Instituto Superior de Engenharia

Politécnico de Coimbra

Integração de Dados

CTeSP Tecnologias e Programação de Sistemas de Informação (Cantanhede)

Professor: João Leal

joao.leal@isec.pt

UNIgreen The green European University

PRR Plano de Recuperação e Resiliência

REPÚBLICA PORTUGUESA

Financiado pela União Europeia NextGenerationEU

Técnicas de Validação de Dados

Instituto Superior de Engenharia
Politécnico de Coimbra

- A validação de dados é um processo crucial para garantir que os dados que entram ou são processados num sistema cumprem um conjunto de regras predefinidas, assegurando a sua qualidade e integridade.

- Esta etapa é fundamental para prevenir a introdução de erros e inconsistências que poderiam comprometer a fiabilidade das informações.

Técnicas de Validação de Dados

- As técnicas de validação podem ser aplicadas em diferentes fases do ciclo de vida dos dados, desde a entrada até à integração.

Definição de Regras de Validação

- As regras de validação são condições ou critérios que os dados devem satisfazer para serem considerados válidos.

- Estas regras são derivadas dos requisitos de negócio e das especificações do sistema.

Definição de Regras de Validação

- Podem ser simples, como "o campo 'idade' deve ser um número inteiro", ou complexas, como "o campo 'data de fim' deve ser posterior ao campo 'data de início’”.

- A definição clara e abrangente destas regras é o ponto de partida para qualquer processo de validação eficaz.

Restrições de Integridade (Integrity Constraints) em Bases de Dados

- As bases de dados relacionais oferecem mecanismos robustos para impor a qualidade e consistência dos dados através de restrições de integridade.

- Estas restrições são definidas ao nível do esquema da base de dados e são automaticamente aplicadas pelo sistema de gestão de base de dados (SGBD).

Restrições de Integridade (Integrity Constraints) em Bases de Dados

- **Chave Primária (Primary Key):** Garante a unicidade e não nulidade de um identificador para cada registo.

- **Chave Estrangeira (Foreign Key):** Mantém a integridade referencial, assegurando que os valores numa coluna (chave estrangeira) correspondem a valores existentes numa chave primária de outra tabela.

Restrições de Integridade *(Integrity Constraints)* em Bases de Dados

- **Restrição UNIQUE:** Garante que todos os valores numa coluna (ou conjunto de colunas) são distintos.

- **Restrição NOT NULL:** Impede que uma coluna contenha valores nulos.

- **Restrição CHECK:** Permite definir uma condição que todos os valores numa coluna devem satisfazer (ex: idade > 0).

Validação de Formato (ex: expressões regulares)

- A validação de formato verifica se os dados estão de acordo com um padrão específico.

- É frequentemente utilizada para campos como endereços de e-mail, números de telefone, códigos postais, números de identificação fiscal, etc.

Validação de Formato (ex: expressões regulares)

- As expressões regulares (regex) são uma ferramenta poderosa para definir e aplicar estas regras de formato, permitindo a correspondência de padrões complexos em cadeias de caracteres.

- *Exemplo de Expressão Regular:*

  Para validar um endereço de e-mail simples:

  ^[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}$

Validação de Intervalo e Tipo de Dados

- Validação de Tipo de Dados:
  - Assegura que os dados inseridos correspondem ao tipo de dados esperado (ex: um campo numérico deve conter apenas números, um campo de data deve conter uma data válida).

  - Muitos SGBDs e linguagens de programação impõem isto automaticamente, mas é importante verificar em interfaces de utilizador ou na ingestão de dados de fontes externas.

Validação de Intervalo e Tipo de Dados

- Validação de Intervalo:
  - Verifica se um valor numérico ou de data se encontra dentro de um intervalo aceitável.
  - Por exemplo, a idade de um cliente deve estar entre 0 e 120 anos, ou uma data de encomenda não pode ser no futuro.

Validação Cruzada (Cross-field Validation)

- A validação cruzada envolve a verificação da consistência entre múltiplos campos de dados.

- Esta técnica é mais complexa e verifica relações lógicas entre diferentes atributos. Por exemplo:

  - Se o campo "Estado Civil" é "Casado", então o campo "Nome do Cônjuge" não deve estar vazio.

Validação Cruzada (Cross-field Validation)

- A "Data de Fim" de um projeto deve ser sempre posterior à "Data de Início".

- A soma dos valores de várias colunas deve ser igual a um total especificado.

Data Profiling: Análise e Descoberta de Padrões e Anomalias

- O Data Profiling é o processo de examinar os dados existentes para recolher estatísticas e informações descritivas sobre eles.

- Ajuda a entender a estrutura, conteúdo, qualidade e relações dos dados.

Data Profiling: Análise e Descoberta de Padrões e Anomalias

- É uma técnica exploratória que permite identificar padrões, anomalias, valores em falta, formatos inconsistentes e outras questões de qualidade antes mesmo de definir regras de validação formais.

- O Data Profiling é essencial para a fase de avaliação da DQM e para o planeamento de processos ETL.

Data Profiling: Análise e Descoberta de Padrões e Anomalias

- O Data Profiling pode revelar:
  - Distribuição de valores em colunas.
  - Valores únicos e contagens de valores distintos.
  - Valores mínimos, máximos e médios.
  - Percentagem de valores nulos.
  - Padrões de formato de dados.
  - Dependências funcionais entre colunas.

Limpeza e Transformação de Dados (Data Cleansing/Scrubbing)

- A limpeza de dados (ou data cleansing/scrubbing) é o processo de detetar e corrigir (ou remover) registos incorretos, corrompidos, formatados incorretamente, duplicados ou incompletos de um conjunto de dados.

- É uma etapa fundamental para melhorar a qualidade dos dados e garantir que estes são adequados para análise e utilização em sistemas de informação.

Limpeza e Transformação de Dados (Data Cleansing/Scrubbing)

- A limpeza de dados é frequentemente realizada como parte da fase de Transformação num processo ETL.

Identificação e Tratamento de Duplicados

- Os dados duplicados são um dos problemas mais comuns e podem distorcer significativamente os resultados de qualquer análise. O tratamento de duplicados envolve:

- **Identificação:** Utilizar algoritmos de correspondência exata ou difusa para encontrar registos que representam a mesma entidade.

Identificação e Tratamento de Duplicados

- **Remoção:** Eliminar os registos duplicados, mantendo apenas uma versão (a mais completa ou mais recente).

- **Fusão:** Combinar informações de múltiplos registos duplicados num único registo consolidado.

- **Marcação:** Marcar os registos como duplicados para análise posterior, sem os remover imediatamente.

Correção de Erros e Inconsistências

- Esta etapa foca-se na retificação de dados que não cumprem as regras de validação ou que contêm erros evidentes:

- **Correção Manual:** Para pequenos volumes de dados ou erros complexos que exigem discernimento humano.

- **Correção Automatizada:** Utilização de regras predefinidas ou algoritmos para corrigir erros comuns (ex: correção ortográfica de nomes de cidades, padronização de códigos postais).

Correção de Erros e Inconsistências

- **Validação contra Fontes Externas:** Comparar dados com fontes de dados de referência (ex: bases de dados de endereços válidos, listas de países) para corrigir imprecisões.

Normalização e Padronização de Dados

- A normalização e padronização visam transformar os dados para um formato uniforme e consistente, facilitando a sua comparação e integração:

- **Padronização de Formatos:** Converter datas, horas, moedas e unidades de medida para um formato consistente (ex: todas as datas para AAAA-MM-DD).

Normalização e Padronização de Dados

- **Padronização de Valores:** Unificar a representação de valores que podem ter múltiplas formas (ex: "Masc." e "M" para "Masculino"; "Rua" e "R." para "Rua").

- **Normalização de Texto:** Converter texto para minúsculas/maiúsculas, remover caracteres especiais ou espaços em excesso.

Tratamento de Valores em Falta (Imputation Techniques)

- Valores em falta podem ser problemáticos para análises e modelos. O tratamento pode ser feito de várias formas:

- **Remoção:** Eliminar registos ou colunas com um grande número de valores em falta. Esta abordagem deve ser usada com cautela para evitar perda de informação valiosa.

.

Tratamento de Valores em Falta (Imputation Techniques)

- **Imputação por Constante:** Preencher os valores em falta com uma constante (ex: 0, "Desconhecido", a média ou mediana da coluna). A escolha da constante depende do contexto e do tipo de dados.

- **Imputação por Modelos:** Utilizar modelos estatísticos ou de machine learning para prever e preencher os valores em falta com base nos outros dados disponíveis (ex: regressão, k-NN). Esta é uma abordagem mais sofisticada, mas também mais complexa.

Tratamento de Valores em Falta (Imputation Techniques)

- **Imputação por Valor Anterior/Posterior:** Para dados de séries temporais, preencher com o último valor conhecido ou o próximo valor conhecido.

Monitorização e Relatórios de Qualidade de Dados

- A monitorização contínua e a geração de relatórios são componentes essenciais de um programa eficaz de Gestão da Qualidade de Dados (DQM).

- Não basta apenas limpar os dados uma vez; a qualidade dos dados é um estado dinâmico que requer vigilância constante para garantir que os padrões são mantidos ao longo do tempo e que novos problemas são identificados e resolvidos proativamente.

Métricas de Qualidade de Dados

- Para monitorizar a qualidade dos dados, é fundamental definir e medir métricas de qualidade de dados.

- Estas métricas quantificam o desempenho dos dados em relação às dimensões da qualidade (precisão, completude, atualidade, consistência, unicidade, validade).

- As métricas devem ser relevantes para os objetivos de negócio e facilmente compreensíveis.

Métricas de Qualidade de Dados

- *Exemplos de Métricas:*
  - *Taxa de Completude:* Percentagem de campos preenchidos em relação ao total de campos esperados (ex: % de endereços de e-mail não nulos).
  - *Taxa de Unicidade:* Percentagem de registos únicos em relação ao total de registos (ex: % de clientes sem duplicados).

Métricas de Qualidade de Dados

- *Taxa de Validade:* Percentagem de dados que cumprem as regras de validação (ex: % de datas de nascimento válidas).
- *Taxa de Consistência:* Percentagem de dados que são consistentes entre diferentes fontes ou sistemas (ex: % de nomes de clientes que correspondem em CRM e ERP).
- *Taxa de Atualidade:* Frequência ou atraso na atualização dos dados (ex: idade média dos dados de inventário).
- *Número de Erros por Categoria:* Contagem de erros específicos (ex: número de códigos postais inválidos).

Dashboards e Relatórios de Qualidade

Instituto Superior de Engenharia
Politécnico de Coimbra

- Os dashboards e relatórios de qualidade de dados são ferramentas visuais que apresentam as métricas de qualidade de forma clara e concisa, permitindo que as partes interessadas compreendam rapidamente o estado da qualidade dos dados.

- Estes relatórios devem ser acessíveis e adaptados às necessidades de diferentes públicos, desde analistas de dados a gestores de negócio.

Dashboards e Relatórios de Qualidade

- Características de Dashboards e Relatórios Eficazes:

  - **Visualização Clara:** Utilização de gráficos, tabelas e indicadores de desempenho (KPIs) para representar as métricas.

  - **Atualização Regular:** Os relatórios devem ser gerados e atualizados periodicamente (diariamente, semanalmente, mensalmente) para refletir o estado atual da qualidade dos dados.

Dashboards e Relatórios de Qualidade

- **Capacidade de Detalhe (Drill-down):** Permitir que os utilizadores explorem os dados subjacentes para entender a causa-raiz dos problemas.

- **Alertas e Notificações:** Configurar alertas automáticos para quando as métricas de qualidade caem abaixo de um limiar aceitável.

- **Contexto de Negócio:** Relacionar as métricas de qualidade de dados com o impacto nos processos de negócio e nos objetivos estratégicos.

Processos de Auditoria de Dados

- Os processos de auditoria de dados são avaliações sistemáticas e independentes da qualidade dos dados e dos processos de DQM.

- As auditorias ajudam a verificar a conformidade com as políticas de qualidade de dados, identificar vulnerabilidades e recomendar melhorias.

Processos de Auditoria de Dados

- Podem ser realizadas interna ou externamente e são cruciais para manter a confiança nos dados e nos sistemas que os utilizam.

- Objetivos da Auditoria de Dados:
  - **Verificar a Conformidade:** Assegurar que os dados e os processos de gestão de dados cumprem as políticas internas e os requisitos regulamentares.

Processos de Auditoria de Dados

- **Identificar Riscos:** Detetar potenciais riscos associados à má qualidade de dados, como riscos financeiros, operacionais ou de reputação.

- **Avaliar a Eficácia:** Medir a eficácia das iniciativas de DQM e identificar áreas para otimização.

- **Fornecer Recomendações:** Propor ações corretivas e preventivas para melhorar a qualidade dos dados e os processos de gestão.

Exemplos

- *Possíveis resoluções em Python:*

```python
#Exemplo 1 - Tratar dados em falta

import pandas as pd

clientes = pd.read_csv("clientes.csv")

# Substituir nulos no campo email
clientes["email"] = clientes["email"].fillna("email_desconhecido")

# Preencher valores em falta na idade com a média
clientes["idade"] = clientes["idade"].fillna(clientes["idade"].mean())

print(clientes)
```

Exemplos

- *Possíveis resoluções em Python:*

```python
#Exemplo 2 - Detetar e remover duplicados

clientes = clientes.drop_duplicates(subset=["nome", "email"])


#Exemplo 3 - Corrigir formatos e tipos

# Converter nomes para maiúsculas
clientes["nome"] = clientes["nome"].str.upper()

# Converter idades para inteiro
clientes["idade"] = clientes["idade"].astype("Int64")
```

Exemplos

- *Possíveis resoluções em Python:*

```
#Exemplo 4 - Validar integridade entre tabelas
vendas = pd.read_csv("vendas.csv")

# Filtrar vendas com id_cliente existente
vendas_validas = vendas[vendas["id_cliente"].isin(clientes["id_cliente"])]

print(vendas_validas)
```