as populações-alvo por comportamentos, hábitos ou preferências específicas, como preferências por produtos e marcas, associações a organizações, atividades de lazer, estilo de transporte, preferências de horário de compras, etc.
- **Informações de avaliação:** use esse tipo de aprimoramento para avaliação de ativos, inventário e venda.

1.3.10.3 Análise e formatação de dados Análise

de dados é o processo de analisar dados usando regras pré-determinadas para definir seu conteúdo ou valor. A análise de dados permite que o analista de dados defina conjuntos de padrões que alimentam um mecanismo de regras usado para distinguir entre valores de dados válidos e inválidos. A correspondência de padrões específicos aciona ações.

A análise de dados atribui características aos valores de dados que aparecem em uma instância de dados, e essas características ajudam a determinar fontes potenciais para benefícios adicionais. Por exemplo, se um atributo chamado 'nome' puder ser determinado como tendo valores pertencentes a 'nome comercial' incorporados a ele, então o valor de dados é identificado como o nome de uma empresa em vez do nome de uma pessoa. Use a mesma abordagem para qualquer situação em que os valores de dados se organizem em hierarquias semânticas, como subpartes, partes e montagens.

Muitos problemas de qualidade de dados envolvem situações em que a variação nos valores de dados que representam conceitos semelhantes introduz ambiguidade. Extraia e reorganize os componentes separados (comumente chamados de

'tokens') podem ser extraídos e reorganizados em uma representação padrão para criar um padrão válido. Quando um padrão inválido é reconhecido, o aplicativo pode tentar transformar o valor inválido em um que atenda às regras. Realize a padronização mapeando dados de algum padrão de origem em uma representação de destino correspondente.

Por exemplo, considere as diferentes maneiras pelas quais os números de telefone esperados para estar em conformidade com um plano de numeração são formatados. Enquanto alguns têm dígitos, alguns têm caracteres alfabéticos e todos usam caracteres especiais diferentes para separação. As pessoas podem reconhecer cada um como um número de telefone. No entanto, para determinar se esses números são precisos (talvez comparando-os a um diretório mestre de clientes) ou para investigar se existem números duplicados quando deveria haver apenas um para cada fornecedor, os valores devem ser analisados em seus segmentos componentes (código de área, central e número de linha) e então transformados em um formato padrão.

Outro bom exemplo é o nome de um cliente, já que nomes podem ser representados em milhares de formas diferentes. Uma boa ferramenta de padronização será capaz de analisar os diferentes componentes de um nome de cliente, como nome próprio, nome do meio, sobrenome, iniciais, títulos, designações geracionais e, então, reorganizar esses componentes em uma representação canônica que outros serviços de dados serão capazes de manipular.

A capacidade humana de reconhecer padrões familiares contribui para uma capacidade de caracterizar valores de dados variantes pertencentes à mesma classe abstrata de valores; as pessoas reconhecem diferentes tipos de números de telefone porque eles estão em conformidade com padrões usados com frequência. Um analista descreve os padrões de formato que todos representam um objeto de dados, como **Nome da Pessoa, Descrição do Produto** e assim por diante. Uma ferramenta de qualidade de dados analisa valores de dados que estão em conformidade com qualquer um desses padrões e até os transforma em um único formato padronizado que simplificará os processos de avaliação, análise de similaridade e remediação. A análise sintática baseada em padrões pode automatizar o reconhecimento e a subsequente padronização de componentes de valor significativos

1.3.10.4 Transformação e Padronização de Dados

Durante o processamento normal, as regras de dados ação e transformam os dados em um formato que é legível pela arquitetura de destino. No entanto, legível nem sempre significa aceitável. As regras são criadas diretamente dentro de um fluxo de integração de dados ou dependem de tecnologias alternativas incorporadas ou acessíveis de dentro de uma ferramenta.

A transformação de dados se baseia nesses tipos de técnicas de padronização.

Oriente transformações baseadas em regras mapeando valores de dados em seus formatos e padrões originais em uma representação de destino.

Os componentes analisados de um padrão são submetidos a rearranjos, correções ou quaisquer alterações conforme orientado pelas regras na base de conhecimento.

Na verdade, a padronização é um caso especial de transformação, empregando regras que capturam contexto, linguística e expressões idiomáticas reconhecidas como comuns ao longo do tempo, por meio de análises repetidas pelo analista de regras ou fornecedor de ferramentas. ([Consulte o Capítulo 3.](#))

2. Atividades

2.1 Defina dados de alta qualidade

Muitas pessoas reconhecem dados de baixa qualidade quando os veem. Menos pessoas conseguem definir o que querem dizer com dados de alta qualidade. Alternativamente, eles os definem em termos muito gerais: "Os dados têm que estar certos."

"Precisamos de dados precisos." Dados de alta qualidade são adequados para os propósitos dos consumidores de dados. Antes de lançar um programa de Qualidade de Dados, é benéfico entender as necessidades do negócio, definir termos, identificar pontos problemáticos organizacionais e começar a construir consenso sobre os motivadores e prioridades para a melhoria da qualidade dos dados. Faça um conjunto de perguntas para entender o estado atual e avaliar a prontidão organizacional para a melhoria da qualidade dos dados:

- O que as partes interessadas querem dizer com "dados de alta qualidade"?
- Qual é o impacto de dados de baixa qualidade nas operações e estratégias de negócios?

- Como dados de maior qualidade viabilizarão a estratégia de negócios?
- Quais prioridades impulsionam a necessidade de melhoria da qualidade dos dados?
- Qual é a tolerância para dados de baixa qualidade?
- Que governança está em vigor para dar suporte à melhoria da qualidade dos dados?
- Quais estruturas de governança adicionais serão necessárias?

Obter uma visão abrangente do estado atual da qualidade dos dados em uma organização requer abordar a questão de diferentes perspectivas:

- Uma compreensão da estratégia e dos objetivos do negócio
- Entrevistas com as partes interessadas para identificar pontos problemáticos, riscos e motivadores de negócios
- Avaliação direta de dados, por meio de criação de perfis e outras formas de análise
- Documentação de dependências de dados em processos de negócios
- Documentação de arquitetura técnica e suporte de sistemas para processos de negócios

Esse tipo de avaliação pode revelar um número significativo de oportunidades. Elas precisam ser priorizadas com base no benefício potencial para a organização. Usando a contribuição de stakeholders, incluindo Data Stewards e SMEs de negócios e técnicas, a equipe de Data Quality deve definir o significado de qualidade de dados e propor prioridades de programa.

2.2 Defina uma estratégia de qualidade de dados Melhorar

a qualidade dos dados requer uma estratégia que considere o trabalho que precisa ser feito e a maneira como as pessoas o executarão. Qualidade de dados

as prioridades devem se alinhar com a estratégia de negócios. Adotar ou desenvolver uma estrutura e metodologia ajudará a orientar tanto a estratégia quanto as táticas, ao mesmo tempo em que fornece um meio de medir o progresso e os impactos. Uma estrutura deve incluir métodos para:

- Entenda e priorize as necessidades do negócio
- Identifique os dados essenciais para atender às necessidades do negócio
- Definir regras de negócios e padrões de qualidade de dados com base nos requisitos de negócios
- Avalie os dados em relação às expectativas
- Compartilhe descobertas e obtenha feedback das partes interessadas
- Priorizar e gerenciar problemas
- Identificar e priorizar oportunidades de melhoria
- Medir, monitorar e relatar a qualidade dos dados
- Gerenciar metadados produzidos por meio de processos de qualidade de dados
- Integrar controles de qualidade de dados em processos comerciais e técnicos

Uma estrutura também deve levar em conta como organizar a qualidade dos dados e como alavancar as ferramentas de qualidade dos dados. Conforme observado na introdução do capítulo, melhorar a qualidade dos dados requer que uma equipe do programa de Qualidade dos Dados envolva a equipe comercial e técnica e defina um programa de trabalho que aborde questões críticas, defina as melhores práticas e coloque em prática processos operacionais que suportem o gerenciamento contínuo da qualidade dos dados. Frequentemente, essa equipe fará parte da Organização de Gerenciamento de Dados. Os analistas de DQ precisarão trabalhar em estreita colaboração com os Administradores de Dados em todos os níveis. Eles também devem influenciar a política, incluindo a política sobre processos comerciais e desenvolvimento de sistemas. No entanto, essa equipe não será capaz de resolver todos os desafios de qualidade de dados de uma organização. O trabalho de DQ e um compromisso com dados de alta qualidade precisam ser incorporados às práticas organizacionais. A Estratégia de DQ deve levar em conta como estender as melhores práticas.

(Consulte

Capítulo 17.)

2.3 Identificar Dados Críticos e Regras de Negócios Nem todos os dados são de igual importância. Os esforços de Gerenciamento de Qualidade de Dados devem se concentrar primeiro nos dados mais importantes da organização: dados que, se fossem de maior qualidade, forneceriam maior valor para a organização e seus clientes. Os dados podem ser priorizados com base em fatores como requisitos regulatórios, valor financeiro e impacto direto sobre os clientes. Frequentemente, os esforços de melhoria da qualidade dos dados começam com os Dados Mestres, que estão, por definição, entre os dados mais importantes de qualquer organização. O resultado da análise de importância é uma lista classificada de dados, que a equipe de Qualidade de Dados pode usar para concentrar seus esforços de trabalho.

Tendo identificado os dados críticos, os analistas de Qualidade de Dados precisam identificar regras de negócios que descrevam ou impliquem expectativas sobre as características de qualidade dos dados. Frequentemente, as regras em si não são explicitamente documentadas. Elas podem precisar de engenharia reversa por meio da análise de processos de negócios existentes, fluxos de trabalho, regulamentações, políticas, padrões, edições de sistema, código de software, gatilhos e procedimentos, atribuição e uso de código de status e bom senso. Por exemplo, se uma empresa de marketing deseja direcionar esforços para pessoas em um grupo demográfico específico, então os índices potenciais de qualidade de dados podem ser o nível e a razoabilidade da população em campos demográficos como data de nascimento, idade, sexo e renda familiar.

A maioria das regras de negócios está associada a como os dados são coletados ou criados, mas a medição da qualidade dos dados gira em torno de se os dados são adequados para uso. Os dois (criação de dados e uso de dados) estão relacionados. As pessoas querem usar dados por causa do que eles representam e por que foram criados. Por exemplo, entender o desempenho de vendas de uma organização durante um trimestre específico ou ao longo do tempo depende de ter dados confiáveis sobre o processo de vendas (número e tipo de unidades vendidas, volume vendido a clientes existentes vs. novos clientes, etc.).

Não é possível conhecer todas as formas como os dados podem ser utilizados, mas é possível compreender o processo e as regras pelas quais os dados foram

criados ou coletados. As medições que descrevem se os dados são adequados para uso devem ser desenvolvidas em relação a usos conhecidos e regras mensuráveis com base em dimensões de qualidade de dados: completude, conformidade, validade, integridade, etc. que fornecem a base para métricas significativas.

As dimensões de qualidade permitem que os analistas caracterizem tanto as regras (o campo X é obrigatório e deve ser preenchido) quanto as descobertas (por exemplo, o campo não é preenchido em 3% dos registros; os dados estão apenas 97% completos).

No nível de campo ou coluna, as regras podem ser diretas.

As regras de completude são um reflexo de se um campo é obrigatório ou opcional e, se opcional, as condições sob as quais ele deve ser preenchido. As regras de validade dependem da estipulação do domínio de valores válidos e, em alguns casos, do relacionamento entre os campos. Por exemplo, um código postal dos EUA precisa ser válido, por si só, e corretamente associado a um código de estado dos EUA. As regras também devem ser definidas no nível do conjunto de dados. Por exemplo, todo cliente deve ter um endereço de correspondência válido.

Definir regras de qualidade de dados é desafiador porque a maioria das pessoas não está acostumada a pensar sobre dados em termos de regras. Pode ser necessário chegar às regras indiretamente, perguntando às partes interessadas sobre os requisitos de entrada e saída de um processo de negócios. Também ajuda perguntar sobre pontos problemáticos, o que acontece quando os dados estão faltando ou incorretos, como eles identificam problemas, como eles reconhecem dados ruins, etc. Tenha em mente que não é necessário conhecer todas as regras para avaliar os dados. A descoberta e o refinamento das regras são um processo contínuo. Uma das melhores maneiras de chegar às regras é compartilhar os resultados das avaliações. Esses resultados geralmente dão às partes interessadas uma nova perspectiva sobre os dados a partir da qual elas podem articular regras que lhes dizem o que precisam saber sobre os dados.

2.4 Realizar uma Avaliação Inicial da Qualidade dos Dados Uma vez que as necessidades comerciais mais críticas e os dados que as suportam tenham sido identificados, a parte mais importante da avaliação da qualidade dos dados é realmente olhar para esses dados, consultá-los para entender o conteúdo e os relacionamentos dos dados e comparar os dados reais com regras e expectativas. Na primeira vez que isso for feito, os analistas descobrirão muitas

coisas: relacionamentos e dependências não documentados dentro dos dados, regras implícitas, dados redundantes, dados contraditórios, etc., bem como dados que realmente estão em conformidade com as regras. Com a ajuda de administradores de dados, outras PMEs e consumidores de dados, os analistas de DQ precisarão classificar e priorizar as descobertas.

O objetivo de uma avaliação inicial da qualidade dos dados é aprender sobre os dados para definir um plano açãoável para melhoria. Geralmente é melhor começar com um esforço pequeno e focado – uma prova básica de conceito – para demonstrar como o processo de melhoria funciona. As etapas incluem:

- Defina os objetivos da avaliação; eles orientarão o trabalho
- Identificar os dados a serem avaliados; o foco deve estar em um pequeno conjunto de dados, mesmo um único elemento de dados, ou um problema específico de qualidade de dados
- Identificar os usos dos dados e os consumidores dos dados
- Identificar riscos conhecidos com os dados a serem avaliados, incluindo o impacto potencial de problemas de dados nos processos organizacionais
- Inspecione os dados com base em regras conhecidas e propostas
- Documentar níveis de não conformidade e tipos de problemas
- Realizar análises adicionais e aprofundadas com base nas descobertas iniciais para
 - Quantificar descobertas
 - Priorizar questões com base no impacto comercial
 - Desenvolver hipóteses sobre as causas raiz dos problemas de dados
- Reúna-se com administradores de dados, PMEs e consumidores de dados para confirmar problemas e prioridades

- Use as descobertas como base para o planejamento
 - Remediação de problemas, idealmente nas suas causas raiz
 - Controles e melhorias de processos para evitar que problemas ocorram novamente
 - Controles e relatórios contínuos

2.5 Identificar e priorizar melhorias potenciais Tendo provado que o processo de melhoria pode funcionar, o próximo objetivo é aplicá-lo estrategicamente. Fazer isso requer identificar e priorizar melhorias potenciais. A identificação pode ser realizada por perfis de dados em larga escala de conjuntos de dados maiores para entender a amplitude dos problemas existentes. Também pode ser realizada por outros meios, como entrevistar as partes interessadas sobre os problemas de dados que as afetam e acompanhar com a análise do impacto comercial desses problemas. Por fim, a priorização requer uma combinação de análise de dados e discussão com as partes interessadas.

As etapas para executar um perfil e análise de dados completos são essencialmente as mesmas da execução de uma avaliação em pequena escala: definir metas, entender os usos e riscos dos dados, medir em relação às regras, documentar e confirmar descobertas com PMEs, usar essas informações para priorizar esforços de remediação e melhoria. No entanto, às vezes há obstáculos técnicos para o perfil em grande escala. E o esforço precisará ser coordenado por uma equipe de analistas e os resultados gerais precisarão ser resumidos e compreendidos se um plano de ação eficaz for colocado em prática. Os esforços de perfil em larga escala, como aqueles em menor escala, ainda devem se concentrar nos dados mais críticos.

A criação de perfil de dados é apenas o primeiro passo na análise de problemas de qualidade de dados. Ela ajuda a identificar problemas, mas não identifica as causas raiz, nem determina o impacto dos problemas nos processos de negócios. Determinar o impacto requer a contribuição das partes interessadas ao longo da cadeia de dados. Ao planejar a criação de perfil em larga escala, garanta que o tempo seja alocado para compartilhar

resultados, priorizar problemas e determinar quais questões exigem análise aprofundada.

2.6 Definir metas para melhoria da qualidade de dados

O conhecimento obtido por meio das avaliações preliminares forma a base para metas específicas do programa de qualidade de dados. A melhoria pode assumir diferentes formas, desde a simples remediação (por exemplo, correção de erros em registros) até a remediação de causas raiz. Os planos de remediação e melhoria devem levar em conta acertos rápidos – problemas que podem ser resolvidos imediatamente a baixo custo – e mudanças estratégicas de longo prazo. O foco estratégico de tais planos deve ser resolver as causas raiz dos problemas e colocar em prática mecanismos para prevenir problemas em primeiro lugar.

Esteja ciente de que muitas coisas podem atrapalhar os esforços de melhoria: restrições do sistema, idade dos dados, trabalho de projeto em andamento que usa dados questionáveis, complexidade geral do cenário de dados, resistência cultural à mudança. Para evitar que essas restrições paralisem o programa, defina metas específicas e atingíveis com base na quantificação consistente do valor comercial das melhorias na qualidade dos dados.

Por exemplo, uma meta pode ser melhorar a integridade dos dados do cliente de 90% para 95% com base em melhorias de processo e edições do sistema. Obviamente, mostrar a melhoria envolverá comparar medições iniciais e resultados aprimorados. Mas o valor vem com os benefícios da melhoria: menos reclamações do cliente, menos tempo gasto corrigindo erros, etc. Meça essas coisas para explicar o valor do trabalho de melhoria. Ninguém se importa com os níveis de integridade do campo, a menos que haja um impacto comercial. Deve haver um retorno positivo sobre o investimento para melhorias nos dados. Quando problemas forem encontrados, determine o ROI das correções com base em:

- A criticidade (classificação de importância) dos dados afetados
- Quantidade de dados afetados
- A idade dos dados

- Número e tipo de processos de negócios impactados pelo problema
- Número de clientes, fornecedores ou funcionários afetados pelo problema
- Riscos associados ao problema
- Custos de remediação das causas raiz
- Custos de possíveis soluções alternativas

Ao avaliar problemas, especialmente aqueles em que as causas raiz são identificadas e mudanças técnicas são necessárias, sempre busque oportunidades para evitar que os problemas se repitam. Prevenir problemas geralmente custa menos do que corrigi-los – às vezes, ordens de magnitude menores. (Consulte o [Capítulo 11](#).)

2.7 Desenvolver e implantar operações de qualidade de dados Muitos programas de qualidade de dados começam por meio de um conjunto de projetos de melhoria identificados por meio dos resultados da avaliação de qualidade de dados. Para sustentar a qualidade dos dados, um programa de DQ deve colocar em prática um plano que permita à equipe gerenciar regras e padrões de qualidade de dados, monitorar a conformidade contínua dos dados com as regras, identificar e gerenciar problemas de qualidade de dados e relatar os níveis de qualidade. Em suporte a essas atividades, os analistas de DQ e os administradores de dados também estarão envolvidos em atividades como documentar padrões de dados e regras de negócios e estabelecer requisitos de qualidade de dados para fornecedores.

2.7.1 Gerenciar regras de qualidade de dados

O processo de criação de perfil e análise de dados ajudará uma organização a descobrir (ou fazer engenharia reversa) regras de negócios e qualidade de dados. À medida que a prática de qualidade de dados amadurece, a captura dessas regras deve ser incorporada ao processo de desenvolvimento e aprimoramento do sistema. Definir regras antecipadamente irá:

- Defina expectativas claras para as características de qualidade dos dados

- Fornecer requisitos para edições e controles do sistema que evitem a introdução de problemas de dados
- Fornecer requisitos de qualidade de dados a fornecedores e outras partes externas
- Crie a base para medição e relatórios contínuos da qualidade dos dados

Em resumo, regras e padrões de qualidade de dados são uma forma crítica de Metadados. Para serem eficazes, eles precisam ser gerenciados como Metadados.
As regras devem ser:

- **Documentado consistentemente:** Estabeleça padrões e modelos para documentar regras para que elas tenham um formato e significado consistentes.
- **Definido em termos de dimensões de Qualidade de Dados:** Dimensões de qualidade ajudam as pessoas a entender o que está sendo medido. A aplicação consistente de dimensões ajudará nos processos de medição e gerenciamento de problemas.
- **Vinculado ao impacto comercial:** embora as dimensões de qualidade de dados permitam a compreensão de problemas comuns, elas não são uma meta em si mesmas. Padrões e regras devem ser conectados diretamente ao seu impacto no sucesso organizacional. Medições que não estejam vinculadas a processos comerciais não devem ser tomadas.
- **Apoiado pela análise de dados:** Analistas de Qualidade de Dados não devem adivinhar regras. As regras devem ser testadas em relação a dados reais. Em muitos casos, as regras mostrarão que há problemas com os dados. Mas a análise também pode mostrar que as regras em si não estão completas.
- **Confirmado por PMEs:** O objetivo das regras é descrever como os dados devem parecer. Muitas vezes, é preciso conhecimento dos processos organizacionais para confirmar que as regras funcionam corretamente

descrever os dados. Esse conhecimento vem quando especialistas no assunto confirmam ou explicam os resultados da análise de dados.

- **Acessível a todos os consumidores de dados:** Todos os consumidores de dados devem ter acesso a regras documentadas. Esse acesso permite que eles entendam melhor os dados. Também ajuda a garantir que as regras estejam corretas e completas. Garanta que os consumidores tenham um meio de fazer perguntas e fornecer feedback sobre as regras.

2.7.2 Medir e monitorar a qualidade dos dados

Os procedimentos operacionais de gerenciamento da qualidade dos dados dependem da capacidade de medir e monitorar a qualidade dos dados. Há duas razões igualmente importantes para implementar a qualidade dos dados operacionais

Medidas:

- Para informar os consumidores de dados sobre os níveis de qualidade
- Para gerenciar o risco de que mudanças possam ser introduzidas por meio de alterações em processos comerciais ou técnicos

Algumas medições servem para ambos os propósitos. As medições devem ser desenvolvidas com base em descobertas da avaliação de dados e análise de causa raiz. As medições destinadas a informar os consumidores de dados se concentrarão em elementos e relacionamentos de dados críticos que, se não forem sólidos, impactarão diretamente os processos de negócios. As medições relacionadas ao gerenciamento de risco devem se concentrar em relacionamentos que deram errado no passado e podem dar errado no futuro. Por exemplo, se os dados forem derivados com base em um conjunto de regras ETL e essas regras puderem ser impactadas por mudanças nos processos de negócios, as medições devem ser colocadas em prática para detectar mudanças nos dados.

O conhecimento de problemas passados deve ser aplicado para gerenciar riscos. Por exemplo, se vários problemas de dados estiverem associados a problemas complexos

derivações, então todas as derivações devem ser avaliadas – mesmo aquelas que não foram associadas a problemas de dados. Na maioria dos casos, vale a pena implementar medições que monitorem funções semelhantes àquelas que tiveram problemas.

Os resultados da medição podem ser descritos em dois níveis: o detalhe relacionado à execução de regras individuais e os resultados gerais agregados das regras. Cada regra deve ter um índice padrão, alvo ou limite para comparação. Essa função geralmente reflete a porcentagem de dados corretos ou a porcentagem de exceções, dependendo da fórmula usada. Por exemplo:

$$\text{ValidDQI}(r) = \frac{(TestExecutions(r) - ExceptionsFound(r))}{TestExecutions(r)}$$

$$\text{InvalidDQI}(r) = \frac{(ExceptionsFound(r))}{TestExecutions(r)}$$

R representa a regra que está sendo testada. Por exemplo, 10.000 testes de uma regra de negócios (r) encontraram 560 exceções. Neste exemplo, o resultado ValidDQ seria $9440/10.000 = 94,4\%$, e o resultado Invalid DQ seria $560/10.000 = 5,6\%$.

Organizar as métricas e os resultados conforme mostrado na Tabela 30 pode ajudar a estruturar medidas, métricas e indicadores em todo o relatório, revelar possíveis rollups e aprimorar as comunicações. O relatório pode ser mais formalizado e vinculado a projetos que remediarão os problemas.

Relatórios filtrados são úteis para administradores de dados que buscam tendências e contribuições. A Tabela 30 fornece exemplos de regras construídas dessa maneira. Quando aplicável, os resultados das regras são expressos em porcentagens positivas (a parte dos dados que está em conformidade com as regras e expectativas) e porcentagens negativas (a parte dos dados que não está em conformidade com a regra).

As regras de qualidade de dados fornecem a base para o gerenciamento operacional da qualidade de dados. As regras podem ser integradas em serviços de aplicativos ou serviços de dados que complementam o ciclo de vida dos dados, seja por meio de

Ferramentas de qualidade de dados comerciais prontas para uso (COTS), mecanismos de regras e ferramentas de relatórios para monitoramento e relatórios, ou aplicativos desenvolvidos sob medida.

Tabela 30 Exemplos de métricas DQ

Table 30 DQ Metric Examples

Dimension and Business Rule	Measure	Metrics	Status Indicator
Completeness Business Rule 1: Population of field is mandatory	Count the number of records where data is populated, compare to the total number of records	Divide the obtained number of records where data is populated by the total number of records in the table or database and multiply it by 100 to get to percentage complete	Unacceptable: Below 80% populated Above 20% not populated
Example 1: Postal Code must be populated in the address table	Count populated: 700,000 Count not populated: 300,000 Total count: 1,000,000	Positive measure: $700,000/1,000,000 * 100 = 70\%$ populated Negative measure: $300,000/1,000,000 * 100 = 30\%$ not populated	Example result: Unacceptable
Uniqueness Business Rule 2: There should be only one record per entity instance in a table	Count the number of duplicate records identified; report on the percentage of records that represent duplicates	Divide the number of duplicate records by the total number of records in the table or database and multiply it by 100	Unacceptable: Above 0%
Example 2: There should be one and only one current row per postal code on the Postal Codes master list	Count of duplicates: 1,000 Total Count: 1,000,000	$1,000/1,000,000 * 100 = 1.0\%$ of postal codes are present on more than one current row	Example result: Unacceptable
Timeliness Business Rule 3: Records must arrive within a scheduled timeframe	Count the number of records failing to arrive on time from a data service for business transactions to be completed	Divide the number of incomplete transactions by the total number of attempted transactions in a time period and multiply by 100	Unacceptable: Below 99% completed on time Above 1% not completed on time

Example 3: Equity market record should arrive within 5 minutes of being transacted	Count of incomplete transactions: 2000 Count of attempted transactions: 1,000,000	Positive: $(1,000,000 - 2000) / 1,000,000 * 100 = 99.8\%$ 99.8% of transaction records arrived within defined timeframe Negative: $2000 / 1,000,000 * 100 = 0.20\%$ of transactions did not arrive within defined timeframe	Example Result: Acceptable
Validity Business Rule 4: If field X = value 1, then field Y must = value 1-prime	Count the number of records where the rule is met	Divide the number of records that meet the condition by the total number of records	Unacceptable : Below 100% adherence to the rule
Example 4: Only shipped orders should be billed	Count of records where status for shipping = Shipped and status for billing = Billed: 999,000 Count of total records: 1,000,000	Positive: $999,000 / 1,000,000 * 100 = 99.9\%$ of records conform to the rule Negative: $(1,000,000 - 999,000) / 1,000,000 * 100 = 0.10\%$ do not conform to the rule	Example Result: Unacceptable

Forneça monitoramento contínuo incorporando processos de controle e medição ao fluxo de processamento de informações.

O monitoramento automatizado da conformidade com as regras de qualidade de dados pode ser feito no fluxo ou por meio de um processo em lote. As medições podem ser feitas em três níveis de granularidade: o valor do elemento de dados, a instância ou registro de dados ou o conjunto de dados. A Tabela 31 descreve técnicas para coletar medições de qualidade de dados. As medições no fluxo podem ser feitas durante a criação de dados ou a transferência de dados entre os estágios de processamento. As consultas em lote podem ser realizadas em coleções de instâncias de dados reunidas em um conjunto de dados, geralmente em armazenamento persistente. As medições do conjunto de dados geralmente não podem ser feitas no fluxo, pois a medição pode precisar do conjunto inteiro.

A incorporação dos resultados dos processos de controlo e medição nos procedimentos operacionais e nas estruturas de relatórios permite a monitorização contínua dos níveis de qualidade dos dados para

feedback e melhoria nas atividades de geração/coleta de dados.

Tabela 31 Qualidade dos dados

Técnicas de Monitoramento

Table 31 Data Quality Monitoring Techniques

Granularity	In-stream (In-Process Flow) Treatment	Batch Treatment
Data Element	Edit checks in application Data element validation services Specially programmed applications	Direct queries Data profiling or analyzer tool
Data Record	Edit checks in application Data record validation services Specially programmed applications	Direct queries Data profiling or analyzer tool
Data set	Inspection inserted between processing stages	Direct queries Data profiling or analyzer tool

2.7.3 Desenvolver Procedimentos Operacionais para Gerenciar Problemas de Dados Quaisquer que sejam as ferramentas usadas para monitorar a qualidade dos dados, quando os resultados são avaliados pelos membros da equipe de Qualidade de Dados, eles precisam responder às descobertas de forma oportuna e eficaz. A equipe deve projetar e implementar procedimentos operacionais detalhados para:

- **Diagnosticando problemas:** O objetivo é revisar os sintomas do incidente de qualidade de dados, rastrear a linhagem dos dados em questão, identificar o problema e onde ele se originou, e apontar as potenciais causas raiz do problema. O procedimento deve descrever como a equipe de Operações de Qualidade de Dados:

- Rever as questões de dados no contexto dos fluxos de processamento de informações apropriados e

isolar o local no processo onde a falha é introduzida

- Avaliar se houve alguma alteração ambiental que possa causar erros na entrada do sistema
- Avalie se há ou não outros problemas de processo que contribuíram para o incidente de qualidade de dados
- Determinar se há problemas com dados externos que afetaram a qualidade dos dados

NOTA: O trabalho de análise de causa raiz requer a contribuição de SMEs técnicos e de negócios. Embora a equipe de DQ possa liderar e facilitar esse tipo de esforço de trabalho, o sucesso requer colaboração multifuncional

- **Formulação de opções para remediação:** Com base no diagnóstico, avalie alternativas para abordar o problema. Isso pode incluir:

- Abordar causas raiz não técnicas, como falta de treinamento, falta de apoio da liderança, responsabilidade e propriedade pouco claras, etc.
- Modificação dos sistemas para eliminar problemas técnicos causas raiz
- Desenvolvendo controles para evitar o problema
- Introdução de inspeção e monitoramento adicionais
- Corrigindo diretamente dados falhos
- Não tomar nenhuma ação com base no custo e no impacto da correção em relação ao valor da correção de dados

- **Resolução de problemas:** Tendo identificado opções para resolver

o problema, a equipe de Qualidade de Dados deve conferir com os proprietários de dados comerciais para determinar a melhor maneira de resolver o problema. Esses procedimentos devem detalhar como os analistas:

- Avalie os custos e méritos relativos das alternativas
- Recomendar uma das alternativas planejadas
- Fornecer um plano para desenvolver e implementar a resolução
- Implementar a resolução

As decisões tomadas durante o processo de gerenciamento de problemas devem ser rastreadas em um sistema de rastreamento de incidentes. Quando os dados em tal sistema são bem gerenciados, eles podem fornecer insights valiosos sobre as causas e os custos dos problemas de dados. Inclua uma descrição do problema e das causas raiz, opções para correção e a decisão sobre como resolver o problema.

O sistema de rastreamento de incidentes coletará dados de desempenho relacionados à resolução de problemas, atribuições de trabalho, volume de problemas, frequência de ocorrência, bem como o tempo para responder, diagnosticar, planejar uma solução e resolver problemas. Essas métricas podem fornecer insights valiosos sobre a eficácia do fluxo de trabalho atual, bem como sistemas e utilização de recursos, e são pontos de dados de gerenciamento importantes que podem impulsionar a melhoria operacional contínua para o controle de qualidade de dados.

Dados de rastreamento de incidentes também ajudam os consumidores de dados. Decisões baseadas em dados corrigidos devem ser tomadas com conhecimento de que eles foram alterados, por que foram alterados e como foram alterados. Essa é uma das razões pelas quais é importante registrar os métodos de modificação e a justificativa para eles. Disponibilize essa documentação para consumidores de dados e desenvolvedores que pesquisam alterações de código. Embora as alterações possam ser óbvias para as pessoas que as implementam, o histórico de alterações será perdido para dados futuros

consumidores, a menos que seja documentado. O rastreamento de incidentes de qualidade de dados exige que a equipe seja treinada sobre como os problemas devem ser classificados, registrados e rastreados. Para dar suporte ao rastreamento eficaz:

- **Padronize problemas e atividades de qualidade de dados:** Como os termos usados para descrever problemas de dados podem variar entre as linhas de negócios, é valioso definir um vocabulário padrão para os conceitos usados. Isso simplificará a classificação e o relatório. A padronização também facilita a medição do volume de problemas e atividades, a identificação de padrões e interdependências entre sistemas e participantes e o relatório sobre o impacto geral das atividades de qualidade de dados. A classificação de um problema pode mudar à medida que a investigação se aprofunda e as causas raiz são expostas.
- **Forneça um processo de atribuição para problemas de dados:** Os procedimentos operacionais direcionam os analistas a atribuir incidentes de qualidade de dados a indivíduos para diagnóstico e fornecer alternativas para resolução. Conduza o processo de atribuição dentro do sistema de rastreamento de incidentes sugerindo aqueles indivíduos com áreas específicas de especialização.
- **Gerenciar procedimentos de escalonamento de problemas:** O tratamento de problemas de qualidade de dados requer um sistema bem definido de escalonamento com base no impacto, duração ou urgência de um problema. Especifique a sequência de escalonamento dentro do Acordo de Nível de Serviço de qualidade de dados. O sistema de rastreamento de incidentes implementará os procedimentos de escalonamento, o que ajuda a agilizar o tratamento e a resolução eficientes de problemas de dados.
- **Gerenciar fluxo de trabalho de resolução de qualidade de dados:** O SLA de qualidade de dados especifica objetivos para monitoramento, controle e resolução, todos os quais definem uma coleção de fluxos de trabalho operacionais. O sistema de rastreamento de incidentes pode dar suporte ao gerenciamento de fluxo de trabalho para rastrear o progresso com diagnóstico e resolução de problemas.

2.7.4 Estabelecer Acordos de Nível de Serviço de Qualidade de Dados

Dados Um Acordo de Nível de Serviço (SLA) de qualidade de dados especifica as expectativas de uma organização para resposta e correção de problemas de qualidade de dados em cada sistema. As inspeções de qualidade de dados, conforme programadas no SLA, ajudam a identificar problemas a serem corrigidos e, com o tempo, reduzem o número de problemas. Ao permitir o isolamento e a análise da causa raiz de falhas de dados, há uma expectativa de que os procedimentos operacionais forneçam um esquema para correção de causas raiz dentro de um prazo acordado. Ter inspeção e monitoramento de qualidade de dados em vigor aumenta a probabilidade de detecção e correção de um problema de qualidade de dados antes que um impacto comercial significativo possa ocorrer. O controle de qualidade de dados operacionais definido em um SLA de qualidade de dados inclui:

- Elementos de dados abrangidos pelo acordo
- Impactos comerciais associados a falhas de dados
- Dimensões de qualidade de dados associadas a cada elemento de dados
- Expectativas de qualidade para cada elemento de dados para cada uma das dimensões identificadas em cada aplicação ou sistema na cadeia de valor de dados
- Métodos para medir essas expectativas
- Limiar de aceitabilidade para cada medição
- Os administradores devem ser notificados caso o limite de aceitabilidade não seja atingido
- Cronogramas e prazos para resolução ou remediação esperada do problema
- Estratégia de escalada e possíveis recompensas e penalidades

O SLA de qualidade de dados também define as funções e responsabilidades associadas ao desempenho dos procedimentos operacionais de qualidade de dados. Os procedimentos de qualidade de dados operacionais fornecem relatórios em conformidade com as regras de negócios definidas, bem como monitoramento

desempenho da equipe na reação a incidentes de qualidade de dados. Os administradores de dados e a equipe operacional de qualidade de dados, ao mesmo tempo em que mantêm o nível de serviço de qualidade de dados, devem considerar suas restrições de SLA de qualidade de dados e conectar a qualidade de dados a planos de desempenho individuais.

Quando os problemas não são resolvidos dentro dos tempos de resolução especificados, um processo de escalonamento deve existir para comunicar a não observância do nível de serviço na cadeia de gerenciamento e governança. O SLA de qualidade de dados estabelece os limites de tempo para geração de notificação, os nomes daqueles nessa cadeia de gerenciamento e quando o escalonamento precisa ocorrer. Dado o conjunto de regras de qualidade de dados, métodos para medir a conformidade, os limites de aceitabilidade definidos pelos clientes comerciais e os acordos de nível de serviço, a equipe de Qualidade de Dados pode monitorar a conformidade dos dados com as expectativas comerciais, bem como o desempenho da equipe de Qualidade de Dados nos procedimentos associados a erros de dados.

O relatório de SLA pode ser programado, impulsionado por requisitos comerciais e operacionais. Foco particular será na análise de tendências de relatórios em casos focados em recompensas e penalidades periódicas, se tais conceitos forem incorporados à estrutura de SLA.

2.7.5 Desenvolver Relatórios de Qualidade de

Dados O trabalho de avaliar a qualidade dos dados e gerenciar problemas de dados não beneficiará a organização a menos que as informações sejam compartilhadas por meio de relatórios para que os consumidores de dados entendam a condição dos dados. Os relatórios devem se concentrar em torno de:

- Scorecard de qualidade de dados, que fornece uma visão geral das pontuações associadas a várias métricas, relatadas a diferentes níveis da organização dentro de limites estabelecidos
- Tendências de qualidade de dados, que mostram ao longo do tempo como a qualidade dos dados é medida e se a tendência é de alta ou baixa
- Métricas de SLA, como se a equipe de qualidade de dados operacionais diagnostica e responde a incidentes de qualidade de dados em tempo hábil

maneiras

- Gerenciamento de problemas de qualidade de dados, que monitora o status dos problemas e resoluções
- Conformidade da equipe de Qualidade de Dados com as políticas de governança
- Conformidade das equipes de TI e negócios com as políticas de qualidade de dados
- Efeitos positivos dos projetos de melhoria

Os relatórios devem se alinhar às métricas no SLA de qualidade de dados o máximo possível, para que as metas da equipe estejam alinhadas com as de seus clientes. O programa de Qualidade de Dados também deve relatar os efeitos positivos dos projetos de melhoria. É melhor fazer isso em termos comerciais para lembrar continuamente a organização do efeito direto que os dados têm sobre os clientes.

3. Ferramentas

As ferramentas devem ser selecionadas e as arquiteturas de ferramentas devem ser definidas na fase de planejamento do programa Enterprise Data Quality. As ferramentas fornecem um kit inicial de conjunto de regras parcial, mas as organizações precisam criar e inserir suas próprias regras e ações específicas de contexto em qualquer ferramenta.

3.1 Ferramentas de Criação de Perfil

de Dados As ferramentas de criação de perfil de dados produzem estatísticas de alto nível que permitem que analistas identifiquem padrões em dados e realizem avaliações iniciais de características de qualidade. Algumas ferramentas podem ser usadas para realizar monitoramento contínuo de dados. As ferramentas de criação de perfil são particularmente importantes para esforços de descoberta de dados porque permitem a avaliação de grandes conjuntos de dados. Ferramentas de criação de perfil aumentadas com recursos de visualização de dados ajudarão no processo de descoberta. (Consulte os Capítulos 5 e 8 e a Seção 1.3.9.)

3.2 Ferramentas de consulta de dados

O perfil de dados é apenas o primeiro passo na análise de dados. Ele ajuda a identificar

problemas potenciais. Os membros da equipe de Qualidade de Dados também precisam consultar dados mais profundamente para responder a perguntas levantadas pelos resultados de criação de perfil e encontrar padrões que forneçam insights sobre as causas raiz dos problemas de dados. Por exemplo, consultar para descobrir e quantificar outros aspectos da qualidade de dados, como exclusividade e integridade.

3.3 Ferramentas de modelagem e ETL

As ferramentas usadas para modelar dados e criar processos ETL têm um impacto direto na qualidade dos dados. Se usadas com os dados em mente, essas ferramentas podem permitir dados de maior qualidade. Se forem usadas sem conhecimento dos dados, podem ter efeitos prejudiciais. Os membros da equipe de DQ devem trabalhar com as equipes de desenvolvimento para garantir que os riscos de qualidade dos dados sejam abordados e que a organização aproveite ao máximo as maneiras pelas quais a modelagem e o processamento de dados eficazes podem permitir dados de maior qualidade. (Consulte os Capítulos 5, 8 e 11.)



3.4 Modelos de regras de qualidade de dados

Os modelos de regras permitem que o analista capture as expectativas para os dados. Os modelos também ajudam a preencher a lacuna de comunicação entre as equipes de negócios e técnicas. A formulação consistente de regras facilita a tradução das necessidades de negócios em código, seja esse código incorporado em um mecanismo de regras, no componente de análise de dados de uma ferramenta de criação de perfil de dados ou em uma ferramenta de integração de dados. Um modelo pode ter várias seções, uma para cada tipo de regra de negócios a ser implementada.

3.5 Repositórios de Metadados

Conforme observado na [Seção 1.3.4](#), definir a qualidade dos dados requer Metadados e definições de dados de alta qualidade são um tipo valioso de Metadados. As equipes de DQ devem trabalhar em estreita colaboração com as equipes que gerenciam Metadados para garantir que os requisitos de qualidade dos dados, regras, resultados de medição e documentação de problemas sejam disponibilizados aos consumidores de dados.

4. Técnicas

4.1 Ações Preventivas

A melhor maneira de criar dados de alta qualidade é evitar que dados de baixa qualidade entrem em uma organização. Ações preventivas impedem que erros conhecidos ocorram. Inspeccionar dados depois que eles estão em produção não melhorará sua qualidade. As abordagens incluem:

- **Estabeleça controles de entrada de dados:** crie regras de entrada de dados que impeçam a entrada de dados inválidos ou imprecisos em um sistema.
- **Treine os produtores de dados:** garanta que a equipe nos sistemas upstream entenda o impacto de seus dados nos usuários downstream. Dê incentivos ou baseie as avaliações na precisão e integridade dos dados, em vez de apenas na velocidade.
- **Defina e aplique regras:** Crie um 'firewall de dados', que tem uma tabela com todas as regras de qualidade de dados empresariais usadas para verificar se a qualidade dos dados é boa, antes de serem usados em um aplicativo como um data warehouse. Um firewall de dados pode inspecionar o nível de qualidade dos dados processados por um aplicativo e, se o nível de qualidade estiver abaixo dos níveis aceitáveis, os analistas podem ser informados sobre o problema.
- **Exija dados de alta qualidade de fornecedores de dados:** Examine os processos de um provedor de dados externo para verificar suas estruturas, definições e fontes de dados e procedência de dados. Isso permite a avaliação de quão bem seus dados serão integrados e ajuda a evitar o uso de dados não autoritativos ou dados adquiridos sem a permissão do proprietário.
- **Implementar Governança e Administração de Dados:** Garantir que papéis e responsabilidades sejam definidos, descrevendo e aplicando regras de engajamento, direitos de decisão e responsabilidades para gerenciamento eficaz de dados e ativos de informação (McGilvray, 2008). Trabalhar com administradores de dados para revisar o processo e os mecanismos para gerar, enviar e receber dados.

- **Institua o controle formal de mudanças:** garanta que todas as mudanças nos dados armazenados sejam definidas e testadas antes de serem implementadas. Evite mudanças diretamente nos dados fora do processamento normal estabelecendo processos de controle.

4.2 Ações corretivas

Ações corretivas são implementadas após um problema ter ocorrido e sido detectado. Problemas de qualidade de dados devem ser abordados sistematicamente e em suas causas raiz para minimizar os custos e riscos de ações corretivas. "Resolva o problema onde ele acontece" é a melhor prática em Gerenciamento de Qualidade de Dados. Isso geralmente significa que ações corretivas devem incluir a prevenção da recorrência das causas dos problemas de qualidade.

Execute a correção de dados de três maneiras gerais:

- **Correção automatizada:** Técnicas de correção automatizada incluem padronização, normalização e correção baseadas em regras. Os valores modificados são obtidos ou gerados e confirmados sem intervenção manual. Um exemplo é a correção automatizada de endereços, que envia endereços de entrega para um padronizador de endereços que os conforma e corrige usando regras, análise sintática, padronização e tabelas de referência. A correção automatizada requer um ambiente com padrões bem definidos, regras comumente aceitas e padrões de erro conhecidos. A quantidade de correção automatizada pode ser reduzida ao longo do tempo se esse ambiente for bem gerenciado e os dados corrigidos forem compartilhados com sistemas upstream.
- **Correção direcionada manualmente:** use ferramentas automatizadas para remediar e corrigir dados, mas exija revisão manual antes de confirmar as correções no armazenamento persistente. Aplique correção de nome e endereço, resolução de identidade e correções baseadas em padrões automaticamente e use alguns

mecanismo de pontuação para propor um nível de confiança na correção. Correções com pontuações acima de um nível particular de confiança podem ser comprometidas sem revisão, mas correções com pontuações abaixo do nível de confiança são apresentadas ao administrador de dados para revisão e aprovação.

Confirme todas as correções aprovadas e revise aquelas não aprovadas para entender se deve ajustar as regras subjacentes aplicadas. Ambientes nos quais conjuntos de dados sensíveis exigem supervisão humana (por exemplo, MDM) são bons exemplos de onde a correção manual pode ser adequada.

- **Correção manual:** Às vezes, a correção manual é a única opção na ausência de ferramentas ou automação ou se for determinado que a mudança é melhor tratada por meio de supervisão humana. As correções manuais são melhor feitas por meio de uma interface com controles e edições, que fornecem uma trilha de auditoria para mudanças. A alternativa de fazer correções e confirmar os registros atualizados diretamente em ambientes de produção é extremamente arriscada. Evite usar esse método.

4.3 Módulos de código de auditoria e verificação de qualidade

Crie módulos de código compartilháveis, vinculáveis e reutilizáveis que executem verificações de qualidade de dados repetidas e processos de auditoria que os desenvolvedores podem obter de uma biblioteca. Se o módulo precisar mudar, todo o código vinculado a esse módulo será atualizado. Esses módulos simplificam o processo de manutenção. Blocos de código bem projetados podem evitar muitos problemas de qualidade de dados. Tão importante quanto isso, eles garantem que os processos sejam executados de forma consistente. Onde leis ou políticas exigem relatórios de resultados de qualidade específicos, a linhagem dos resultados geralmente precisa ser descrita. Os módulos de verificação de qualidade podem fornecer isso. Para dados que tenham qualquer dimensão de qualidade questionável e que sejam altamente classificados, qualifique as informações nos ambientes compartilhados com notas de qualidade e classificações de confiança.

4.4 Métricas de qualidade de dados eficazes

Um componente crítico do gerenciamento da qualidade dos dados é desenvolver métricas que informem os consumidores de dados sobre as características de qualidade que são importantes para seus usos de dados. Muitas coisas podem ser medidas, mas nem todas valem o tempo e o esforço. Ao desenvolver métricas, os analistas de DQ devem levar em conta essas características:

- **Mensurabilidade:** Uma métrica de qualidade de dados deve ser mensurável – precisa ser algo que possa ser contado. Por exemplo, a relevância dos dados não é mensurável, a menos que critérios claros sejam definidos para o que torna os dados relevantes. Até mesmo a completude dos dados precisa ser definida objetivamente para ser medida. Os resultados esperados devem ser quantificáveis dentro de um intervalo discreto.
- **Relevância comercial:** Embora muitas coisas sejam mensuráveis, nem todas se traduzem em métricas úteis. As medições precisam ser relevantes para os consumidores de dados. O valor da métrica é limitado se não puder ser relacionado a algum aspecto das operações ou desempenho comercial. Cada métrica de qualidade de dados deve se correlacionar com a influência dos dados nas principais expectativas comerciais.
- **Aceitabilidade:** As dimensões de qualidade de dados enquadram os requisitos de negócios para qualidade de dados. Quantificar ao longo da dimensão identificada fornece evidências concretas dos níveis de qualidade de dados. Determine se os dados atendem às expectativas de negócios com base em limites de aceitabilidade especificados. Se a pontuação for igual ou exceder o limite, a qualidade dos dados atende às expectativas de negócios. Se a pontuação estiver abaixo do limite, não.
- **Responsabilidade/Administração:** As métricas devem ser entendidas e aprovadas pelas principais partes interessadas (por exemplo, proprietários de negócios e administradores de dados). Eles são notificados quando a medição da métrica mostra que a qualidade não atende às expectativas. O proprietário dos dados comerciais é responsável, enquanto um administrador de dados toma as devidas

ação corretiva.

- **Controlabilidade:** Uma métrica deve refletir um aspecto controlável do negócio. Em outras palavras, se a métrica estiver fora do intervalo, ela deve desencadear uma ação para melhorar os dados. Se não houver maneira de responder, então a métrica provavelmente não é útil.
- **Tendências:** Métricas permitem que uma organização mensure a melhoria da qualidade de dados ao longo do tempo. O rastreamento ajuda os membros da equipe de Qualidade de Dados a monitorar atividades dentro do escopo de um SLA de qualidade de dados e acordo de compartilhamento de dados, e demonstrar a eficácia das atividades de melhoria. Uma vez que um processo de informação esteja estável, técnicas de controle estatístico de processos podem ser aplicadas para detectar mudanças na previsibilidade dos resultados de medição e nos processos comerciais e técnicos sobre os quais ele fornece insights.

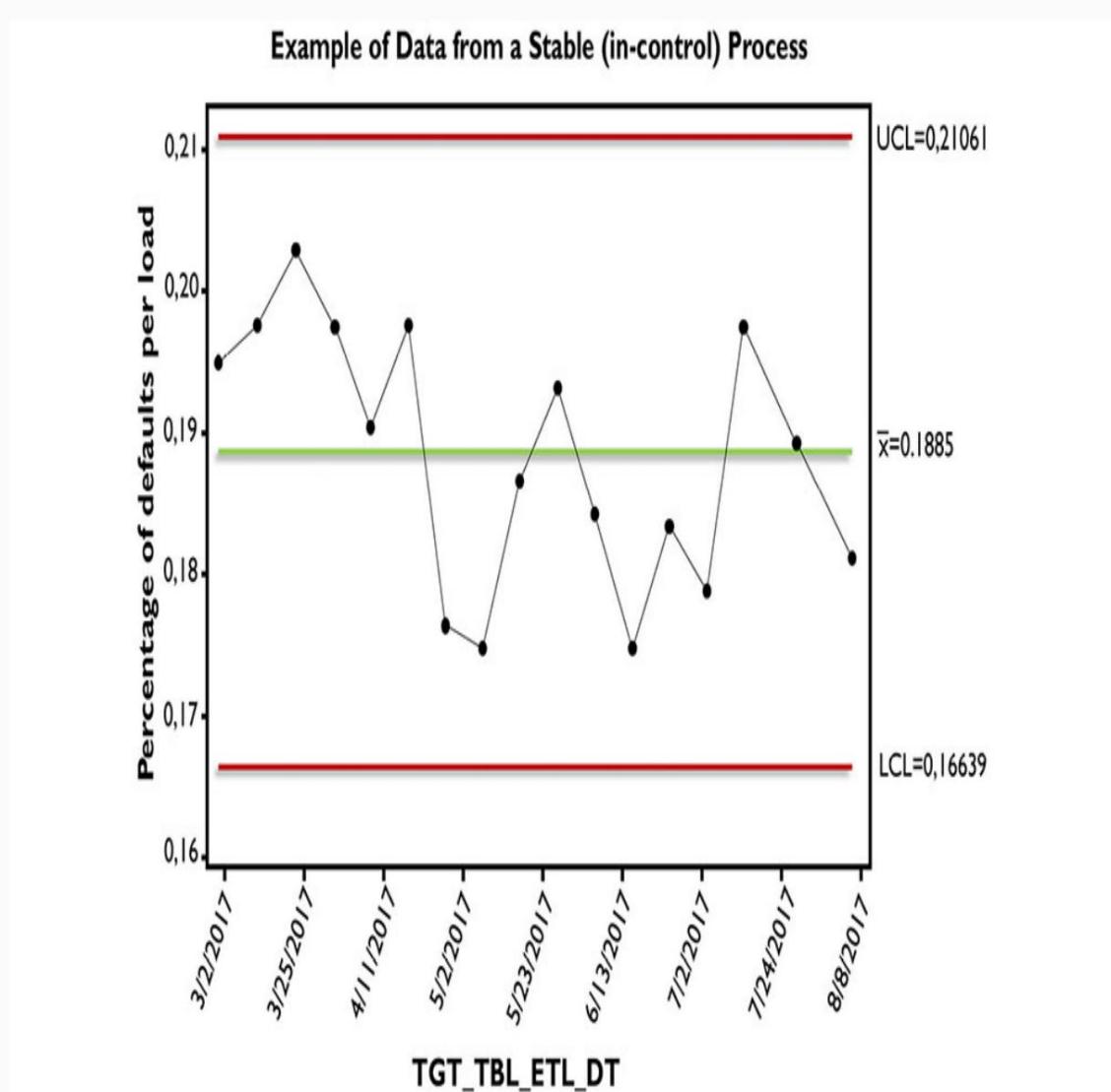
4.5 Controle Estatístico de Processos

O Controle Estatístico de Processos (CEP) é um método para gerenciar processos analisando medições de variação em entradas, saídas ou etapas do processo. A técnica foi desenvolvida no setor de manufatura na década de 1920 e tem sido aplicada em outras indústrias, em metodologias de melhoria como Six Sigma e em Gerenciamento de Qualidade de Dados.⁸⁷ Definido de forma simples, um processo é uma série de etapas executadas para transformar entradas em saídas. O CEP é baseado na suposição de que quando um processo com entradas consistentes é executado consistentemente, ele produzirá saídas consistentes. Ele usa medidas de tendência central (como os valores se agrupam em torno de um valor central, como uma média, mediana ou moda) e de variabilidade em torno de um valor central (por exemplo, intervalo, variância, desvio padrão), para estabelecer tolerâncias para variação dentro de um processo.

A ferramenta primária usada para SPC é o gráfico de controle (Figura 95), que é um gráfico de série temporal que inclui uma linha central para a média (a medida da tendência central) e descreve limites de controle superior e inferior calculados (variabilidade em torno de um valor central). Em um processo estável,

resultados de medição fora dos limites de controle indicam uma condição especial causa.

Figura 95 Gráfico de controle de um Processo em Controle Estatístico



que são imprevisíveis ou intermitentes. Quando as únicas fontes de variação são causas comuns, diz-se que um sistema está em controle (estatístico) e uma faixa de variação normal pode ser estabelecida. Esta é a linha de base contra a qual a mudança pode ser detectada.

A aplicação do SPC à medição da qualidade de dados é baseada na suposição de trabalho de que, como um produto manufaturado, os dados são o produto de um processo. Às vezes, o processo que cria dados é muito simples (por exemplo, uma pessoa preenche um formulário). Outras vezes, os processos são bastante complexos: um conjunto de algoritmos agrupa dados de reivindicações médicas para seguir tendências relacionadas à eficácia de protocolos clínicos específicos. Se tal processo tiver entradas consistentes e for executado consistentemente, ele produzirá resultados consistentes sempre que for executado. No entanto, se as entradas ou a execução mudarem, as saídas também mudarão. Cada um desses componentes pode ser medido. As medições podem ser usadas para detectar causas especiais. O conhecimento das causas especiais pode ser usado para mitigar riscos associados à coleta ou processamento de dados.

O SPC é usado para controle, detecção e melhoria. O primeiro passo é medir o processo para identificar e eliminar causas especiais. Esta atividade estabelece o estado de controle do processo. O próximo passo é colocar em prática medições para detectar variações inesperadas assim que forem detectáveis. A detecção precoce de problemas simplifica a investigação de suas causas raiz. As medições do processo também podem ser usadas para reduzir os efeitos indesejados de causas comuns de variação, permitindo maior eficiência.

4.6 Análise de Causa Raiz Uma

causa raiz de um problema é um fator que, se eliminado, removeria o problema em si. A análise de causa raiz é um processo de compreensão dos fatores que contribuem para os problemas e as maneiras como eles contribuem. Seu propósito é identificar condições subjacentes que, se eliminadas, significariam que os problemas desapareceriam.

Um exemplo de gerenciamento de dados pode esclarecer a definição. Digamos que um processo de dados que é executado a cada mês requer como entrada um arquivo de informações do cliente. A medição dos dados mostra que em abril, julho, outubro e janeiro, a qualidade dos dados cai. Inspeção de

o momento da entrega mostra que em março, junho, setembro e dezembro, o arquivo é entregue no dia 30 do mês, enquanto em outras épocas é entregue no dia 25. Uma análise mais aprofundada mostra que a equipe responsável pela entrega do arquivo também é responsável pelo fechamento dos processos financeiros trimestrais. Esses processos têm precedência sobre outros trabalhos e os arquivos são entregues com atraso durante esses meses, impactando a qualidade. A causa raiz do problema de qualidade de dados acaba sendo um atraso no processo causado por uma prioridade concorrente. Ele pode ser resolvido agendando a entrega do arquivo e garantindo que os recursos possam entregar dentro do cronograma.

Técnicas comuns para análise de causa raiz incluem análise de Pareto (regra 80/20), análise de diagrama de espinha de peixe, rastreamento e localização, análise de processo e os Cinco Porquês (McGilvray, 2008).

5. Diretrizes de implementação

Melhorar a qualidade dos dados dentro de uma organização não é uma tarefa fácil – mesmo quando os esforços de melhoria da qualidade dos dados são lançados de dentro de um programa de governança de dados e com o suporte da alta gerência. Uma discussão acadêmica clássica é se é melhor implementar um programa de Qualidade de Dados de cima para baixo ou de baixo para cima. Normalmente, uma abordagem híbrida funciona melhor – de cima para baixo para patrocínio, consistência e recursos, mas de baixo para cima para descobrir o que está realmente quebrado e obter sucessos incrementais.

Melhorar a qualidade dos dados requer mudanças na forma como as pessoas pensam e se comportam em relação aos dados. A mudança cultural é desafiadora. Ela requer planejamento, treinamento e reforço. (Consulte [o Capítulo 17.](#)) Embora as especificidades da mudança cultural sejam diferentes de organização para organização, a maioria das implementações de programas de Qualidade de Dados precisa planejar:

- **Métricas sobre o valor dos dados e o custo de dados de baixa qualidade:** Uma maneira de aumentar a conscientização organizacional sobre a necessidade de Gerenciamento da Qualidade de Dados é por meio de métricas que descrevem o valor dos dados e o retorno do investimento

de melhorias. Essas métricas (que diferem das pontuações de qualidade de dados) fornecem a base para o financiamento de melhorias e a mudança do comportamento tanto da equipe quanto da gerência.

(Ver [Capítulo 11.](#))

- **Modelo operacional para interações de TI/negócios:** os empresários sabem quais são os dados importantes e o que eles significam. Os custodiantes de dados de TI entendem onde e como os dados são armazenados e, portanto, estão bem posicionados para traduzir definições de qualidade de dados em consultas ou códigos que identificam registros específicos que não estão em conformidade. ([Consulte o Capítulo 11.](#))
- **Mudanças em como os projetos são executados:** A supervisão do projeto deve garantir que o financiamento do projeto inclua etapas relacionadas à qualidade dos dados (por exemplo, criação de perfil e avaliação, definição de expectativas de qualidade, remediação, prevenção e correção de problemas de dados, construção de controles e medições). É prudente garantir que os problemas sejam identificados cedo e criar expectativas de qualidade de dados antecipadamente nos projetos.
- **Mudanças nos processos de negócios:** Melhorar a qualidade dos dados depende da melhoria dos processos pelos quais os dados são produzidos. A equipe de Qualidade de Dados precisa ser capaz de avaliar e recomendar mudanças em processos não técnicos (assim como técnicos) que impactam a qualidade dos dados.
- **Financiamento para projetos de remediação e melhoria:** Algumas organizações não planejam remediar dados, mesmo quando estão cientes de problemas de qualidade de dados. Os dados não se consertarão sozinhos. Os custos e benefícios dos projetos de remediação e melhoria devem ser documentados para que o trabalho de melhoria de dados possa ser priorizado.
- **Financiamento para operações de qualidade de dados:** manter a qualidade dos dados requer operações contínuas para monitorar a qualidade dos dados, relatar descobertas e continuar a gerenciar problemas à medida que são descobertos.

5.1 Avaliação de prontidão / Avaliação de risco

A maioria das organizações que dependem de dados tem muitas oportunidades de melhoria. O quanto formal e bem suportado será um programa de Qualidade de Dados depende de quanto madura a organização é de uma perspectiva de gerenciamento de dados. (Consulte o Capítulo 15.) A prontidão organizacional para adotar práticas de qualidade de dados pode ser avaliada considerando as seguintes características:

- **Compromisso da gerência com o gerenciamento de dados como um ativo estratégico:** Como parte da solicitação de suporte para um programa de Qualidade de Dados, é importante determinar o quanto bem a gerência sênior entende o papel que os dados desempenham na organização. Até que ponto a gerência sênior reconhece o valor dos dados para objetivos estratégicos? Quais riscos eles associam a dados de baixa qualidade? Quão bem informados eles são sobre os benefícios da governança de dados? Quão otimistas sobre a capacidade de mudar a cultura para dar suporte à melhoria da qualidade?
- **A compreensão atual da organização sobre a qualidade de seus dados:** antes que a maioria das organizações comece sua jornada de melhoria da qualidade, elas geralmente entendem os obstáculos e os pontos problemáticos que indicam dados de baixa qualidade. Obter conhecimento sobre eles é importante. Por meio deles, dados de baixa qualidade podem ser diretamente associados a efeitos negativos, incluindo custos diretos e indiretos, na organização. Uma compreensão dos pontos problemáticos também ajuda a identificar e priorizar projetos de melhoria.
- **O estado real dos dados:** Encontrar uma maneira objetiva de descrever a condição dos dados que está causando pontos problemáticos é o primeiro passo para melhorar os dados. Os dados podem ser medidos e descritos por meio de criação de perfil e análise, bem como por meio da quantificação de problemas conhecidos e pontos problemáticos. Se a equipe de DQ não souber o estado real dos dados, será difícil priorizar e agir sobre as oportunidades

para melhoria.

- **Riscos associados à criação, processamento ou uso de dados:** Identificar o que pode dar errado com os dados e o dano potencial a uma organização devido a dados de baixa qualidade fornece a base para mitigar riscos. Se a organização não reconhecer esses riscos, pode ser desafiador obter suporte para o programa Data Quality.
- **Prontidão cultural e técnica para monitoramento escalável da qualidade de dados:** A qualidade dos dados pode ser impactada negativamente por processos comerciais e técnicos. Melhorar a qualidade dos dados depende da cooperação entre as equipes comerciais e de TI. Se o relacionamento entre as equipes comerciais e de TI não for colaborativo, será difícil progredir.

As descobertas de uma avaliação de prontidão ajudarão a determinar onde começar e com que rapidez prosseguir. As descobertas também podem fornecer a base para metas de programa de roteiro. Se houver forte suporte para melhoria da qualidade de dados e a organização conhecer seus próprios dados, então pode ser possível lançar um programa estratégico completo. Se a organização não souber o estado real de seus dados, então pode ser necessário se concentrar em construir esse conhecimento antes de desenvolver uma estratégia completa.

5.2 Organização e Mudança Cultural

A qualidade dos dados não será melhorada por meio de uma coleção de ferramentas e conceitos, mas por meio de uma mentalidade que ajude os funcionários e as partes interessadas a agirem sempre pensando na qualidade dos dados e no que a empresa e seus clientes precisam. Fazer com que uma organização seja consciente sobre a qualidade dos dados geralmente requer uma mudança cultural significativa. Essa mudança requer visão e liderança. (Consulte o Capítulo 17.)

O primeiro passo é promover a conscientização sobre o papel e a importância dos dados para a organização. Todos os funcionários devem agir de forma responsável e levantar questões de qualidade de dados, pedir dados de boa qualidade como consumidores e

fornecer informações de qualidade para outros. Cada pessoa que toca nos dados pode impactar a qualidade desses dados. A qualidade dos dados não é apenas responsabilidade de uma equipe de DQ ou grupo de TI.

Assim como os funcionários precisam entender o custo para adquirir um novo cliente ou manter um cliente existente, eles também precisam saber os custos organizacionais de dados de baixa qualidade, bem como as condições que fazem com que os dados sejam de baixa qualidade. Por exemplo, se os dados do cliente estiverem incompletos, um cliente pode receber o produto errado, criando custos diretos e indiretos para uma organização. O cliente não apenas devolverá o produto, mas também poderá ligar e reclamar, usando o tempo do call center, com potencial para danos à reputação da organização. Se os dados do cliente estiverem incompletos porque a organização não estabeleceu requisitos claros, todos que usam esses dados têm interesse em esclarecer os requisitos e seguir os padrões.

Por fim, os funcionários precisam pensar e agir de forma diferente se quiserem produzir dados de melhor qualidade e gerenciar dados de forma a garantir a qualidade. Isso requer treinamento e reforço. O treinamento deve focar em:

- Causas comuns de problemas de dados
- Relacionamentos dentro do ecossistema de dados da organização e por que melhorar a qualidade dos dados requer uma abordagem empresarial
- Consequências de dados de baixa qualidade
- Necessidade de melhoria contínua (por que a melhoria não é algo pontual)
- Tornando-se 'linguístico de dados', prestes a articular o impacto dos dados na estratégia e sucesso organizacional, relatórios regulatórios, satisfação do cliente

O treinamento também deve incluir uma introdução a quaisquer mudanças no processo, com afirmações sobre como as mudanças melhoram a qualidade dos dados.

6. Qualidade de dados e governança de dados

Um programa de Qualidade de Dados é mais eficaz quando faz parte de um programa de governança de dados. Frequentemente, problemas de qualidade de dados são a razão para estabelecer governança de dados em toda a empresa (veja o [Capítulo 3](#)).

A incorporação de esforços de qualidade de dados ao esforço geral de governança permite que a equipe do programa de Qualidade de Dados trabalhe com uma variedade de partes interessadas e facilitadores:

- Pessoal de risco e segurança que pode ajudar a identificar vulnerabilidades organizacionais relacionadas a dados
- Equipe de engenharia e treinamento de processos de negócios que pode ajudar as equipes a implementar melhorias de processos
- Administradores de dados comerciais e operacionais e proprietários de dados que podem identificar dados críticos, definir padrões e expectativas de qualidade e priorizar a correção de problemas de dados

Uma Organização de Governança pode acelerar o trabalho de uma organização de dados Programa de qualidade por:

- Estabelecendo prioridades
- Identificar e coordenar o acesso daqueles que devem estar envolvidos em várias decisões e atividades relacionadas à qualidade dos dados
- Desenvolver e manter padrões para qualidade de dados
- Relatar medições relevantes da qualidade de dados em toda a empresa
- Fornecer orientação que facilite o envolvimento da equipe
- Estabelecer mecanismos de comunicação para partilha de conhecimentos
- Desenvolver e aplicar políticas de qualidade e conformidade de dados

- Monitoramento e relatórios de desempenho
- Compartilhamento de resultados de inspeção de qualidade de dados para aumentar a conscientização, identificar oportunidades de melhorias e criar consenso para melhorias
- Resolver variações e conflitos; fornecer direção

6.1 Política de Qualidade de Dados

Os esforços de Qualidade de Dados devem ser apoiados por e devem apoiar políticas de governança de dados. Por exemplo, políticas de governança podem autorizar auditorias periódicas de qualidade e exigir conformidade com padrões e melhores práticas. Todas as Áreas de Conhecimento de Gerenciamento de Dados exigem algum nível de política, mas as políticas de qualidade de dados são particularmente importantes, pois frequentemente abordam requisitos regulatórios. Cada política deve incluir:

- Finalidade, âmbito e aplicabilidade da política
- Definições de termos
- Responsabilidades do programa de Qualidade de Dados
- Responsabilidades de outras partes interessadas
- Relatórios
- Implementação da política, incluindo links para risco, medidas preventivas, conformidade, proteção de dados e segurança de dados

6.2 Métricas

Grande parte do trabalho de uma equipe de Qualidade de Dados se concentrará em medir e relatar a qualidade. Categorias de alto nível de métricas de qualidade de dados incluem:

- **Retorno sobre o investimento:** Declarações sobre o custo dos esforços de melhoria versus os benefícios dos dados melhorados

qualidade

- **Níveis de qualidade:** Medidas do número e da porcentagem de erros ou violações de requisitos em um conjunto de dados ou entre conjuntos de dados
- **Tendências de qualidade de dados:** melhoria da qualidade ao longo do tempo (ou seja, uma tendência) em relação a limites e metas, ou incidentes de qualidade por período.
- **Métricas de gerenciamento de problemas de dados:**
 - Contagens de problemas por dimensões de qualidade de dados
 - Problemas por função empresarial e seus status (resolvidos, pendentes, escalonados)
 - Problema por prioridade e gravidade
 - Hora de resolver problemas
- **Conformidade com os níveis de serviço:** unidades organizacionais envolvidas e equipe responsável, intervenções do projeto para avaliações de qualidade de dados, conformidade geral do processo
- **Implementação do plano de qualidade de dados:** como está e roteiro para expansão

7. Trabalhos Citados / Recomendados

Batini, Carlo e Monica Scannapieco. *Qualidade de Dados: Conceitos, Metodologias e Técnicas*. Springer, 2006. Impresso.

Brackett, Michael H. *Qualidade de Recursos de Dados: Transformando Maus Hábitos em Boas Práticas*. Addison-Wesley, 2000. Imprimir.

Deming, W. Edwards. *Fora da Crise*. The MIT Press, 2000. Impresso.

Inglês, Larry. *Melhorando o Data Warehouse e a Qualidade das Informações Empresariais: Métodos para Reduzir Custos e Aumentar Lucros*. John Wiley and Sons, 1999. Impresso.