Governança e Qualidade de Dados

Qualidade de Dados

Arnaldo Vitaliano, MSc.

2017


data
quality
management
reporting
staff
Data
processes
procedures
recording
reviewed
collection
needs
responsibility
information
place
top
results
defined
level
associated
working
support
changes
may
covering
updates
clearly
commitment
needed
service
arrangements
regularly
body's
external
governance
taken
standards
corporate
example
senior
ensure
body
basis
set
used
training
agreement
Systems
necessary
reported
form
delivery
Policies
provision
regular
policy
timely
periodically
systems
analysis
business
action
guidance
charged
risk
address
complete
requirements
well
back
Accuracy
unauthorised
programme
areas
acted
clear
manipulation
relevant
practice
work
policies
objectives
monitoring
assessed
day
supported
Completeness
scrutiny
appraisal
missing
review
levels
system
Internal
corrective
role
necessitate
meet
practices
document
critical
rigorous
extensive
forecasting
member
cleansing
minimum
procedure
compiling
implemented
overall
correction
agreed
descriptions
tailored
guidelines
Procedures
identified
collecting
part
fed
right
erroneous
time
Arrangements
Relevance
performance
although
comprehensive
valid
updated
least
Act
documented
consistently
Corrective
previous
possible
relation
reinforce
understanding
built-in
Accountability
Information
evidence
monitor
communicated
reinforcing
reliable
actions


EU NÃO TENHO NÚMEROS EXATOS, ENTÃO EU INVENTEI ESSE DAQUI.
$4,629,873
www.dilbert.com scottadams@aol.com
ESTUDOS MOSTRAM QUE NÚMEROS EXATOS NÃO SÃO NECESSARIAMENTE MELHORES QUE AQUELES INVENTADOS.
QUANTOS ESTUDOS MOSTRARAM ISSO?
OITENTA E SETE.
5-8-08 ©2008 Scott Adams, Inc./Dist. by UFS, Inc.

Definindo Qualidade de Dados

Definindo Qualidade de Dados

"Dados são de alta qualidade, se aqueles que os utilizam disserem isso. Normalmente, os dados de alta qualidade deve ser tanto livres de defeitos e possuírem características que os clientes desejam." (Thomas Redman)

"O nível de qualidade dos dados é determinado pelos consumidores de dados em termos de quanto cumpre ou excede expectativas." (David Loshin)

"O grau segundo o qual os dados podem ser uma fonte confiável para todo e qualquer uso requerido. É ter o conjunto correto de informações corretas, no momento certo, no lugar certo, para as pessoas certas que tomam decisões e conduzem o negócio." (Danette McGilvray)

Definindo Qualidade de Dados

- Características associadas a dados com alto grau de qualidade

- Processos utilizados para medir ou melhorar a qualidade de dados

O que é alto grau de qualidade?

Definindo Qualidade de Dados

O dado tem alto grau de qualidade quando...

... ele atinge as expectativas e necessidades do consumidor dos dados.

... ele está apto para o propósito do seu uso.

Ou seja, qualidade de dados depende do contexto do uso.

Desafios

- Expectativas dos usuários nem sempre são conhecidas
- Gestores não perguntam sobre requisitos de qualidade
- Analistas de dados precisam entender melhor as necessidades dos usuários dos dados
- Discussão continuada sobre mudança nos requisitos, no tempo, na medida que o negócio evolui

Por quê?

Oportunidades

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

Custos derivados da baixa qualidade

Quem nunca recebeu uma ligação da sua própria operadora...

Custos derivados da baixa qualidade

- Inabilidade de cobrar corretamente
- Aumento de reclamações e diminuição na habilidade de resolvê-las
- Perdas de receitas devidas a oportunidades não identificadas
- Demora em integrações e junções em aquisições
- Exposição a fraudes
- Perdas provocadas por decisões baseadas em dados ruins

Objetivos

1. Desenvolver uma abordagem **governada** de tornar dados aptos ao uso
2. Definir padrões e especificações para **controle** da qualidade de dados (como parte do ciclo de vida da informação)
3. Definir e implementar processos de **medição, monitoramento e reporte** dos níveis de qualidade de dados
4. Identificar oportunidades de **melhoria** da qualidade dos dados

Princípios

1. Criticidade
2. Gestão de Cíclo de Vida
3. Prevenção
4. Remediação das Causas Primárias
5. Governança
6. Foco em Padrões
7. Medição Objetiva e Transparencia
8. Sistematicamente “Forçar”
9. Conectado aos níveis de serviço

Dimensões de Qualidade


Accuracy
Conciseness
Availability
Believability
Completeness
Data
Currency
Amount-of-data
Contextual
Consistency
Dataset
Interlinking
Intrinsic
LinkedData
Licensing
Dynamicity
Dimensions
Metrics
Performance
Reputation
Relevancy
Objectivity
Interpretability
Verifiability
Timeliness
Versatility
Understandability
Representational-consistency
Representational-conciseness
Representational
Volatility
SecurityTrust
Quality

Strong-Wang Framework (1996)

1. Intrínsecas
    1. Precisão
    2. Objetividade
    3. Credibilidade
    4. Reputação

2. Contextuais
    1. Valor agregado
    2. Relevância
    3. Tempestividade
    4. Completude
    5. Quantidade apropriada

3. Representacionais
    1. Interpretabilidade
    2. Facilidade de entendimento
    3. Consistência na representação
    4. Representação concisa

4. Acessíveis
    1. Acessibilidade
    2. Segurança de acesso

Larry English (1999)

- Características inerentes
  - Conformidade
  - Completude
  - Conformidade com regras de negócio
  - Precisão
  - Unicidade
  - Equivalência
  - Concorrência

- Características pragmáticas
  - Acessibilidade
  - Tempestividade
  - Clareza
  - Usabilidade
  - Integridade
  - Corretude

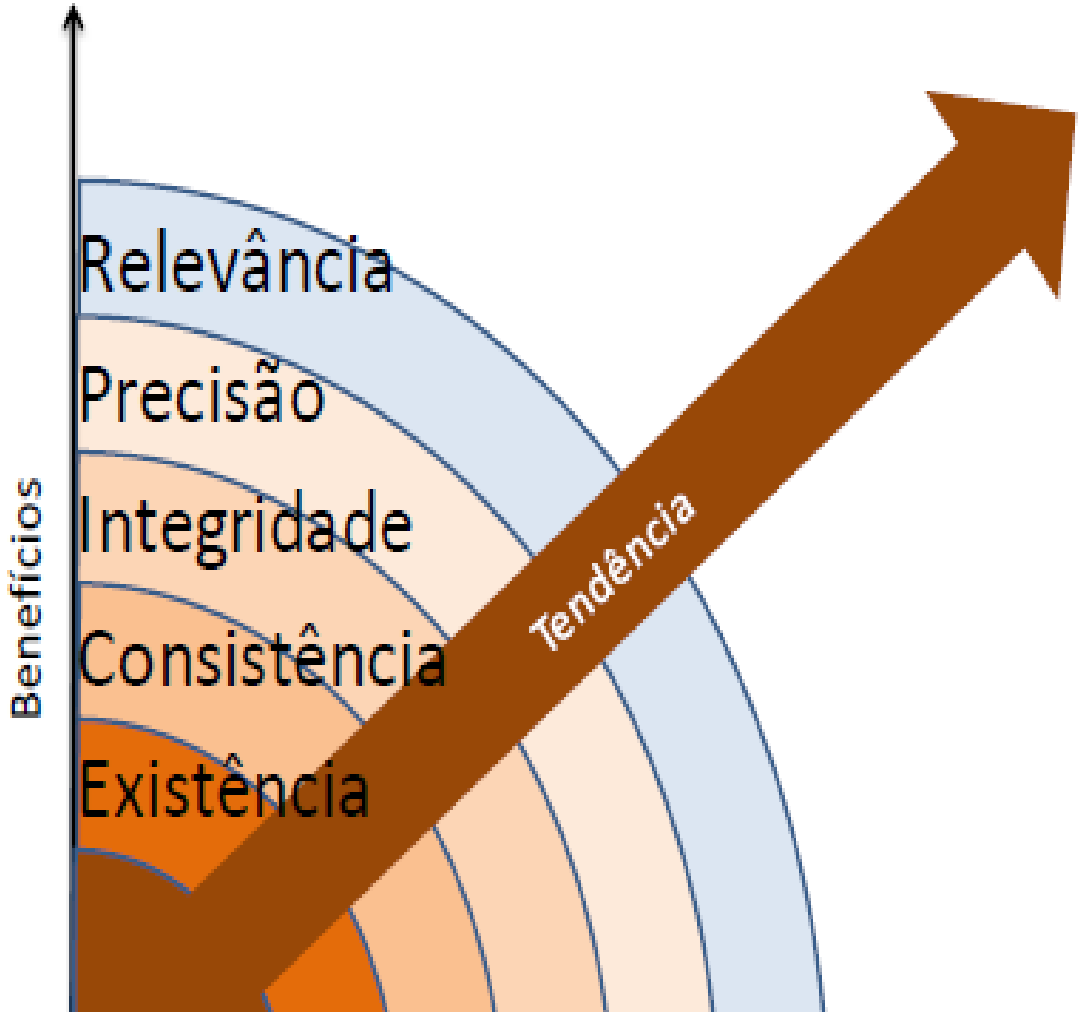
Dama UK (2003)

1. Completude
2. Unicidade
3. Tempestividade
4. Validade/Conformidade
5. Precisão
6. Consistência

Dimensões de Qualidade de Dados

- Definindo dimensões.

Dimensões de Qualidade de Dados



um defeito é qualquer situação onde os valores dos dados não são acessíveis ou não correspondem com exatidão a uma referência estabelecida.

Dimensões de Qualidade de Dados

Completude

Nome	Completude (*Completeness*)
Definição	O dado esperado é devidamente preenchido, diferente de nulo ou vazio.
Medição	Um dado existe quando seu valor é diferente de nulo, espaços em branco ou *strings* vazias.
Exemplo	Em uma tabela de operações de crédito, o cliente da operação é identificado pelo seu CPF ou CNPJ, no campo CLIENTE_COD. Uma medição nesta dimensão pode mostrar que em 14% dos registros, os dados de cliente não estão preenchidos. Logo, para este conjunto de dados, para esta coluna, temos um nível de qualidade de 86%.

Unicidade

Nome	Unicidade (*Uniqueness*)
Definição	O dado é representado uma única vez, baseando-se nas características que definem a sua identificação.
Medição	O dado é medido dentro do seu conjunto de dados (tabela), utilizando todos os atributos (campos) que compõem sua identificação única.
Exemplo	Na tabela de municípios, não pode haver duplicidade. Os atributos que identificam unicamente um município são (i) seu nome, e (ii) o estado a qual pertence. Em uma medição nesta tabela, um mesmo nome de município – São Francisco – pode ser encontrado em 3 diferentes estados, o que não viola esta dimensão.

Consistência

Nome	Consistência (*Consistency*)
Definição	A mesma representação é utilizada em todos os conjuntos de dados (tabelas/bases). Outra definição é a respeito de regras de negócio, ou seja, comportamentos esperados dentro do negócio.
Medição	O dado deve ser conferido em todos os pontos em que aparece. Ele deve ser representado da mesma maneira em todos os conjuntos.
Exemplo	Em uma tabela que contém o campo de CNPJ, a coluna é definida do tipo *integer*. Em outra tabela, o mesmo campo CNPJ é documentado como *char(8).* O campo não está consistente entre as duas tabelas.

Conformidade

Nome	Conformidade (*Validity*)
Definição	O dado é representado no formato, tamanho e domínio de dados esperado.
Medição	O dado deve ser conferido contra os metadados esperados ou documentados: tipos (texto, número, data), formatos: (dia/mês/ano), domínios: (de 0 a 1, de 10 a 100, masculino/feminino).
Exemplo	Em uma tabela que contém uma lista de CPFs, o padrão esperado é 999.999.999-99, ou seja, um texto (*string*), de tamanho 14, com pontos e hífen. Os exemplos 837726827-23 e 83772682723 são inválidos e o exemplo 837.726.827-23 é válido.

Dimensões de Qualidade de Dados

Integridade

Nome	Integridade (*Integrity*)
Definição	O conjunto de dados contém referências a outros dados – ou é referenciado, e o dado referenciado existe no outro conjunto de dados e pode ser rastreado.
Medição	Os dados que referenciam e são referenciados devem ser cruzados para validar as integridades.
Exemplo	Em uma tabela de operações de crédito há referências a contas contábeis. Um cruzamento é feito entre as tabelas para conferir se (i) as referências existem, e se (ii) as referências são as corretas.

Precisão/Validade

Nome	Precisão (*Accuracy*) ou Validade (*Validity*)
Definição	O grau em que o dado representa o objeto ou evento no “mundo real”.
Medição	A medição deve ser comparada com dados confiáveis – dados mestre, ou dados oficiais – IBGE, RFB, ISBN, IBAN.
Exemplo	A precisão de um dado se dá quando o valor avaliado se encontra dentro do domínio de valores esperados para o campo.

Tempestividade

Nome	Tempestividade (*Timeliness*)
Definição	Indica se o dado está disponível no ponto requerido do tempo.
Medição	Um dado deve refletir o instante no tempo em que realmente aconteceu.
Exemplo	Um registro de operação de câmbio realizada no primeiro dia do mês deve ser consolidado no mês corrente. Uma mensagem do sistema de pagamento deve ser processada no máximo 15 minutos depois de chegarem à fila de mensagens. As operações de crédito do mês corrente devem ser enviadas ao BCB pelas entidades supervisionadas até o 10º dia útil do mês subsequente.

Dimensões de Qualidade de Dados

Quiz

1. Qual das alternativas abaixo pode ser considerada sinônimo da dimensão de “Existência”?

a. Validade
b. Completude
c. Conformidade

2. Em uma tabela VENDAS qualquer, o cliente é identificado pelo CPF. Na tabela CADASTRO, o cliente é identificado por um código interno sequencial. Qual dimensão não está correta?

a. Conformidade
b. Consistência
c. Integridade

3. Em uma tabela de endereços qualquer, a coluna CEP é do tipo inteiro. Qual dimensão é impactada?

a. Conformidade
b. Consistência
c. Integridade

4. Que tipo de dado deve ser armazenado em uma coluna numérica (int/float/numeric/decimal)?

a. Qualquer dado representado por números, como CEP e CPF.
b. Dados que representam números no mundo real.
c. Dados que representam números no mundo real e em que são aplicadas operações matemáticas.

5. Em uma tabela OPERACOES_POR_MUNICIPIO, existe uma coluna chamada NUMERO_TOTAL_HABITANTES. Nas operações do município Brasília/DF, consta o valor de 300.000 habitantes. Segundo o IBGE, Brasília possui mais de 2 milhões de habitantes. Qual dimensão está comprometida?

a. Consistência
b. Integridade
c. Precisão

6. Em uma tabela OPERACOES_CAMBIO, existem 3.404 registros do banco X, cnpj: 11.222.333. Na tabela de referência de instituições financeiras, não há registro relativo ao banco X. Qual dimensão não está correta?

a. Consistência
b. Integridade
c. Precisão

7. Na tabela OPERACOES_CREDITO, existem alguns registros com o CEP 70360-707. Mas este CEP não consta na tabela de referência de CEPs. O CEP é real. Quais dimensões possuem problemas neste caso?

a. Consistência e Precisão
b. Integridade e Existência
c. Precisão e Consistência

8. Na tabela OPERACOES, a nota da operação deve ser entre F e HH quando há valores atrasados. Se uma operação com atrasos possui nota B, qual dimensão possui problema?

a. Conformidade
b. Consistência
c. Precisão

9. O DEPTO 01 mede um índice X pela fórmula A + B. O DEPTO 02 mede um índice X pela fórmula A + B * K.

Quando os resultados são armazenados em tabelas, qual dimensão não é assegurada?

a. Conformidade
b. Consistência
c. Precisão
d. Validade

ISO 8000

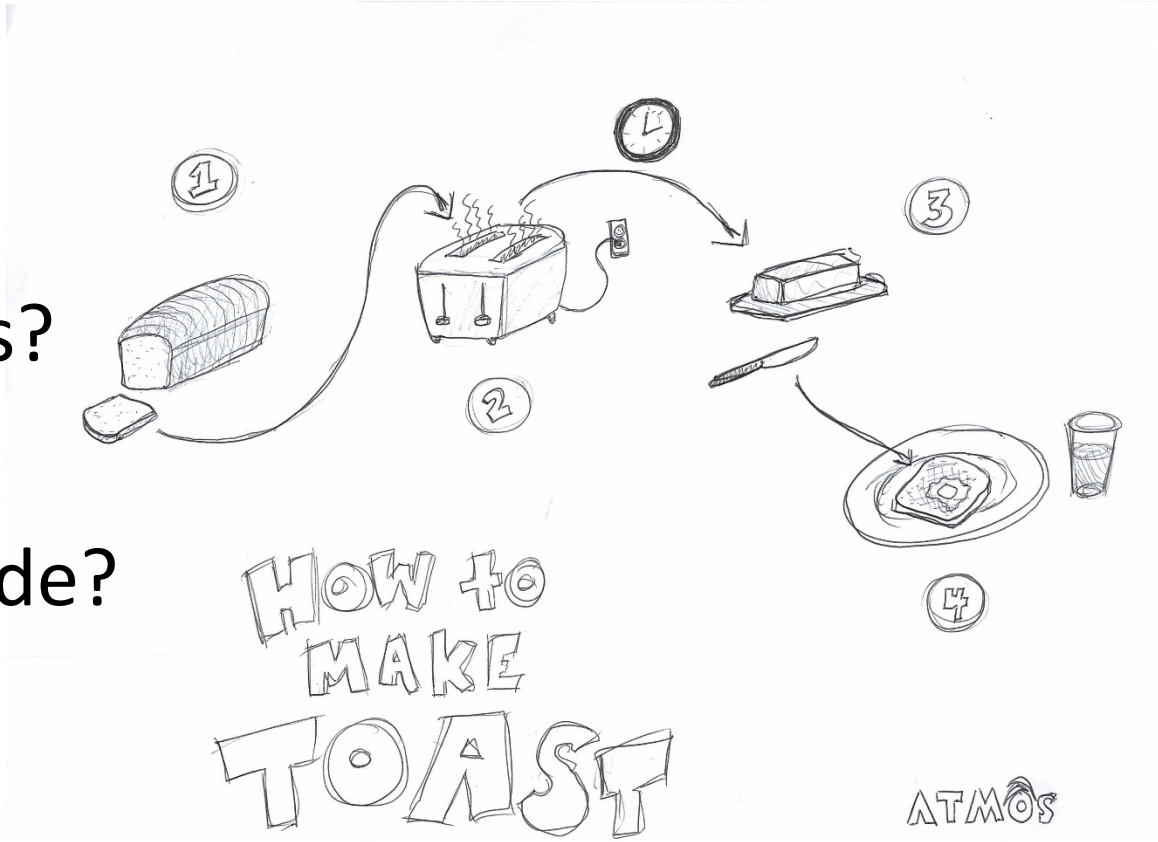
"A habilidade de criar, coletar, armazenar, manter, transferir, processar e apresentar dados para apoiar processos de negócio de uma maneira efetiva no tempo e no custo requer entendimento das características dos dados que determinam sua qualidade e a habilidade de medir, gerenciar e reportar qualidade nos dados."

Data Quality Planning
Data Quality Control
Data Quality Assurance
Data Quality Improvement

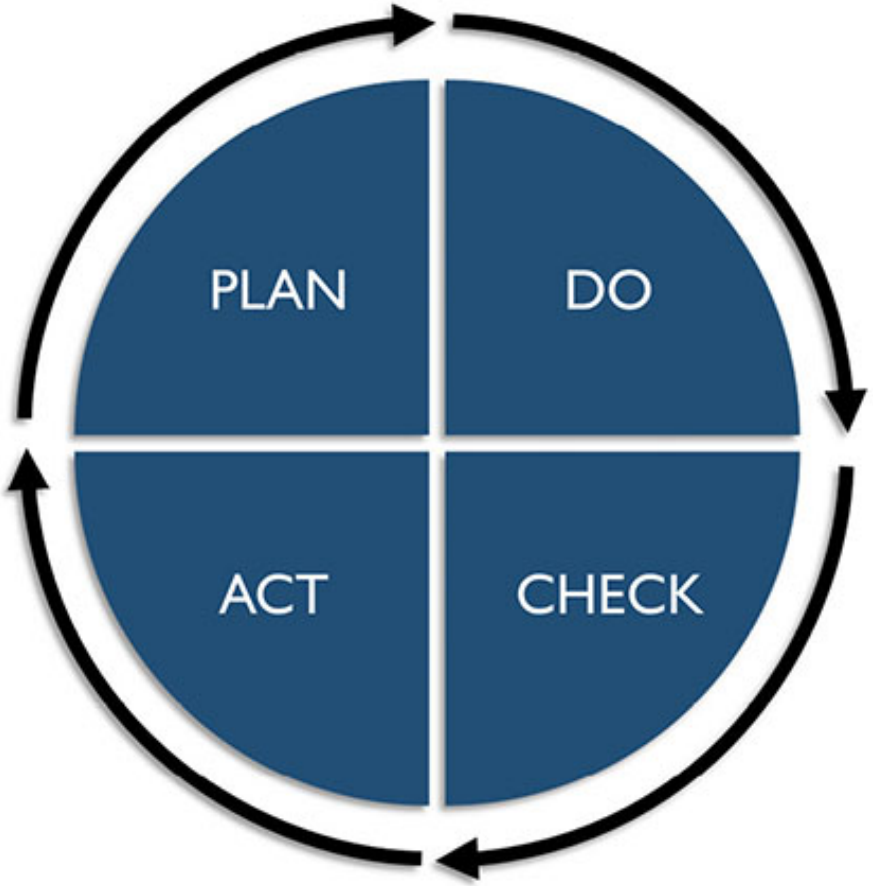
Processo de Qualidade

Processo de Qualidade

- Quais as atividades?
- Qual a sequência?
- Qual a granularidade?
- O que observar?

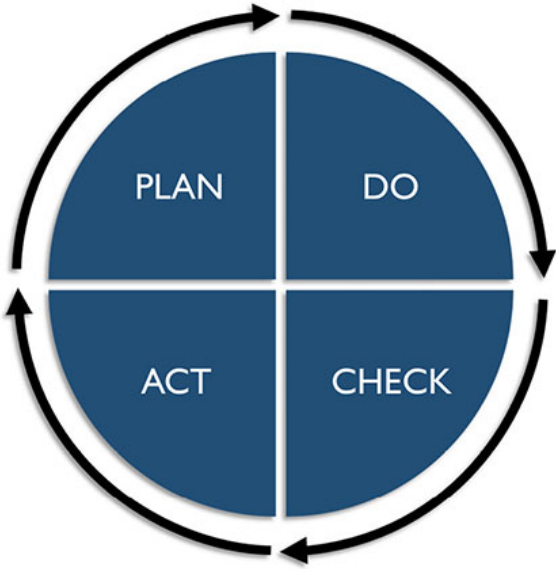


Plan/Do/Check/Act



93 The Shewhart Chart

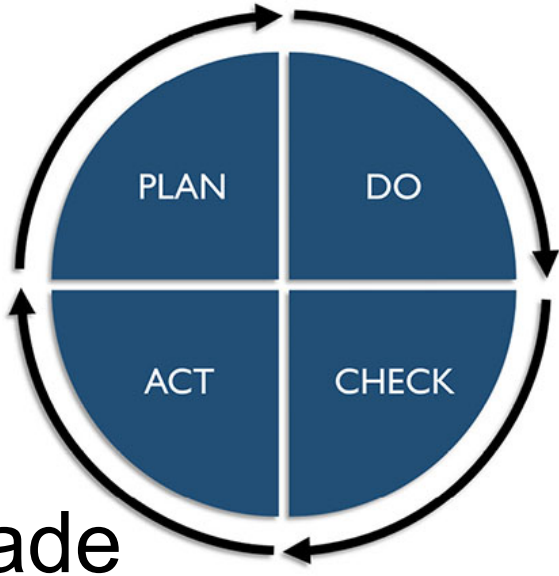
Planejar



- Avaliar
  - Escopo
  - Impacto
  - Prioridade
  - Problemas conhecidos &
  - Alternativas para endereçá-los

O plano deve ser baseado em uma análise profunda das causas raízes dos problemas. Do conhecimento das causas e impactos dos problemas, podemos entender a relação custo/benefício, e prioridades podem ser determinadas e um plano básico pode ser formulado para endereçar os problemas.

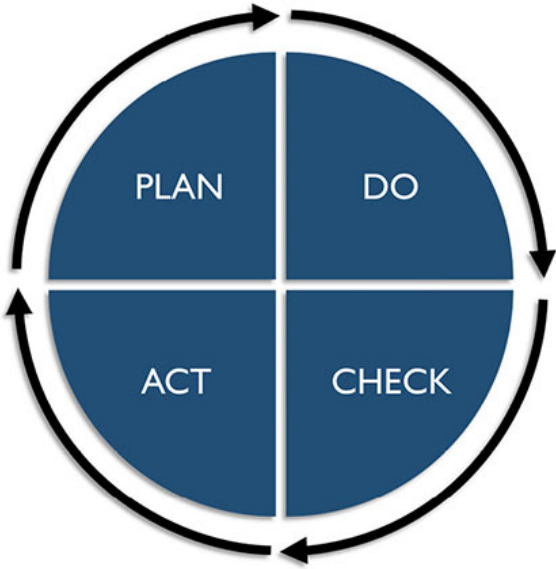
Fazer



Construir os processos para:

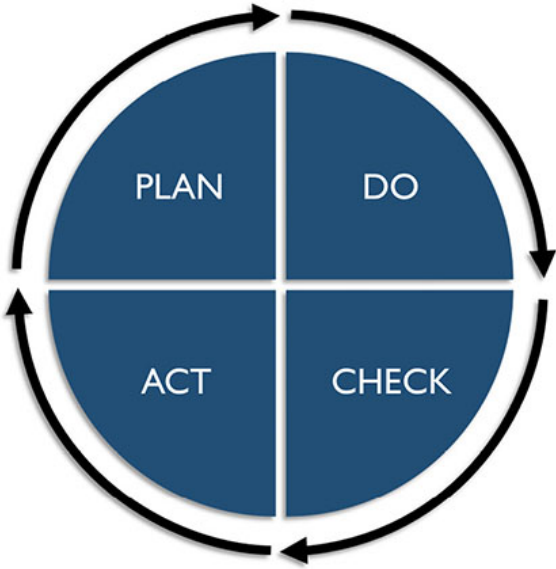
- Monitorar os níveis de qualidade
- Corrigir os problemas

Checar



- Monitoramento ativo
- Níveis de tolerância
- Gatilhos & Notificações

Agir (Tomar decisões)



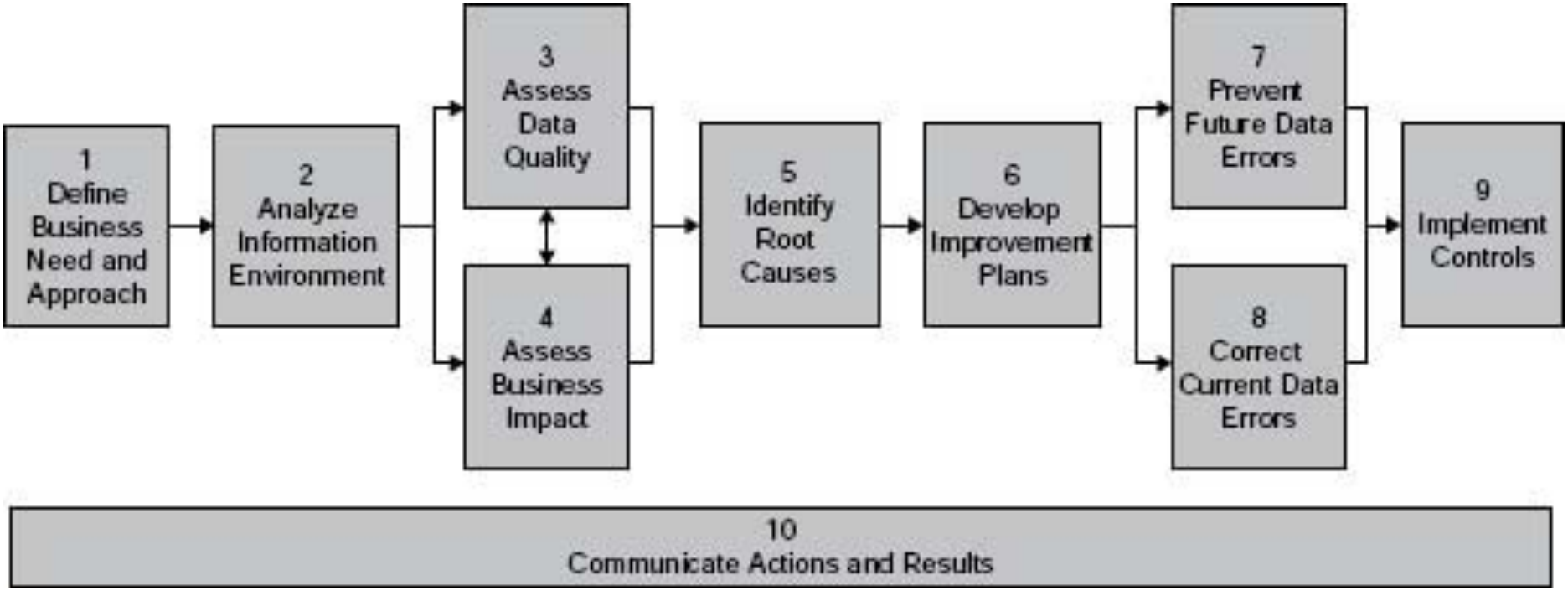
- Endereçar problemas novos
- Um novo ciclo se inicia
- Melhoria contínua

Arquitetura de Qualidade de Dados

- Que atividades fazem sentido para Qualidade de Dados para cada etapa PLAN/DO/CHECK/ ACT?

- Atividade: Desenhar uma arquitetura de qualidade de dados que contemple estas atividades.

Arquitetura de Qualidade de Dados



Source: Copyright © 2005–2008 Danette McGilvray, Granite Falls Consulting, Inc.

Atividades de Qualidade

Avaliando os dados: Perfil

- Fotografia dos dados
- Ideal para análise inicial em qualquer projeto
- Ideal quando não se conhece os dados
- Simples de executar
- Estatísticas básicas

Avaliando os dados: Perfil

- Que informações devemos buscar em um perfil de dados?

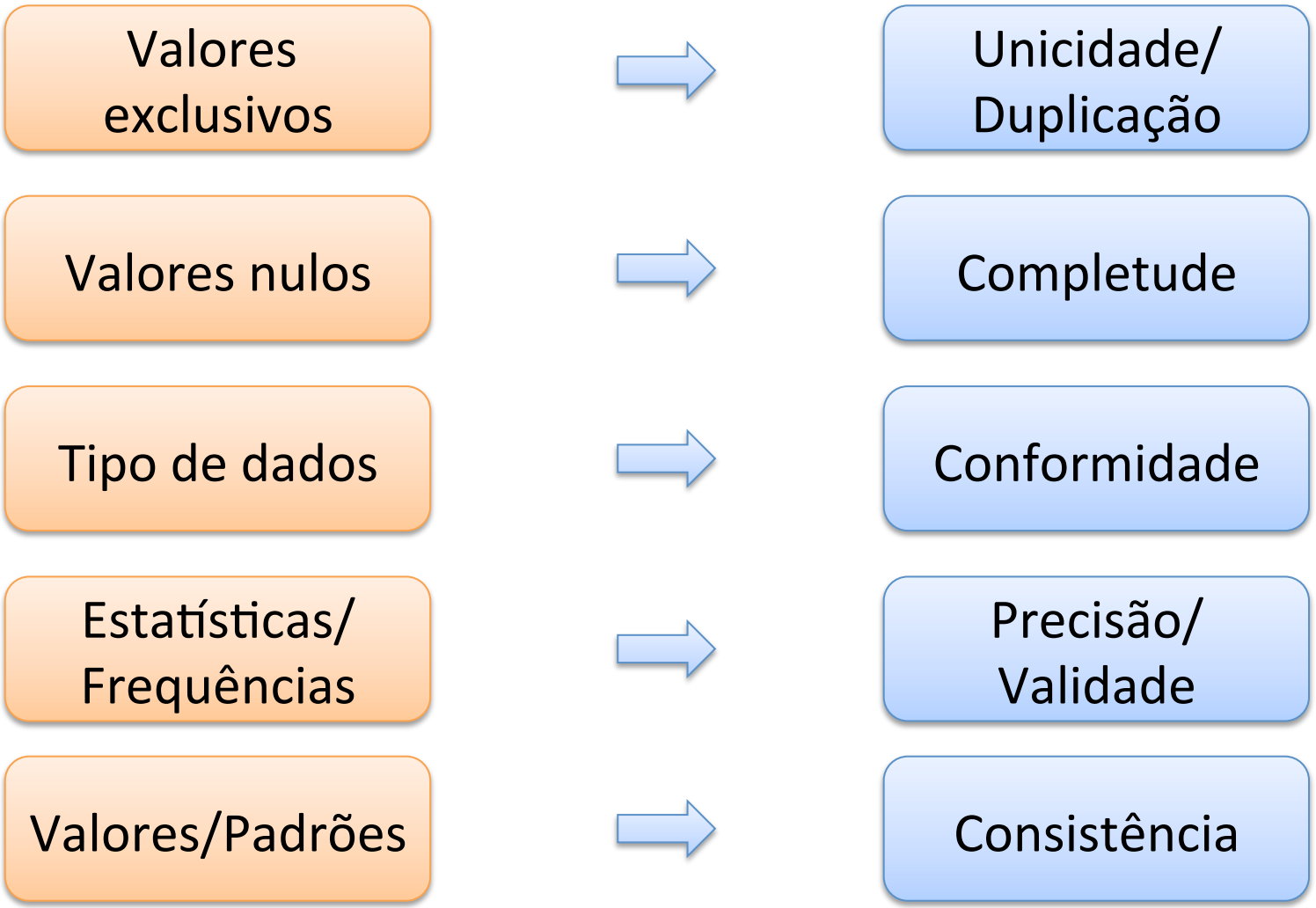
Avaliando os dados: Perfil

1. Contagem de valores únicos
2. Contagem de valores nulos
3. Valores mínimos/máximos
4. Tamanhos mínimos/máximos
5. Tipo e formato de dados
6. Distribuição e frequência de valores
7. Distribuição e frequência de padrões

Avaliando os dados: Perfil

1. Intersecção entre tabelas
2. Intersecção entre colunas
3. Identificação de domínios de dados
    1. Telefone
    2. Email
    3. Endereço
    4. Nome de pessoa
    5. Etc...

Perfil e Dimensões



Limpeza dos Dados (Data cleansing)

- Transformar dados para obter conformidade com padrões e domínios
- Remover ruídos/caracteres inválidos
- Remover dados irrelevantes
- Comparar com dados mestres/de referência
- Deduplicar
- Corrigir inconsistências

Padronização dos Dados (Data standardization)

- Tornar a representação homogênea
- Seguir padrões corporativos

- caixa baixa, CAIXA ALTA, Caixa Camelo
- M/F, Masculino/Feminino, 1/2
- True/False, 1/0, OK/ERRO
- Código IBGE, Receita ou Correios

Deduplicação

- Identificar inequivocamente indivíduos
- Remover ambiguidades

- João de Sousa Jr.
- João Sousa Jr.
- João de Souza Jr.
- João M. Souza Jr.

mesma pessoa?

Resolução de Identidade

Resolução de Identidade (*Identity Resolution*)

- Deduplicação específica para pessoas
- Leva em consideração documentos de identificação
  - CPF, RG, CNH, Passaporte, etc.
- Leva em consideração informações de relacionamentos
  - Emprego, Endereços, Assinaturas
- Uso de algoritmos de distância e técnicas de correspondência

Correspondência (*match*)

Operações conhecidas como “*match*” ou “correspondência”.

- Geração de pares mais “parecidos”
- Nota de similaridade
- Grupamento (*clustering*) de registros por similaridade

Correspondência (*match*)

- Considere os registros seguintes. Quantos registros duplicados existem?
- Existem 2 registros que podem ser considerados correspondências. Conseguem ver?
- O processo de correspondência segue por 3 fases lógicas:
  - Geração dos pares
    - Tipo de correspondência
  - Definição da nota
    - Estratégia
  - Processamento
    - Saídas

Nome	Endereço
David W Adams	Texas
Bill F Hawthorn	New York
Helen Sarah Hawthorn	New York
Nancy Smith	San Francisco
David H W Adams	Texas

Correspondência (*match*)

- Neste exemplo, cada linha da tabela irá ser comparada com todas as outras (CROSS JOIN). Isso nos dá um total de 10 pares.

Nome1	Endereço1	Nome2	Endereço2
David W Adams	Texas	Bill J Hawthorn	New York
David W Adams	Texas	Helen Sarah Hawthorn	New York
David W Adams	Texas	Nancy Smith	San Francisco
David W Adams	Texas	David H W Adams	Texas
Bill J Hawthorn	New York	Helen Sarah Hawthorn	New York
Bill J Hawthorn	New York	Nancy Smith	San Francisco
Bill J Hawthorn	New York	David H W Adams	Texas
Helen Sarah Hawthorn	New York	Nancy Smith	San Francisco
Helen Sarah Hawthorn	New York	David H W Adams	Texas
Nancy Smith	San Francisco	David H W Adams	Texas

Correspondência (*match*)

- A próxima fase atribui notas (1 indica que são idênticos) aos pares, indicando o grau de similaridade.

Nome1	Endereço1	Nome2	Endereço2	Nota
David W Adams	Texas	Bill J Hawthorn	New York	0
David W Adams	Texas	Helen Sarah Hawthorn	New York	0
David W Adams	Texas	Nancy Smith	San Francisco	0
David W Adams	Texas	David H W Adams	Texas	0.9
Bill J Hawthorn	New York	Helen Sarah Hawthorn	New York	0.6
Bill J Hawthorn	New York	Nancy Smith	San Francisco	0
Bill J Hawthorn	New York	David H W Adams	Texas	0
Helen Sarah Hawthorn	New York	Nancy Smith	San Francisco	0
Helen Sarah Hawthorn	New York	David H W Adams	Texas	0
Nancy Smith	San Francisco	David H W Adams	Texas	0

Correspondência (*match*)

- O mesmo número de linhas é gerado com um identificador em cada linha. Linhas similares terão o mesmo identificador de grupo – ClusterID.
- Para determinar se duas linhas são relacionadas, nós especificamos um limite de aceitação. Quanto menor o limite, maior a chance de "falsos positivos".

- Nosso limite é 0.8 Somente 1 par superou este limite.

Name	Address	ClusterID
David W Adams	Texas	1
Bill J Hawthorn	New York	2
Helen Sarah Hawthorn	New York	3
Nancy Smith	San Francisco	4
David H W Adams	Texas	1

Exemplo de Algoritmo de Distância

- EDIT DISTANCE: Para textos de tamanho arbitrário

- Deriva a pontuação de similaridade, calculando o menor custo de se transformar um texto no outro: inserindo deletando ou trocando letras.
- Por exemplo:

Texto 1	Texto 2
College St.	Collage St

  - A estratégia é calcular o custo de (i) transformar a letra **a** em Collage em um **e** e (ii) inserir um ponto depois de St.
  - A fórmula é: 1 – (#edições/tamanho do texto) : 1 − (2/11). O resultado de similaridade é 0.8181.

Algoritmo de distância

- Atividade:
- Calcular a distância entre os nomes abaixo:
  - João Souza Paiva
  - João Sousa e Paiva

Enriquecimento dos Dados

Enriquecimento dos Dados (Data enrichment)

- Agregar valor aos dados
  - Adicionando mais informações
- Aumento da qualidade e usabilidade (relevância)
- Exemplos:
  - Data de atualização/responsável atualização
  - Dados de auditoria
  - Vocabulários de referência (ontologias)
  - Indicador de qualidade

Medição e Monitoramento (Controle)

Medição e Monitoramento (Controle)

- Regras de qualidade
  - Geralmente associadas a uma ou mais dimensões
- Indicadores de valores válidos/inválidos

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

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

Regras de Medição

Nome da regra	regra_Consistencia_CPF
Entrada (CPF)	**Saída**
73692873782	TRUE
627.287.876-23	FALSE (DV INVÁLIDO)
238774609	FALSE (TAMANHO INVÁLIDO)
827.123123-12	TRUE
88298823566	TRUE

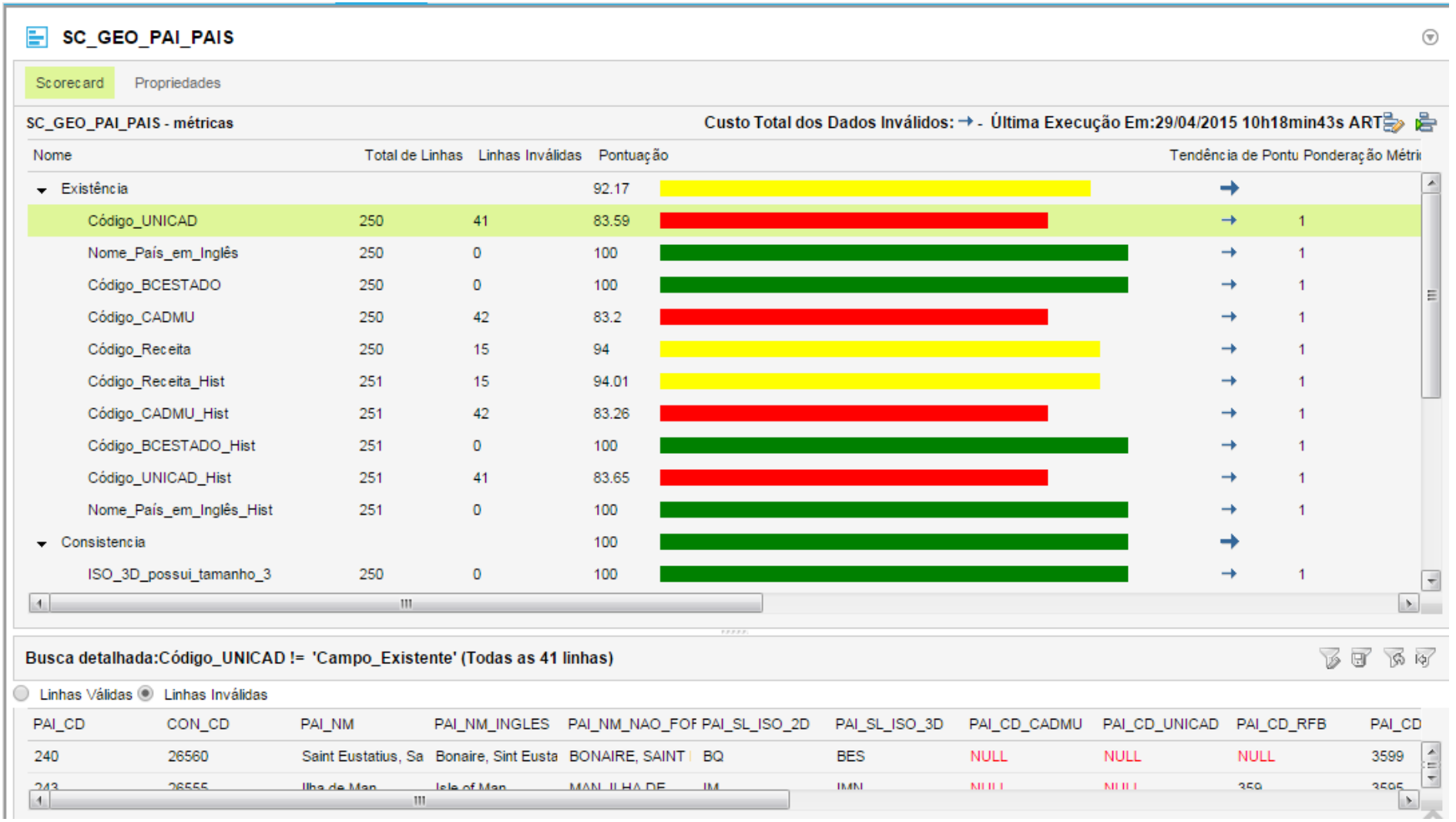
- Entrada: campo com CPF
- Saída: Escala de corretude
  - tamanho inválido/DV inválido/CPF válido
  - 0/1, FALSE/TRUE, etc..

Regras de Medição

Nome da regra		Regra_Precisao_IMC
Entrada		**Saída**
Altura	**Peso**	
182	76	OK
72	76	ERRO
180	150	ATENÇÃO

- Entrada: Altura e Peso
- Saída: Escala de corretude
  - IMC OK/ERRO IMC/ATENÇÃO IMC
  - 0/1/2
  - 0/1
  - IMC OK/IMC fora dos limites
  - Etc...

Monitoramento de Indicadores



SC_GEO_PAI_PAIS

Scorecard | Propriedades

SC_GEO_PAI_PAIS - métricas

Custo Total dos Dados Inválidos: → - Última Execução Em:29/04/2015 10h18min43s ART

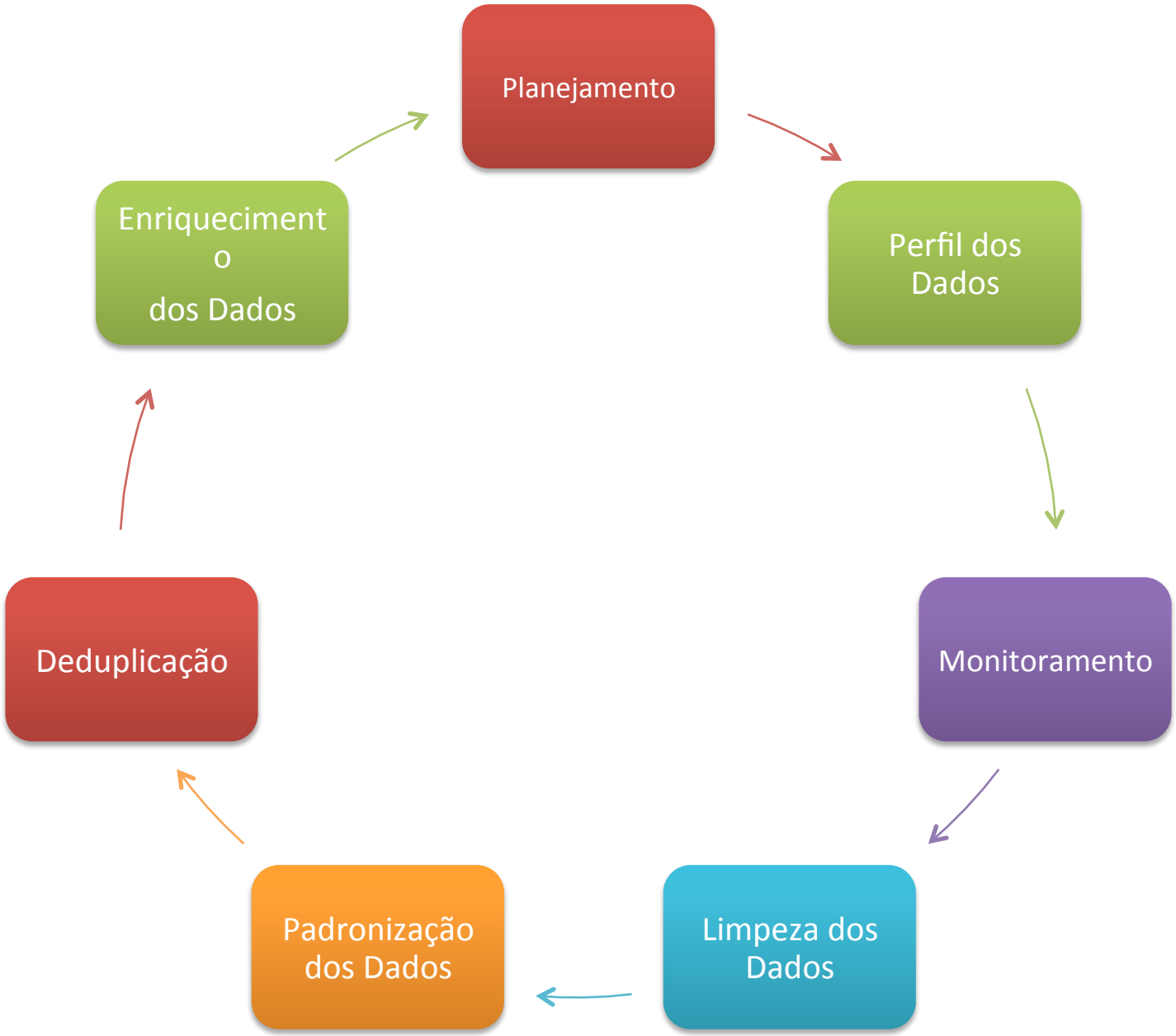
Nome	Total de Linhas	Linhas Inválidas	Pontuação	Tendência de Pontu	Ponderação Métri
Existência			92.17	→	
Código_UNICAD	250	41	83.59	→	1
Nome_País_em_Inglês	250	0	100	→	1
Código_BCESTADO	250	0	100	→	1
Código_CADMU	250	42	83.2	→	1
Código_Receita	250	15	94	→	1
Código_Receita_Hist	251	15	94.01	→	1
Código_CADMU_Hist	251	42	83.26	→	1
Código_BCESTADO_Hist	251	0	100	→	1
Código_UNICAD_Hist	251	41	83.65	→	1
Nome_País_em_Inglês_Hist	251	0	100	→	1
Consistencia			100	→	
ISO_3D_possui_tamanho_3	250	0	100	→	1

Busca detalhada:Código_UNICAD != 'Campo_Existente' (Todas as 41 linhas)

Linhas Válidas | Linhas Inválidas

PAI_CD	CON_CD	PAI_NM	PAI_NM_INGLES	PAI_NM_NAO_FOI	PAI_SL_ISO_2D	PAI_SL_ISO_3D	PAI_CD_CADMU	PAI_CD_UNICAD	PAI_CD_RFB	PAI_CD
240	26560	Saint Eustatius, Sa	Bonaire, Sint Eusta	BONAIRE, SAINT	BQ	BES	NULL	NULL	NULL	3599
243	26555	Ilha de Man	Isle of Man	MAN, ILHA DE	IM	IMN	NULL	NULL	359	3595

Arquitetura de Qualidade de Dados



Avaliando os dados: Perfil

- Atividade:
- Disponível no blackboard Atividade Perfil Dados

"Qualidade não é um ato, é um hábito." Aristóteles