TSW017
Grandes Sistemas de Dados
Projeto e Gerenciamento

Prof. Dr. Jorge Rady de Almeida Jr.
Escola Politécnica da USP

Qualidade de Dados

Qualidade de Dados

- ✓ Medida da concordância entre os dados armazenados e seus valores reais
- ✓ Consequência de má qualidade: impactos nos lucros
- ✓ Exemplos: ataques a locais errados, perda de equipamentos, questões médicas

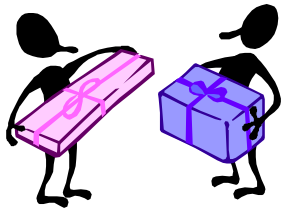
Qualidade de Dados

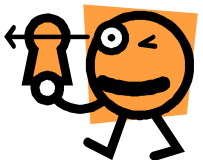
- ✓ Baixa QD: principais causas


entrada inicial


obsolescência

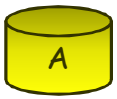

não conhecimento


reestruturações

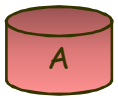

sem monitoramento

Qualidade de Dados

- Baixa QD: principais causas

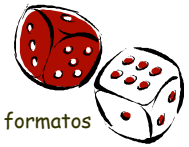


BD1



BD2

dados replicados



formatos

cultura

legados

custos

Qualidade de Dados - Definição

- Data Quality The Field Guide, Thomas Redman

Data are of high quality if they are fit for their intended uses in operations, decision making, and planning. Data are fit for use if they are free of defects and possess desired features. (See Figure 14.2)

Data are of high quality if they are fit for their intended uses in operations, decision making, and planning

Data that are fit for use

free of defects:
- accessible
- accurate
- timely
- complete
- consistent with other sources

etc.

possess desired features:
- relevant
- comprehensive
- proper level of detail
- easy to read
- easy to interpret

etc.

Modelo de Maturidade de Qualidade

- **CMMI** – Modelo Integrado de Maturidade de Capacitação: 5 níveis de maturidade: inicial, repetível, definido, gerenciado e otimizado
- **CMMI** específico p/ DW (Data Warehousing Process Maturity: An Exploratory Study of Factors Influencing User Perceptions, IEEE Transactions on Engineering Management, Sem; Sinha e Ramamurthy – ago/2006): 5 níveis de maturidade de data warehouse
- **IQMM** – Maturidade de Gerenciamento de Qualidade de Informação (Information Quality Management Maturity: Toward the Inteligent Learning Organization, TDAN.com, Larry English - 2004): 5 níveis de maturidade de qualidade de informação

Data Quality and Systems Theory

Data Quality Rules

There are a number of general data quality rules one can deduce from a FCS view of information systems:

DQ1. Unused data cannot remain correct for very long;
DQ2. Data quality in an information system is a function of its use, not its collection;
DQ3. Data quality will, ultimately, be no better than its most stringent use;
DQ4. Data quality problems tend to become worse as the system ages;
DQ5. The less likely some data attribute (element) is to change, the more traumatic it will be when it finally does change;
DQ6. Laws of data quality apply equally to data and metadata (the data about the data).

Dimensões da Qualidade de Dados

Algumas Dimensões Básicas

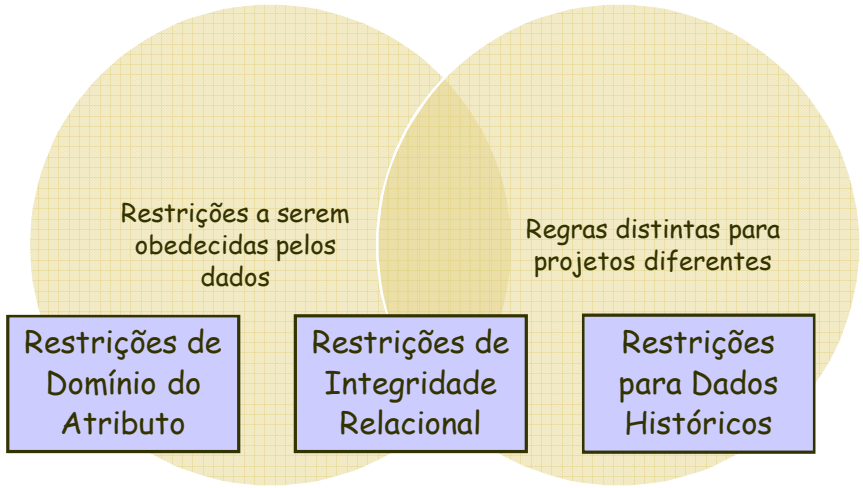
- ✓ Características mensuráveis / avaliáveis de um conjunto de dados → permite determinar a qualidade desse conjunto
- ✓ **Completeza (completeness)** (%): proporção de dados obtidos em comparação com o total possível (100%)
- ✓ **Unicidade Originalidade (uniqueness)** (%): nenhum dado será armazenado mais de uma vez
- ✓ **Temporalidade (timeliness)**: atualização do dado é suficiente para a tarefa a ser realizada (dado representa a realidade no instante de tempo requerido)
- ✓ **Validade (validity)** (%): medida da conformidade com a sintaxe de sua definição (formato, tipo, faixa)
- ✓ **Acurácia (accuracy)** (%): o grau com que um dado descreve corretamente o objeto ou evento
- ✓ **Consistência (consistency)** (%): manutenção de padrão entre duas ou mais representações

Outras Dimensões

- ✓ **Credibilidade / Reputação**: quando o dado pode ser considerado verdadeiro / confiável
- ✓ **Segurança no Acesso**: o acesso ao dado tem as restrições apropriadas
- ✓ **Relevância**: o dado é aplicável e útil para a tarefa a ser realizada

Qualidade de Dados Recomendações

Recomendações



Restrições de Domínio

- Valores **permitidos** para cada atributo
- Valores proibidos: não refletem a realidade do domínio
- Restrições presentes nos modelos / dicionários de dados (porém nem sempre atualizadas)
- **Opcionalidade / Obrigatoriedade**
- **Restrições de Formato**
- **Restrições de Precisão**: casas decimais / unidades de medida

Restrições de Integridade Referencial

Regras de Identidade: chave primária

Regras de Referência: chave estrangeira

Regras de Cardinalidade

- Geralmente presentes nos modelos → nem sempre implementadas → relaxamento em dados com problemas

Restrições para Dados Históricos

Granularidade: mesmo período de representação dos dados

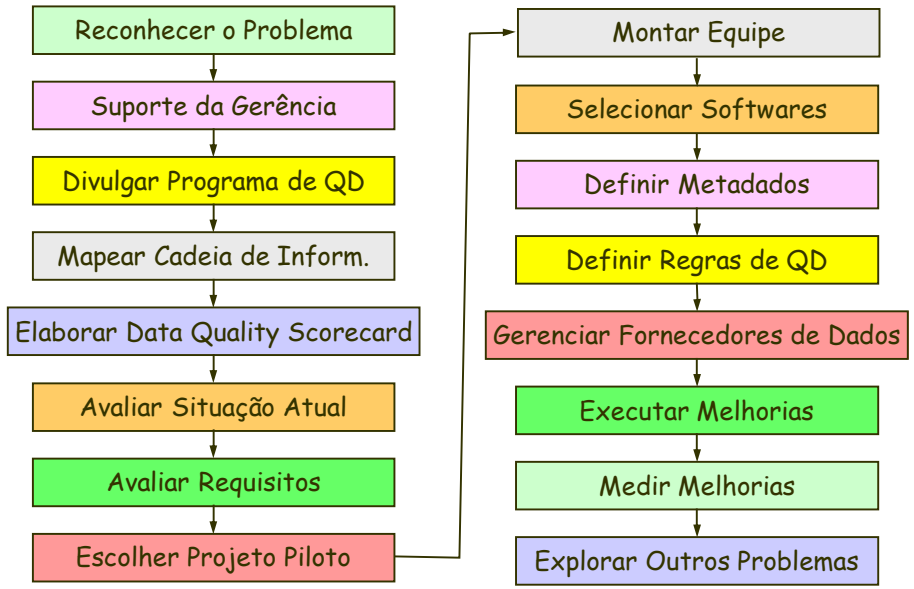
Continuidade: sem lacuna na evolução histórica dos dados

Variação de Valores

- Direção permitida/proibida para mudança de valores
- Magnitude: faixa de valores mínimo e máximo
- Volatilidade: período mínimo obtenção de novos valores

Fluxograma para a Qualidade de Dados

David Loshin - **Enterprise Knowledge Management: The Data Quality Approach**. 2001



Data Provenance

Data Provenance
Proveniência de Dados

- ✓ Descrever pessoas, instituições, entidades e atividades que envolvam, influenciem ou entreguem um dado
- ✓ Motivação
  - ❖ Ter dados mais acreditados
  - ❖ Garantir confiança e qualidade dos dados
  - ❖ Repetibilidade e verificabiliade nos experimentos
- ✓ Objetivos: obter meta informação sobre os dados
  - ❖ Autor e revisor dos dados
  - ❖ Ter a cadeia completa de revisão do dado
  - ❖ No caso de dados integrados, saber a origem de cada parte e a qual processo foram submetidos

Data Provenance
Proveniência de Dados

- ✓ Necessidade de ter a proveniência
  - ❖ Sistemas em geral: origem dos dados, responsável por sua criação
  - ❖ Ciência: como os resultados foram obtidos
  - ❖ Notícias: origem e referências
  - ❖ Leis: documentos de origem
- ✓ Tipos de Proveniência
  - ❖ Qual? Atributos/fatos de origem do dado
  - ❖ Porque? Razão para a seleção da origem do dado
  - ❖ Como? Processo de modificação/produção do dado

Data Provenance
Proveniência de Dados

- ✓ Data Lineage – Linhagem dos Dados
  - ❖ Processo de obter e transformar os dados de sua origem até o destino final
- ✓ Data Provenance
  - ❖ Manter o registro histórico seguido pelo processo de data lineage
- ✓ W3C – World Wide Web Consortium
  - ❖ W3C PROV – Família de normas sobre Provenance

Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.

Proveniência são **informações** sobre **entidades**, **atividades** e **pessoas** envolvidas na **produção** de um **dado** ou **recurso** que servem como **registro** para avaliar sua **qualidade, previsibilidade** e **confiabilidade**.