

PÓS-GRADUAÇÃO

Integração e fluxo de dados



Extração de dados

Bloco 1

Thiago Salhab Alves





► Extração de dados

Objetivos

- Compreender as definições e conceitos de extração de dados.
- Aprender a criar um mapa de dados lógico, que documente a relação entre os campos de origem e os campos de destino da tabela.
- Conhecer os diferentes tipos de fontes de dados de origem que devem ser extraídos.



► Extração de dados

Extração de dados

- De acordo com Kimball e Caserta (2009), o primeiro passo da integração é extrair com sucesso dados dos principais sistemas de origem.
- Cada fonte de dados possui um conjunto distinto de características que precisam ser gerenciadas para extrair de forma efetiva os dados para o processo de ETL.



► Extração de dados

- De acordo com Kimball e Caserta (2009), antes de construir um sistema de extração, é necessário criar um mapa de dados lógico, que documente a relação entre os campos de origem e os campos de destino da tabela.

► Extração de dados

- Antes de iniciar qualquer desenvolvimento de ETL físico, certifique-se de que as etapas seguintes são atendidas:
 - Tenha um plano.
 - Identifique as fontes de dados candidatas.
 - Analise os sistema de origem com uma ferramenta de criação de perfil de dados.
 - Receba as instruções para a linhagem dos dados e regras de negócio.
 - Receba as instruções do modelo de dados do *data warehouse*.
 - Valide cálculos e fórmulas.

► Extração de dados

- Segundo Kimball e Caserta (2009), antes de conhecer os detalhes das várias fontes de dados que serão extraídos, é necessário conhecer o documento de mapeamento de dados lógicos. O mapeamento de dados lógicos é apresentado com os seguintes elementos:
 - Nome da tabela de destino.
 - Nome da coluna de destino.
 - Tipo de tabela.
 - Banco de dados de origem.
 - Nome da tabela de origem.
 - Nome da coluna de origem.
 - Transformação.



► Extração de dados

- Na fase de descoberta de dados, a equipe de ETL deve aprofundar mais na descoberta dos dados para determinar cada sistema, tabela e atributo de origem necessário para carregar o *data warehouse*.



► Extração de dados

- Segundo Kimball e Caserta (2009), pode-se encontrar dados de diferentes fontes de dados que necessitam ser usadas no *data warehouse*, levando ao processo de integração de diferentes fontes de dados.




► Extração de dados

- São atividades para integração de dados:
 - Identificar os sistemas de origem.
 - Compreender os sistemas de origem.
 - Criar e registrar a lógica de correspondência.
 - Estabelecer as regras de negócio de atributos não chave.
 - Carregar dimensão conformada.



► Extração de dados

- De acordo com Kimball e Caserta (2009), cada fonte de dados pode estar em um Sistema Gerenciador de Banco de Dados (SGBD) diferente e em uma plataforma diferente.
 - Em um projeto de *data warehouse* pode haver a necessidade de se comunicar com sistemas de diferentes origens. O ODBC (*Open Database Connectivity*) foi criado para permitir que os usuários acessassem bancos de dados a partir de seus aplicativos.
- 

Extração de dados

Bloco 2

Thiago Salhab Alves



► Sistemas e tipos de arquivos para extração de dados

- Vários são os sistemas e tipos de dados que podem ser extraídos.
- Alguns sistemas e tipos de arquivos que merecem atenção:
 - Dados armazenados em mainframes: apresentam caracteres no formato EBCDIC que devem ser convertido para ASCII:
 - Arquivos simples.
 - Documentos XML.
 - Fonte de Log da Web.
 - ERP.
 - Carregamento inicial.

► Sistemas e tipos de arquivos para extração de dados

- Considere os seguintes pontos, por Kimball e Caserta (2009), sobre o processo de extração:
 - Restringir colunas indexadas.
 - Recupere os dados que necessita.
 - Utilize a cláusula DISTINCT com moderação.
 - Utilize o operador SET com moderação.
 - Utilize HINT conforme necessário.
 - Evite NOT.
 - Evite funções em sua cláusula WHERE.

PÓS-GRADUAÇÃO

Teoria em Prática

Bloco 3

Thiago Salhab Alves





► Teoria em Prática

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 desses produtos, 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.



► Teoria em Prática

- Hoje, a empresa conta com um sistema de vendas e controle de estoque, com um banco de dados relacional e com um processo de marketing pelas redes sociais. Porém, está tendo dificuldades para a tomada de decisões relacionada ao que o seu público-alvo realmente consome, a fim de evitar gastos desnecessários. Como podemos organizar um processo de extração de dados do sistema de vendas e controle de dados, e do marketing das redes social para poder criar um *data warehouse*?



► Teoria em Prática

- R: o primeiro passo seria criar um mapa de dados lógico, identificando os dados. Para o processo de integração, deve-se identificar o sistema de origem e criar regras de correspondência e carregar as dimensões conformadas.

Dica do Professor

Bloco 4

Thiago Salhab Alves





► Dica do Professor

Indicação de leitura do capítulo 3 de Kimball e Caserta (2009), disponível na Biblioteca Virtual:

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

