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

Propostas de Resolução


```python
import pandas as pd

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

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

# Substituir emails vazios
clientes["email"] = clientes["email"].fillna("email_desconhecido")

# Corrigir idades inválidas
clientes.loc[(clientes["idade"] < 0) | (clientes["idade"] > 120), "idade"] = None

# Guardar ficheiro limpo
clientes.to_csv("clientes_final.csv", index=False)

print(clientes)
```

Propostas de Resolução


```python
clientes = pd.read_csv("clientes_final.csv")
vendas = pd.read_csv("vendas.csv")

# Validar integridade referencial
vendas_validas = vendas[vendas["id_cliente"].isin(clientes["id_cliente"])]

# Total de vendas válidas
total = vendas_validas["valor"].sum()

vendas_validas.to_csv("vendas_validas.csv", index=False)

print(vendas_validas)
print(f"Total de vendas válidas: {total}")
```

Qualidade e Consistência de Dados

- Qualidade de dados e consistência de dados são fundamentais para o sucesso de qualquer processo de integração, análise ou tomada de decisão nas organizações.

- Dados de má qualidade podem gerar erros, distorções em análises e impactos nos negócios.

Qualidade de Dados

- A Qualidade de Dados refere-se ao grau em que os dados são adequados para o seu uso pretendido.

- A qualidade dos dados é um conceito multifacetado, abrangendo diversas dimensões que, em conjunto, determinam a sua utilidade e fiabilidade para a tomada de decisões e operações empresariais.

Qualidade de Dados

- A qualidade de dados mede o grau em que os dados são:

  - **Precisos** – refletem corretamente a realidade.
  - **Completos** – sem valores em falta.
  - **Consistentes** – coerentes entre diferentes fontes.
  - **Atualizados** – representam o estado mais recente.
  - **Únicos** – sem duplicações.

Consistência de Dados

- A Consistência de Dados é uma dimensão crítica da qualidade de dados que garante que os dados permaneçam uniformes e sem contradições em diferentes sistemas, bases de dados ou ao longo do tempo.

- Em ambientes de integração de dados, a consistência assegura que a mesma informação, quando replicada ou distribuída, mantém o mesmo valor e formato em todas as suas ocorrências.

Consistência de Dados

- A inconsistência pode surgir de múltiplas fontes de dados, diferentes formatos de armazenamento ou regras de negócio divergentes

Importância da Qualidade e Consistência de Dados em Processos de Integração (ETL)

- Nos processos de Extração, Transformação e Carregamento (ETL), a qualidade e consistência dos dados são fundamentais para o sucesso.

- O objetivo principal do ETL é consolidar dados de diversas fontes num repositório central (como um *data warehouse* ou *data lake*) para análise e *reporting*.

Importância da Qualidade e Consistência de Dados em Processos de Integração (ETL)

- Se os dados extraídos forem de má qualidade ou inconsistentes, os resultados da transformação e do carregamento serão comprometidos, levando a análises erróneas e decisões inadequadas.

- A importância pode ser resumida em:
  - **Fiabilidade das Decisões:** Dados de alta qualidade fornecem uma base sólida para decisões estratégicas e operacionais.

Importância da Qualidade e Consistência de Dados em Processos de Integração (ETL)

- **Eficiência Operacional:** Reduz a necessidade de retrabalho e correção manual de dados, otimizando os processos.

- **Conformidade Regulatória:** Ajuda as organizações a cumprir regulamentos de privacidade e integridade de dados.

- **Confiança do Cliente:** Melhora a experiência do cliente através de informações precisas e personalizadas.

- **Redução de Custos:** Evita custos associados a erros, multas regulatórias e perda de oportunidades de negócio.

Consistência de Dados

- A inconsistência pode surgir de múltiplas fontes de dados, diferentes formatos de armazenamento ou regras de negócio divergentes

Impacto de Dados de Má Qualidade nas Organizações

- Dados de má qualidade podem ter um impacto devastador nas organizações, afetando diversas áreas.

- Os problemas podem variar desde erros operacionais e ineficiências até perdas financeiras significativas e danos à reputação.

Impacto de Dados de Má Qualidade nas Organizações

- Principais impactos incluem:
  - **Decisões Erradas:** Análises baseadas em dados incorretos levam a estratégias falhas.
  - **Perda de Receita:** Oportunidades de vendas perdidas devido a informações incompletas ou imprecisas sobre clientes.
  - **Aumento de Custos:** Gastos adicionais com a correção manual de dados, armazenamento de dados redundantes e ineficiências operacionais.

Impacto de Dados de Má Qualidade nas Organizações

- **Insatisfação do Cliente:** Erros em faturas, comunicações ou serviços devido a dados de cliente incorretos.

- **Danos à Reputação:** Perda de confiança de clientes e parceiros devido a falhas na gestão de dados. Incapacidade de cumprir requisitos regulatórios, resultando em multas e sanções.

- **Não Conformidade**: Incapacidade de cumprir requisitos regulatórios, resultando em multas e sanções.

Dimensões e Problemas Comuns de Qualidade

Tipo de Problema	Descrição	Exemplo
Dados em falta	Campos nulos ou vazios	email = ""
Duplicação	Registos repetidos	Duas linhas com o mesmo cliente
Inconsistência	Formatos diferentes	Lisboa vs lisboa
Erro de tipo	Dados com tipo incorreto	idade = "vinte"
Valor inválido	Fora dos limites esperados	idade = -5
Violação de integridade	Chaves que não existem na tabela relacionada	id_cliente sem correspondência

Dimensões da Qualidade de Dados

- A qualidade dos dados não é um conceito singular, mas sim um conjunto de características ou dimensões que, em conjunto, determinam a sua adequação para um determinado propósito.

- Compreender estas dimensões é crucial para avaliar, gerir e melhorar a qualidade dos dados em qualquer organização.

Precisão (*Accuracy*)

- A Precisão refere-se ao grau em que os dados representam corretamente a realidade do objeto ou evento que descrevem.

- Dados precisos são livres de erros e refletem os valores verdadeiros.

Precisão (*Accuracy*)

- Por exemplo, um registo de cliente com o endereço correto e o número de telefone atualizado é considerado preciso.

- A imprecisão pode surgir de erros de entrada de dados, medições incorretas ou dados desatualizados.

Completude (*Completeness*)

- A Completude indica a extensão em que todos os dados esperados estão presentes e não há valores em falta.

- Dados incompletos podem levar a análises tendenciosas ou impossibilitar a execução de processos de negócio.

Completude (*Completeness*)

- Por exemplo, um formulário de registo de cliente que exige um endereço de e-mail, mas esse campo está vazio, demonstra falta de completude.

- A completude é frequentemente medida pela percentagem de valores não nulos em campos obrigatórios.

Atualidade (*Timeliness*)

- A Atualidade (ou Timeliness) refere-se à medida em que os dados estão disponíveis e são relevantes para o momento em que são necessários.

- Dados podem ser precisos e completos, mas se estiverem desatualizados, perdem a sua utilidade.

Atualidade (*Timeliness*)

- Por exemplo, informações de inventário que não refletem as vendas mais recentes não são atuais.

- A atualidade é crítica em sistemas que dependem de informações em tempo real ou quase real para a tomada de decisões.

Consistência (*Consistency*)

- A Consistência garante que os dados são uniformes e não contraditórios em diferentes sistemas, bases de dados ou ao longo do tempo.

- Dados consistentes seguem as mesmas regras de negócio e definições em todas as suas ocorrências.

Consistência (*Consistency*)

- Por exemplo, se o nome de um cliente é "João Silva" num sistema e "Silva, João" noutro, há uma inconsistência de formato.

- A consistência também se aplica a regras de negócio, como garantir que a data de nascimento de um cliente seja anterior à data atual.

Unicidade (Uniqueness)

- A Unicidade assegura que não existem registos duplicados no conjunto de dados.

- Dados duplicados podem distorcer análises, inflacionar contagens e levar a ineficiências operacionais.

Unicidade (Uniqueness)

- Por exemplo, ter duas entradas separadas para o mesmo cliente na base de dados de CRM viola a unicidade.

- A identificação e eliminação de duplicados é um aspeto fundamental da limpeza de dados.

Nota: CRM significa Customer Relationship Management, ou em português, Gestão de Relacionamento com o Cliente. Portanto, uma base de dados de CRM é uma base de dados que armazena todas as informações sobre clientes e interações que uma empresa tem com eles.

Validade (*Validity*)

- A Validade refere-se à conformidade dos dados com as regras de negócio, formatos e domínios predefinidos.

- Dados válidos respeitam as restrições impostas para garantir a sua integridade.

Validade (*Validity*)

- Exemplos incluem garantir que um campo de idade contém apenas números inteiros positivos, que um código postal segue um formato específico ou que um valor está dentro de um intervalo aceitável (e.g., uma percentagem entre 0 e 100).

Relevância (*Relevance*)

- A Relevância avalia se os dados são apropriados e úteis para o propósito específico para o qual estão a ser recolhidos e utilizados.

- Dados podem ser precisos, completos e consistentes, mas se não forem relevantes para a questão de negócio em causa, a sua utilidade é limitada.

Relevância (*Relevance*)

- Por exemplo, para uma análise de vendas, o histórico de compras do cliente é relevante, mas a cor do seu carro pode não ser.

Problemas Comuns de Qualidade de Dados

- Apesar da crescente importância da gestão da qualidade de dados, as organizações continuam a enfrentar uma série de desafios que comprometem a fiabilidade das suas informações.

- Compreender os problemas comuns de qualidade de dados é o primeiro passo para desenvolver estratégias eficazes de prevenção e correção.

Problemas Comuns de Qualidade de Dados

- Estes problemas podem surgir em qualquer fase do ciclo de vida dos dados, desde a sua recolha até ao seu armazenamento e utilização.

Dados Duplicados

- Os dados duplicados são um dos problemas mais prevalentes e dispendiosos de qualidade de dados.

- Ocorrem quando o mesmo registo ou entidade é representado múltiplas vezes num sistema ou em sistemas diferentes.

- As duplicações podem surgir devido a erros de entrada de dados, fusões de bases de dados, ou a falta de identificadores únicos robustos.

Dados Duplicados

- Por exemplo, um cliente pode ter várias entradas com pequenas variações no nome ou endereço.

- Os dados duplicados levam a contagens incorretas, relatórios enviesados e desperdício de recursos.

Dados Incorretos/Imprecisos

- Dados incorretos ou imprecisos são aqueles que não refletem a realidade ou contêm erros factuais.

- Podem ser resultado de erros humanos na entrada de dados, falhas em sistemas de recolha automatizada, ou desatualização.

Dados Incorretos/Imprecisos

- Exemplos incluem um número de telefone errado, um endereço postal inexistente, ou um valor de venda registado incorretamente.

- A imprecisão compromete a confiança nos dados e pode levar a decisões erradas com consequências significativas.

Dados em Falta (*Missing Data*)

- Os dados em falta (ou missing data) referem-se a campos ou atributos que não contêm qualquer valor quando deveriam.

- A ausência de dados pode ser intencional (por exemplo, um campo opcional não preenchido) ou não intencional (por exemplo, erro no processo de recolha, falha de sistema).

Dados em Falta (*Missing Data*)

- A gestão de dados em falta é crucial, pois pode distorcer análises estatísticas, reduzir a eficácia de modelos preditivos e impedir a conclusão de processos de negócio.

- Técnicas como a imputação de dados (*data imputation*) são frequentemente utilizadas para lidar com este problema.

Dados Desatualizados

- Dados desatualizados são aqueles que já não são válidos ou relevantes devido à passagem do tempo.

- A informação que era precisa num determinado momento pode tornar-se obsoleta rapidamente, especialmente em ambientes dinâmicos.

Dados Desatualizados

- Por exemplo, um endereço de cliente que mudou, um preço de produto que foi atualizado, ou um estado de inventário que não reflete as transações mais recentes.

- A falta de atualidade pode levar a ineficiências operacionais e a uma má experiência do cliente.

Inconsistências de Formato e Representação

- As inconsistências de formato e representação ocorrem quando os mesmos dados são armazenados ou apresentados de diferentes maneiras em vários sistemas ou mesmo dentro do mesmo sistema.

- Por exemplo, datas formatadas como "DD-MM-AAAA" num sistema e "MM/DD/YY" noutro, ou nomes de cidades escritos de forma diferente (por exemplo, "Lisboa" vs. "LISBOA").

Inconsistências de Formato e Representação

- Estas inconsistências dificultam a integração de dados e a realização de análises unificadas, exigindo processos complexos de padronização.

Dados Ambíguos ou Mal Interpretados

- Dados ambíguos ou mal interpretados surgem quando a informação não é clara ou pode ser entendida de diferentes formas.

- Isto pode ser devido a definições de dados vagas, falta de metadados adequados, ou uso inconsistente de terminologia.

Dados Ambíguos ou Mal Interpretados

- Por exemplo, um campo chamado "Código" pode referir-se a um código de produto, código de cliente ou código de região, sem uma especificação clara.

- A ambiguidade leva a erros de interpretação e a uma utilização ineficaz dos dados.

Gestão da Qualidade de Dados (Data Quality Management - DQM)

- A Gestão da Qualidade de Dados (DQM) é um conjunto abrangente de processos, políticas, padrões e tecnologias implementadas por uma organização para garantir que os dados sejam adequados para o seu uso pretendido.

- O objetivo da DQM é melhorar e manter a qualidade dos dados ao longo de todo o seu ciclo de vida, desde a criação até ao arquivo, garantindo que sejam precisos, completos, atuais, consistentes, únicos e válidos.

Princípios e Ciclo de Vida da DQM

- A DQM não é um evento único, mas um processo contínuo e iterativo que segue um ciclo de vida bem definido.

- Os princípios fundamentais da DQM incluem a responsabilidade pelos dados, a medição contínua da qualidade, a melhoria proativa e a integração da qualidade de dados nos processos de negócio.

Princípios e Ciclo de Vida da DQM

- O ciclo de vida da DQM pode ser dividido em várias fases:

1. **Definição:** Identificar os requisitos de qualidade de dados com base nas necessidades de negócio e definir métricas de qualidade.

2. **Avaliação:** Medir a qualidade dos dados existentes em relação às métricas definidas, identificando problemas e suas causas-raiz (Data Profiling).

Princípios e Ciclo de Vida da DQM

3. **Melhoria:** Implementar ações corretivas para resolver os problemas de qualidade de dados, como limpeza, padronização e enriquecimento.

4. **Monitorização:** Acompanhar continuamente a qualidade dos dados para garantir que os padrões são mantidos e para detetar novos problemas.

5. **Governança:** Estabelecer políticas, papéis e responsabilidades para a gestão da qualidade de dados em toda a organização.

Estratégias para Melhorar a Qualidade de Dados

- A melhoria da qualidade de dados requer uma abordagem multifacetada, combinando estratégias técnicas e organizacionais:
  - **Definição de Padrões de Dados:** Estabelecer formatos, domínios de valores e regras de negócio claros para todos os dados.
  - **Validação na Fonte:** Implementar validações no ponto de entrada dos dados para prevenir a introdução de erros.
  - **Limpeza de Dados (*Data Cleansing*):** Processos para identificar e corrigir dados incorretos, incompletos, duplicados ou inconsistentes.

Estratégias para Melhorar a Qualidade de Dados

- **Padronização e Normalização:** Transformar dados para um formato uniforme e consistente.
- **Enriquecimento de Dados:** Adicionar informações de fontes externas para tornar os dados mais completos e úteis.
- **Governança de Dados:** Criar uma estrutura organizacional com papéis e responsabilidades claras para a gestão de dados, incluindo a qualidade.
- **Formação e Consciencialização:** Educar os utilizadores sobre a importância da qualidade de dados e as melhores práticas.

Ferramentas e Tecnologias para DQM

- Existem diversas ferramentas e tecnologias que suportam os processos de DQM, desde soluções empresariais abrangentes até bibliotecas de código aberto:
  - **Ferramentas de Data Profiling:** Analisam os dados para descobrir padrões, anomalias e identificar problemas de qualidade (*ex: Talend Open Studio, Atlan*).
  - **Ferramentas de Data Cleansing e Padronização:** Automatizam a correção e transformação de dados (*ex: Informatica Data Quality, OpenRefine*).

Ferramentas e Tecnologias para DQM

- **Ferramentas de Gestão de Metadados:** Armazenam informações sobre os dados, incluindo definições, regras de negócio e linhagem (*ex: Collibra, Alation*).

- **Ferramentas de Governança de Dados:** Suportam a implementação de políticas e a gestão de acesso e conformidade.

- **Linguagens de Programação:** Python (com bibliotecas como *Pandas, Great Expectations*) e SQL são amplamente utilizadas para desenvolver rotinas personalizadas de validação e limpeza de dados.