

PÓS-GRADUAÇÃO

Integração e fluxo de dados



PÓS-GRADUAÇÃO

Conformação de Dados

Bloco 1

Thiago Salhab Alves





► **Conformação de Dados**

Objetivos

- Compreender as definições e conceitos de conformação de dados.
- Aprender a tratar problemas como redundância e conflito de valores.
- Aprender as técnicas de consolidação de dados.



► **Conformação de Dados**

Conformação de Dados

- A conformação ou integração de dados diz respeito a criação de dimensões e instâncias de fatos configuradas, combinando as melhores informações de várias fontes de dados em uma visão mais abrangente.
- Para serem conformados, os dados recebidos precisam ser estruturalmente idênticos, filtrados de registros inválidos, padronizados em termos de conteúdo e não duplicados.



► Conformação de Dados

- De acordo com Han e Kamber (2006), a conformação de dados é a combinação de dados de diferentes fontes em uma base de dados única e coerente.
- Problemas:
- Identificação de entidades: como garantir que um atributo presente em duas fontes tem o mesmo significado?
- Exemplo: `costumer_id`.



► Conformação de Dados

- Que valores os campos podem assumir?
- Exemplo: campo sexo pode ter valores H/M ou M/F.
- Redundância
- **Dados duplicados.**
- Vários atributos podem ser obtidos a partir de um outro atributo ou conjunto deles.
- Pode-se tentar identificar redundância a partir de análises de correlação.

► Conformação de Dados

- Segundo Han e Kamber (2006), a transformação de dados consiste em transformar ou consolidar os dados em um formato mais adequado para o data warehouse. Vários tipos de transformações são possíveis:
 - Suavização: visa eliminar ruídos.
 - Agregação: operações de resumo ou agregação são realizadas.
 - Exemplo: dados de horários de chuva (em mm) são resumidos em um único atributo correspondente ao total acumulado em um dia.
 - Generalização: consiste em substituir dados de baixo nível por dados de alto nível.
 - Exemplo: idade passa a ser jovem, adulto ou idoso.



► **Conformação de Dados**

- Normalização: atributos são escalados para um novo intervalo mais adequado a ser usado.
- Exemplo: $[0;1]$ ou $[-1;1]$.
- Construção de atributos: novos atributos são construídos e adicionados ao conjunto de dados, a fim de auxiliar o data warehouse.
- Exemplo: criar um atributo volume, a partir dos atributos base, volume e profundidade.



► Conformação de Dados

- De acordo com Han e Kamber (2006), a redução dos dados é uma técnica que busca obter uma representação significativamente menor dos dados (em volume), mas que mantenha a integridade dos dados originais. Algumas técnicas de redução de dados:
 - Seleção de atributos: irrelevantes, pouco relevantes ou redundantes, são removidos da base de dados.
 - Redução de dimensão: técnicas de codificação são usadas para reduzir a dimensão (número de atributos) dos dados.

► Conformação de Dados

- Redução de número: dados são substituídos ou estimados por representações alternativas e menores.
- Discretização de atributos:
 - É usado para reduzir o número de valores para um dado atributo contínuo.
 - Domínio do atributo é dividido em intervalos.
 - A cada intervalo é associado um rótulo.
 - Após a discretização, pode substituir os dados por categorias mais genéricas, que facilitam a interpretação dos dados.
 - Exemplo: atributo numérico *idade* substituído por *jovem*, *adulto* e *idoso*.

PÓS-GRADUAÇÃO

Conformação de Dados

Bloco 2

Thiago Salhab Alves






► Conformação de Dados

- A conformidade de atributos descritivos em várias fontes de dados, vários data marts que participam de um *data warehouse* distribuído, é uma das principais etapas de desenvolvimento para o arquiteto do *data warehouse* e a equipe ETL.
- As dimensões conformadas são extremamente importantes para o *data warehouse*. Sem uma adesão restrita às dimensões conformadas, o data warehouse não pode funcionar como um todo integrado.



► **Conformação de Dados**

- A correspondência ou não duplicação consiste na eliminação de registros padronizados duplicados.
 - Em alguns casos, os dados duplicados podem ser facilmente detectados por meio da aparência de valores idênticos em algumas colunas-chave, como número de telefone ou cartão de crédito.
- 

PÓS-GRADUAÇÃO

Teoria em Prática

Bloco 3

Thiago Salhab Alves





► Conformação de Dados

- Uma empresa nacional de revenda de cosméticos está enfrentando alguns problemas financeiros. Dado o grande volume de produtos lançados pelo setor de cosméticos, a empresa está tendo dificuldades em acompanhar a demanda por produtos e respectivos lançamentos, o que, muitas vezes, acaba por comprometer o resultado financeiro por investir em produtos com baixa procura. Outro problema são os produtos que possuem prazo de validade curto, que acabam por vencer e não podem ser trocados pelos fornecedores.



► Conformação de Dados

- Hoje, a empresa conta com um sistema de vendas e controle de estoque, com banco de dados relacional e um processo de marketing pelas redes sociais, porém está tendo dificuldades para a tomada de decisões relacionado ao que seu público-alvo realmente consome e, assim, evitar gastos desnecessários. Após o processo de extração e limpeza de dados do sistema de vendas e controle, de dados e do marketing das redes social, constatou-se que os dados necessitavam passar por uma consolidação. Como podemos auxiliar a organização no processo de consolidação dos dados?

Resposta: para que isso seja possível, os dados devem ser estruturalmente idênticos, filtrados, padronizados e não duplicados. Para isso, deve-se criar uma base de dados única, sem redundância e conflito de valores.

Dica do Professor

Bloco 4

Thiago Salhab Alves





► Conformação de Dados

Indicação de Leitura de Capítulo de Livro da Biblioteca Virtual:

Leitura do capítulo 4 (parte 4) do livro:

- KIMBALL, L., R.; CASERTA, J. **The Data Warehouse ETL Toolkit**: Practical Techniques for Extracting, Cleaning, Conforming, and Data Delivering Data. Indianapolis: Wiley Publishing, 2009.



► Referências Bibliográficas

HAN, J.; KAMBER, M. **Data Mining:** Concepts and Techniques. Elsevier, 2006.

KIMBALL, R.; CASERTA, J. **The Data Warehouse ETL Toolkit:** Practical Techniques for Extracting, Cleaning, Conforming, and Data Delivering Data. Indianapolis: Wiley Publishing, 2009.

