

Arquitetura de Data Warehouse e Data Marts

Aula 7: Processo ETL – Transformação e Carga dos Dados das tabelas Dimensões

Apresentação

Nesta aula serão apresentadas: as etapas de transformação dos dados e carga dos dados nas tabelas definitivas; a implementação da validação dos dados nas tabelas temporárias utilizando a ferramenta de ETL Pentaho Data Integration (PDI); e a carga dos dados nas dimensões do modelo de dados dimensional.

Objetivos

- Descrever a etapa de transformação dos dados e aplicação de regras de negócio;
- Examinar os tipos de transformações aplicados aos dados no processo de ETL;
- Demonstrar a etapa de carga dos dados nas tabelas dimensões do modelo de dados dimensional.

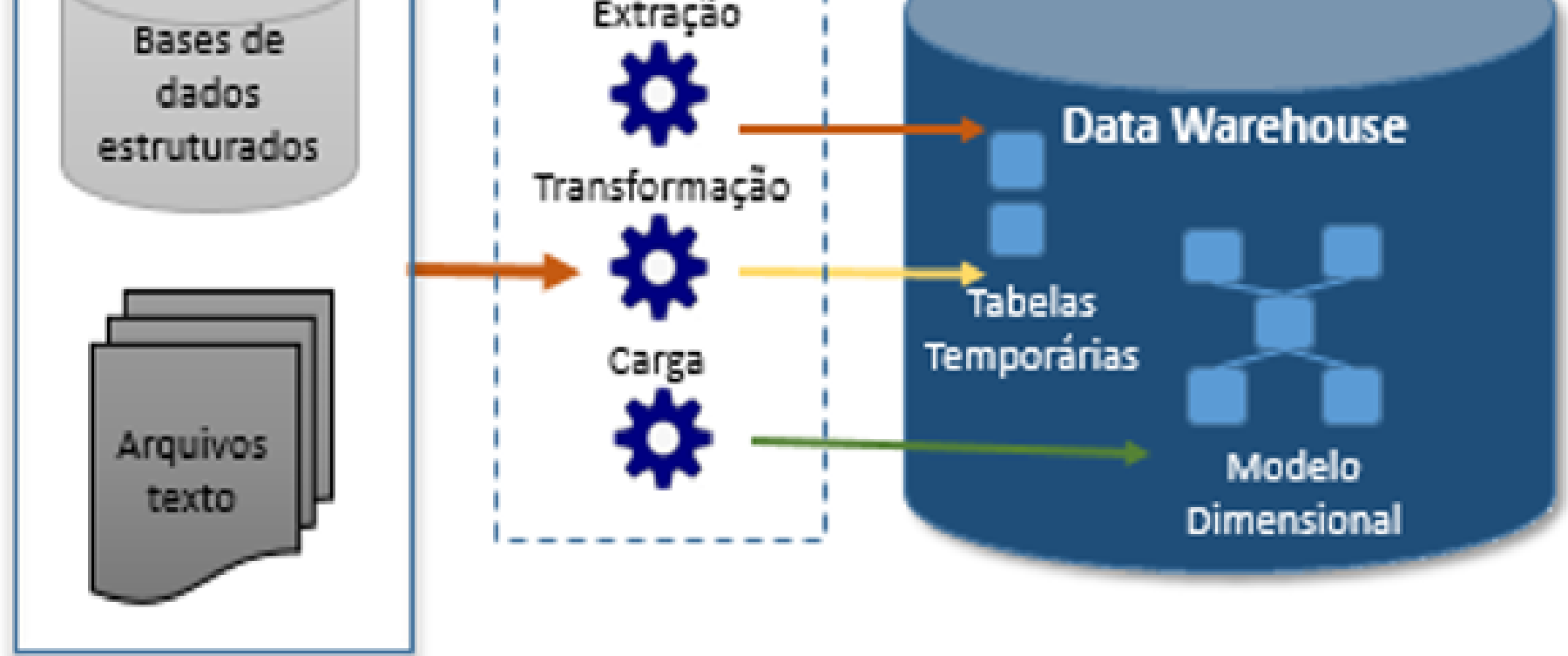
Processo ETL – Extração, Transformação e Carga dos dados

Na aula anterior conhecemos o processo ETL, como a extração dos dados pode ser realizada, quais os tipos de fontes de dados podem ser utilizados para a extração — como fontes estruturadas, semiestruturadas e não estruturadas — e conhecemos também a ferramenta de ETL que facilita o desenvolvimento dos passos a serem realizados no processo ETL.

Nesta aula vamos trabalhar a transformação dos dados, que é realizada após sua extração. Essa etapa limpa os dados e aplica transformações para a conformidade dos dados que serão carregados no DW/DM. Por fim, veremos a carga dos dados nas tabelas definitivas do DW/DM, que acontece após limpeza e transformação, quando os dados já estão preparados para a acomodação na visão dimensional.

A imagem a seguir ilustra a coleta dos dados das fontes de dados, o armazenamento nas tabelas temporárias, os tratamentos e aplicação de regras de negócios aos dados armazenados e a carga nas tabelas Dimensões e tabelas Fato desenhadas no modelo de dados dimensional.





 Processo ETL. Fonte: autor.

A etapa de extração, representada pela seta laranja, acessa a base de dados origem, copia os dados e os insere nas tabelas temporárias. A etapa transformação acessa as tabelas temporárias, altera os dados para deixá-los em conformidade com as regras definidas; e a etapa carga, representada pela seta verde, lê os dados tratados nas tabelas temporárias e insere-os nas tabelas Dimensões e tabelas fatos definidas no modelo de dados dimensional.


A seguir, vamos aprofundar o conhecimento sobre a etapa de transformação dos dados.

Transformação dos Dados (Cleaning and conforming)

O segundo passo do processo ETL é a transformação dos dados. As tarefas de limpeza e transformação dos dados são muito importantes, pois é nesse momento que eles sofrem as transformações necessárias para a apresentação no DW/DM. Além disso, essa etapa contribui para a melhoria dos sistemas origens, devido às validações dos dados que apontam problemas, que não foram verificados em sua entrada no sistema origem.

Nessa etapa são aplicadas as regras de negócios definidas no levantamento de requisitos e as transformações para adequar dados a serem carregados no DW/DM.

A etapa de transformação dos dados conta com alguns tipos comuns de tratamentos, como:

 Clique nos botões para ver as informações.

[Seleção de colunas/Exclusão de colunas](#)



Ao carregar os dados de uma origem, nem todas as colunas podem ser necessárias, então é possível selecionar apenas algumas colunas para serem carregadas no DW/DM.

[Tradução de dados](#)



Identificar o significado de um dado e adequar a representação do dado no DW/DM. Por exemplo: 1 para o sexo Masculino e 2 para o sexo Feminino; 1 para Tipo Moradia Casa e 2 para Tipo Moradia Apartamento.

Derivação de valores



Construir novos valores a partir de valores existentes. Por exemplo, o Valor do Lucro calculado a partir das métricas Preço de Venda e Preço de Custo.

Junção de dados



Padronização dos dados que possuem o mesmo conceito com visualização diferente, com o objetivo de unificar dados de diferentes sistemas.

Criação de chaves



Criação das chaves auxiliares (SK – Surrogate Key) de identificação única nas tabelas Dimensões.

Transposição ou rotação (pivoteamento)



Transformar linhas em colunas ou colunas em linhas.

Divisão de dados



A partir de uma coluna criar outras colunas. Por exemplo, a coluna de data pode ser dividida na coluna Mês e Ano.

Agregações



Sumarização de dados.

Os tratamentos aplicados aos dados visam obter a conformidade esperada para o ambiente analítico que está sendo construído, buscando a qualidade dos dados desejada para que as análises ofereçam valor à organização.

Saiba mais

A transformação aplica testes que verificam a qualidade dos dados, como valores nulos, códigos de elementos existentes, validade de datas, entre outros. Caso sejam encontrados problemas nos dados, devem ser registrados para que o responsável possa avaliar e corrigir o dado na base origem, se for o caso. Em casos mais delicados, o processo de ETL pode ser suspenso para que o problema seja analisado. Em outros casos, é atribuído o valor Não Informado.

Geralmente, os problemas encontrados são registrados em uma tabela de log, que indica em qual Dimensão ou tabela Fato ocorreu o erro, o registro que gerou o erro, o tipo do erro, entre outras informações que permitam rastrear sua origem.

Há sistemas que enviam notificações aos gestores dos dados e/ou aos responsáveis pelos processos, alertando sobre os erros ocorridos. Isso é muito interessante, pois não permite que os erros gerados sejam esquecidos.

A validação dos dados tem como objetivo verificar se o dado está coerente com as regras estabelecidas. Uma série de tratamentos é verificada nesse passo. Tabelas Dimensões e tabelas Fatos possuem validações e registros de problema diferentes. Em tabelas Dimensões, por exemplo, é necessário verificar se o elemento a ser inserido já está na tabela ou se deve ser realizada apenas uma atualização do dado. Já em tabelas Fatos, é necessário verificar se os elementos que compõem a chave primária possuem valores válidos, se o registro não está duplicado, entre outros, além da derivação de dados muito comum nos sistemas DW/DM.

Após a transformação dos dados, eles estão prontos para serem carregados nas tabelas definitivas na etapa Carga dos Dados.

Carga dos Dados (Loading ou Delivering)

A última etapa, e não menos importante, é a de carga dos dados. Essa etapa tem como finalidade verificar os dados que estão prontos para serem carregados nas tabelas Dimensões e tabelas Fatos do modelo de dados dimensional definido para o projeto.

Conforme visto na aula anterior, a periodicidade da carga dos dados no DW/DM varia conforme a necessidade da organização. A latência dos dados, referente à rapidez com que os dados do sistema origem são disponibilizados no DW/DM, deve ser analisada para que o processo atenda às necessidades da organização.

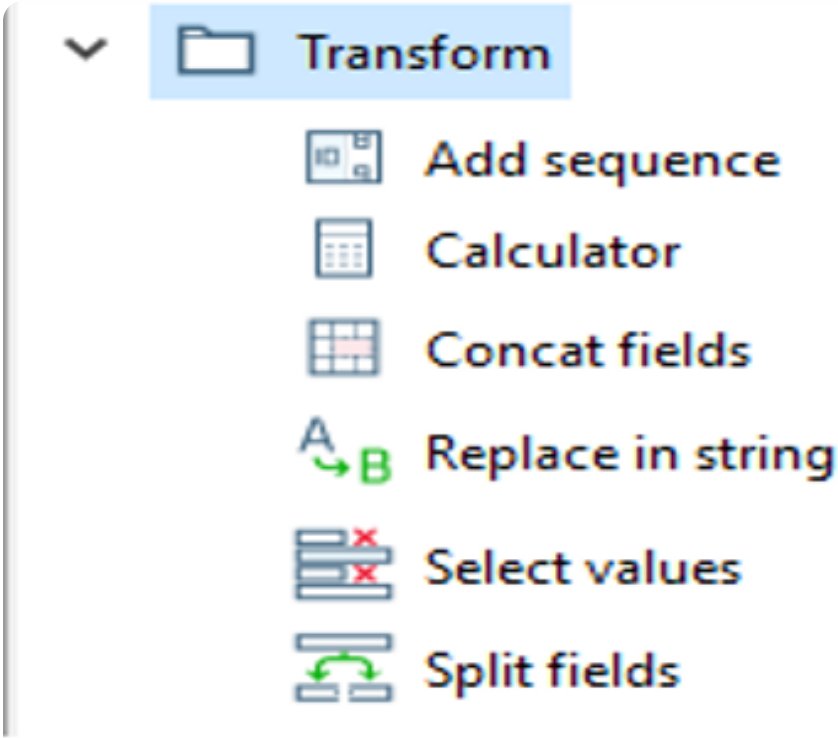
Comentário

Outra preocupação é como os dados são carregados no ambiente analítico. Há negócios que precisam que os dados sejam apagados para serem novamente carregados, enquanto há outras situações em que os dados novos são apenas incrementados na base e os antigos são somente atualizados. Grandes sistemas mantêm registros com o rastreamento de cada mudança realizada nos dados. Isso ajuda o entendimento do negócio nas modificações realizadas e a explicação em casos de auditoria de dados.

Esses elementos devem ser cuidadosamente analisados para que sejam aplicados, pois a implementação depende da necessidade de cada organização e de cada assunto tratado.

A transformação e carga dos dados na ferramenta de ETL

Como vimos na aula anterior, a ferramenta de ETL facilita a construção do processo de ETL, muito custoso em projetos de DW/DM. Assim como a leitura dos dados e inserção de registros em tabelas, a transformação conta com alguns steps para realizar as diversas tarefas de tratamento de dados.



Onde:

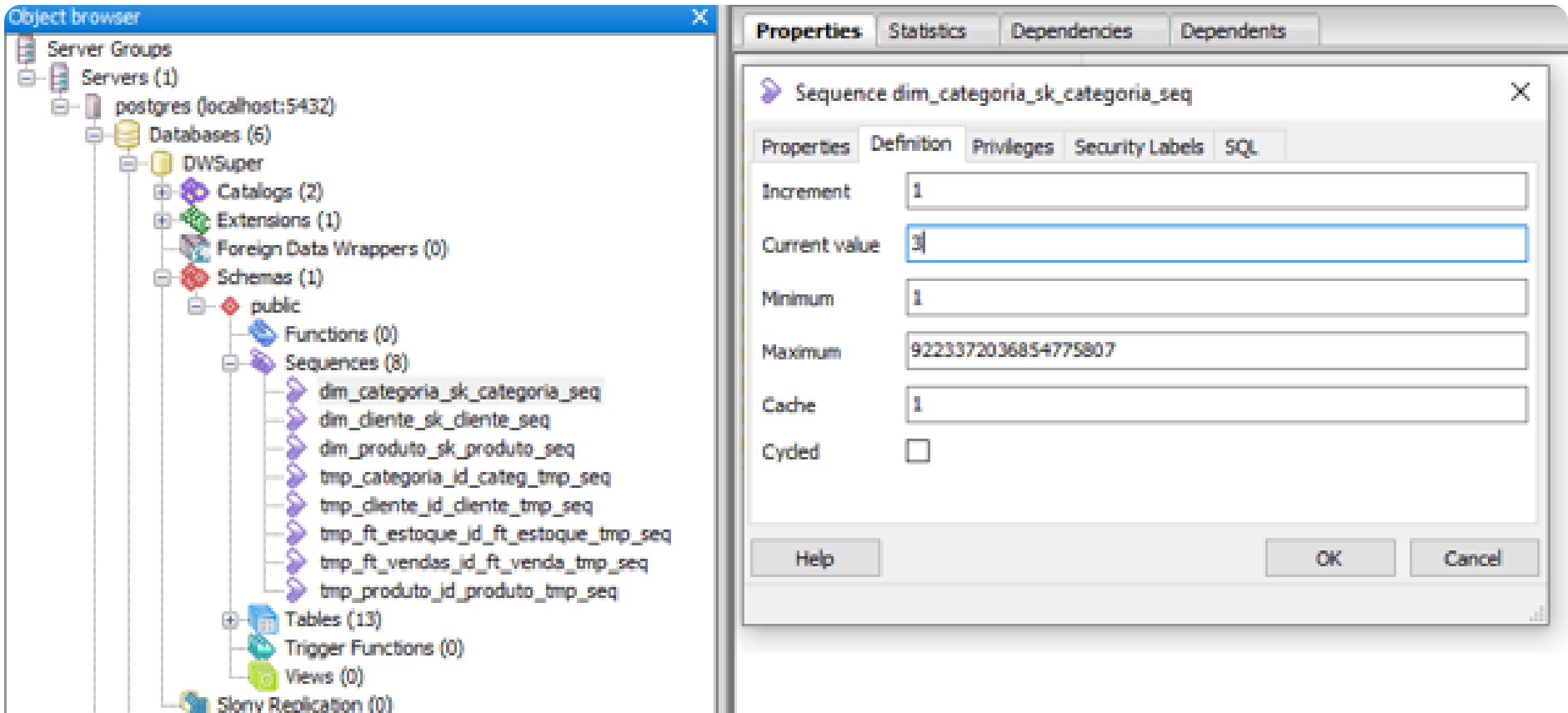
- Add Sequence – Obtém o próximo valor de um objeto *sequence* criado na base de dados.
- Calculator – Cria novos campos calculados.
- Concat Fields – Concatena vários campos em um campo.
- Replace in string – Substitui a ocorrência de uma palavra por outra palavra em uma *string*.
- Select Values – Seleciona ou remove campos em uma linha.
- Split fields – Divide um campo em outros campos.

As dimensões devem conter dois registros para tratar situações em que o código de um elemento Não Informado – por exemplo uma venda que não possui a identificação do cliente, ou Não se Aplica para as situações em que o dado não se aplica a um contexto de análise. Então, insira os elementos nas dimensões Categoria, Produto e Cliente.

Exemplo

```
INSERT INTO dim_categoria(sk_categoria, cd_categoria, ds_categoria) VALUES (1, 0, 'NÃO INFORMADO');  
INSERT INTO dim_categoria(sk_categoria, cd_categoria, ds_categoria) VALUES (2, 0, 'NÃO SE APLICA');
```

Observe que os elementos receberam os códigos 1 e 2, então a *sequence* deve ser iniciada no código 3. A imagem a seguir ilustra um exemplo de como alterar o valor atual da *sequence* no SGBD PostgreSQL.

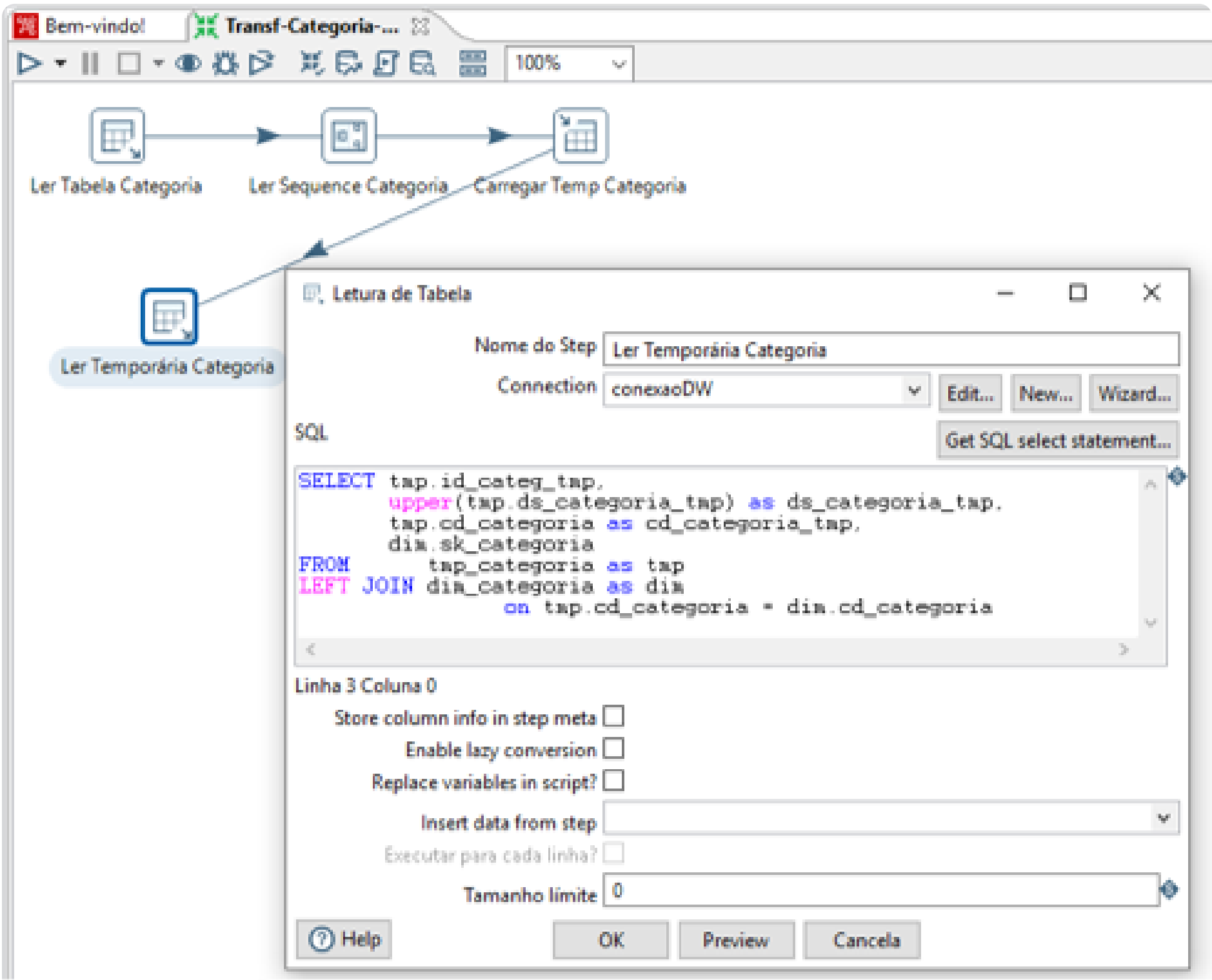


No SGBD, abra o grupo Sequences e clique com o botão direito na sequence desejada, escolha a opção Properties e altere o campo Current value.

Atenção! Aqui existe uma videoaula, acesso pelo conteúdo online

DWSuper – Implementação das etapas de transformação e carga dos dados

Dando sequência aos exemplos criados nas aulas anteriores, vamos implementar a transformação dos dados e a carga nas dimensões do modelo de dados dimensional. A imagem a seguir ilustra a transformação da Dimensão Categoria. O *step* Ler Temporária Categoria seleciona os elementos carregados na tabela temporária categoria e verifica se são novos elementos ou se já existem na tabela Dimensão Categoria. A consulta retornará todos os registros que estão presentes na tabela temporária e o elemento correspondente na Dimensão Categoria.



Por meio do código da categoria (cd_categoria) será verificado se o elemento já está cadastrado na dimensão. Se ele não for localizado, o elemento deve ser inserido. Para gerar uma chave SK para o elemento, será necessário chamar a sequence criada para a Dimensão Categoria. Caso o elemento seja encontrado na dimensão, ele deverá ser atualizado. Para realizar o passo de inserção ou atualização do elemento, deve ser utilizado o *step* Insert/Update.

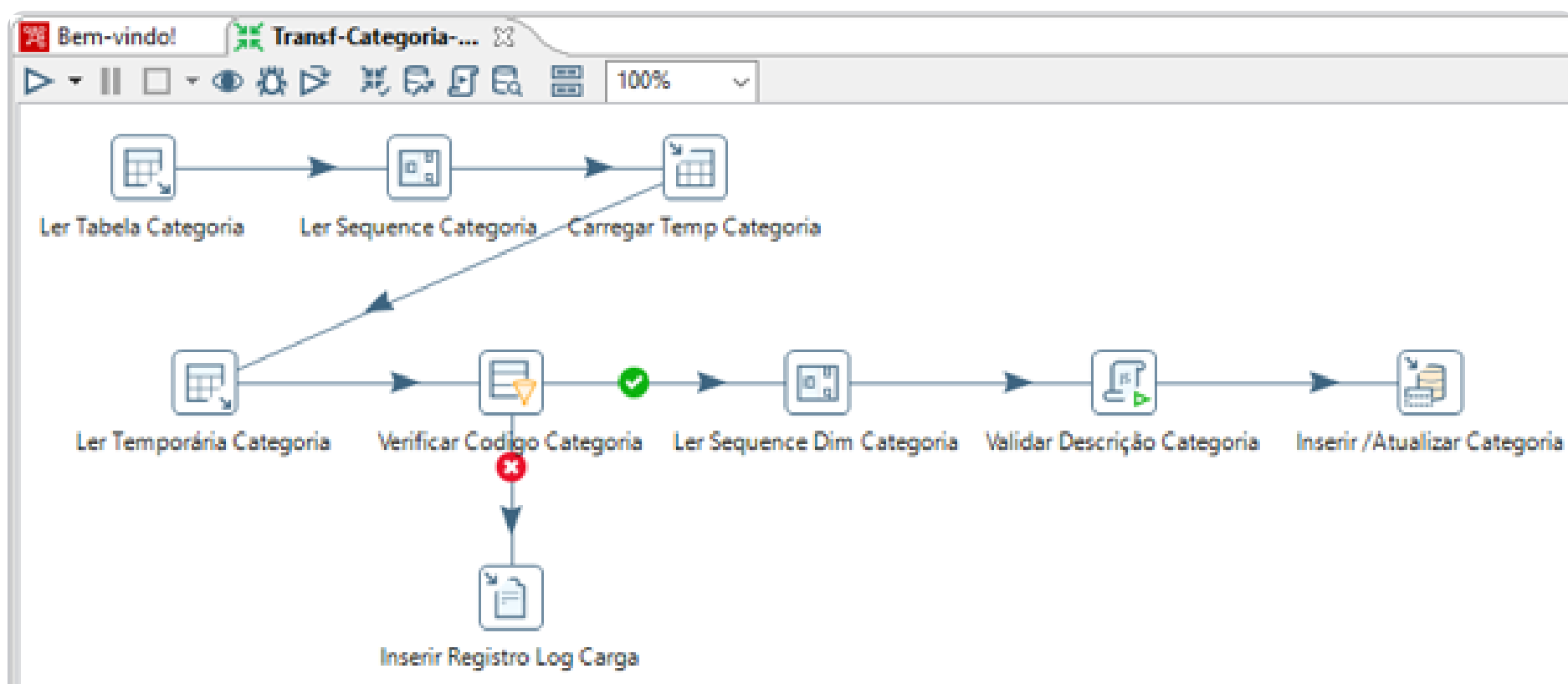
Em sistemas complexos alguns erros comuns podem ocorrer e devem ser tratados, como o elemento contido na tabela temporária, que pode não possuir o código do registro no sistema origem ou o registro não possui descrição preenchida. Cada caso deve ser tratado conforme a necessidade. No nosso exemplo, se o código (cd_categoria) não estiver preenchido, o registro será descartado e adicionado a um arquivo texto de log. Se a descrição não estiver preenchida, ela receberá o valor “DESCRIÇÃO EM BRANCO” e será inserido na dimensão.

Para testar as validações anteriores insira as duas linhas abaixo na base de dados origem (Tabela Categoria):

```
INSERT INTO public.categoria (codigo, descricao) VALUES (NULL, 'PADARIA');
```

```
INSERT INTO public.categoria (codigo, descricao) VALUES (220, NULL);
```

Veja todos os objetos inseridos para o tratamento dos dados da Dimensão Categoria.



Cada um dos steps possui uma função específica de tratamento a ser aplicado. Os *steps* devem ser usados nas transformações conforme a necessidade dos tratamentos que devem ser realizados nos dados. Veja a seguir cada um dos steps utilizados para tratar os dados da dimensão categoria.

Step Verificar Codigo Categoria (Filter Rows).

A imagem ilustra o teste de nulidade do campo `cd_categoria_tmp`. Se a condição de não ser nulo (Is Not Null) for verdadeira, o próximo passo é ler a *sequence* criada para a Dimensão Categoria (Send 'true' data to step). Caso contrário (Send 'false' data to step), o próximo passo é inserir o registro no arquivo de log da carga da Dimensão Categoria.

Filter rows

Step name

Verificar Codigo Categoria

Send 'true' data to step:

Ler Sequence Dim Categoria

Send 'false' data to step:

Inserir Registro Log Carga

The condition:

cd_categoria_tmp

IS NOT NULL

-

-

?

Help

OK

Cancela

Ler Sequence Dim Categoria (Add Sequence)

Veja o step Add Sequence que lê o próximo código para o campo sk_categoria a ser inserido na Dimensão Categoria.

Obtém o valor da sequencia do banco de dados

Nome do step

Ler Sequence Dim Categoria

Nome do valor

id_dimcategoria_seq

Use a database to generate the sequence

Usa BD para obter a sequencia ?

☒

Connection

conexaoDW

Edit...

New...

Wizard...

Schema name

public

Schemas...

Nome da sequência

dim_categoria_sk_categoria_seq

Sequences...

Use a transformation counter to generate the sequence

Usa o contador para calcular a sequência?

☐

Counter name (optional)

Inicia no valor

1

Incremento de

1

Valor máximo

999999999

Help

OK

Cancela

Validar Descrição Categoria (Modified JavaScript value).

Veja agora o *step* Modified JavaScript value que é usado na construção de expressões JavaScript para modificar os dados. O código digitado na área do *script* é executado uma vez em cada linha para tratar os valores indicados. Então, para verificar se a descrição da categoria está vazia, o campo *ds_categoria_tmp* deve testado como *null*. Se a resposta for verdadeira, a variável *vardescricao* recebe o valor "DESCRIÇÃO EM BRANCO"; caso contrário, a variável recebe o valor contido no campo.

Modified JavaScript value

Step nameValidar Descrição Categoria

Java script functions :

> Transform Scripts

> Transform Constants

> Transform Functions

> Input fields

codigo

descricao

id_categoria_seq

id_categ_tmp

ds_categoria_tmp

cd_categoria_tmp

sk_categoria

id_dimcategoria_seq

> Output fields

Please use the 'Repla

Java script :

Script 1 23

//Script here

var vardescricao;

if (ds_categoria_tmp == null) {

vardescricao = "DESCRIÇÃO EM BRANCO";

}

else {

vardescricao = ds_categoria_tmp;

}

<

Linenn: 0

Compatibility mode? ☐ Optimization level9

Fields

#	Fieldname	Rename to	Type	Length	Precision	Replace value 'Fieldname' or 'Rename to'
1	vardescricao		String			N

Help

OK

Cancela

Get variables

Test script

Inserir /Atualizar Categoria (Insert/Update).

A figura ilustra o *step* Insert/Update, que verifica se o registro deve ser inserido ou se ele deve ser alterado com base no campo código da categoria (*cd_categoria*) no *grid*. The Key(s) to look up the value(s). No *grid* Update Fields devem ser listados os campos que serão alterados ou inseridos. A coluna Update é um informativo sobre se o campo listado deve ser alterado. Observe que os campos *cd_categoria* e *sk_categoria* não podem ser alterados, somente inseridos. Observe também que o campo de descrição escolhido (Stream Field) é a variável que tratou a descrição da categoria no passo anterior.

Insert / update

Step name

Inserir /Atualizar Categoria

Connection

conexaoDW

▼

Edit...

New...

Wizard...

Target schema

public

\$

Navega...

Target table

dim_categoria

\$

Browse...

Commit size

100

\$

Don't perform any updates:

☐

The key(s) to look up the value(s):

#	Table field	Comparator	Stream field1	Stream field2
1	cd_categoria	=	cd_categoria_tmp	

Get fields

Update fields:

#	Table field	Stream field	Update
1	ds_categoria	vardescricao	Y
2	cd_categoria	cd_categoria_tmp	N
3	sk_categoria	id_dimcategoria_seq	N

Get update fields

Edit mapping

Help

OK

Cancela

SQL

Inserir Registro Log Carga (Text File Output).

Veja o step que irá adicionar os dados do registro descartado na validação. Para isso, deve ser informado na aba File a extensão do arquivo, o caminho e o nome do arquivo que receberá o *log*; na aba Content deve ser informado o tipo do separador dos campos, conforme a extensão do arquivo escolhido e na aba Fields devem ser adicionados os campos que serão inseridos no arquivo de *log*.

Text file output

Nome do Step

Inserir Registro Log Carga

File

Content

Fields

#	Name	Type	Format
1	id_categ_tmp	None	
2	ds_categoria_tmp	None	
3	cd_categoria_tmp	None	

<

>

Obtem campos

Minimal width

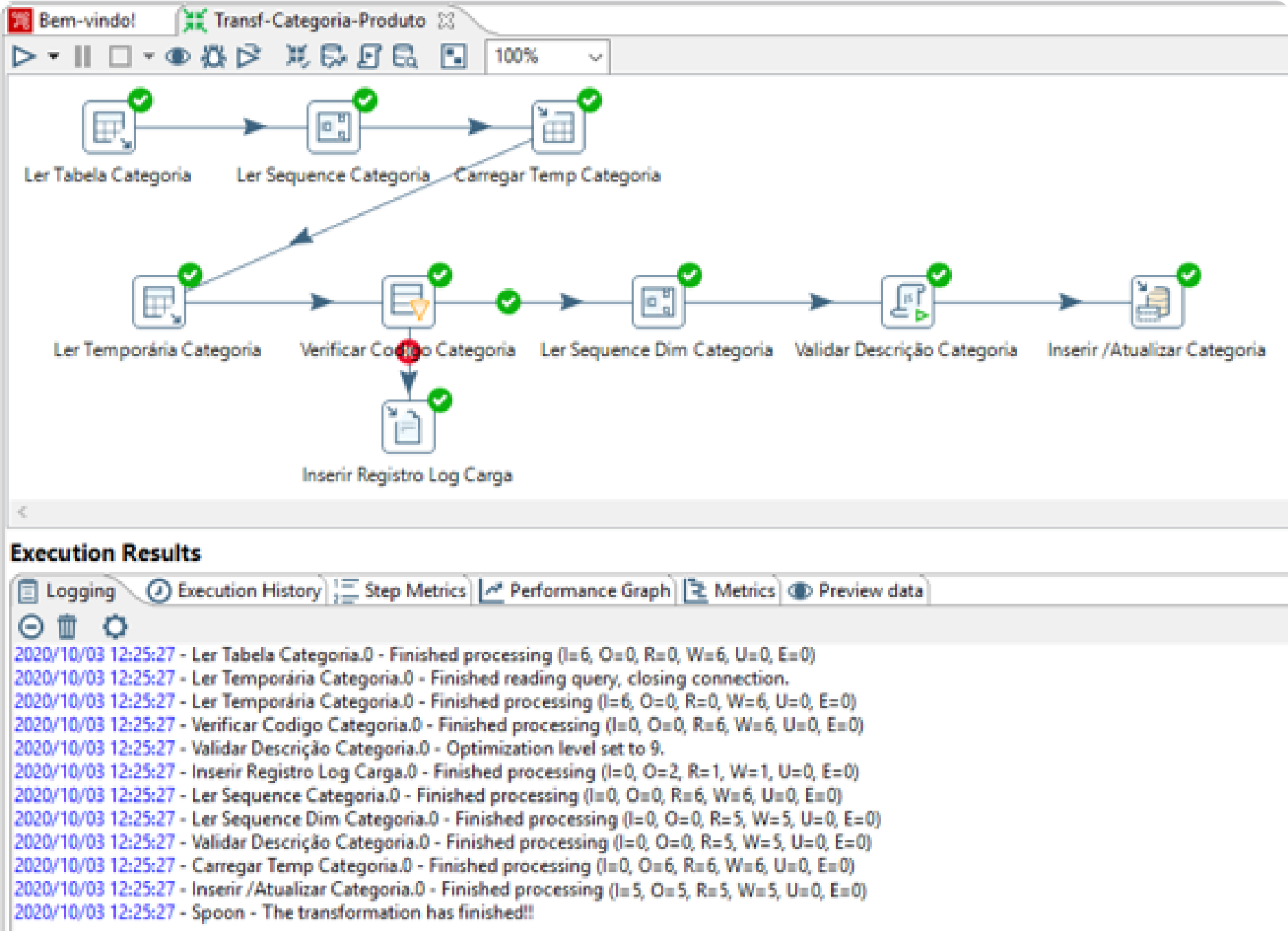
?

Help

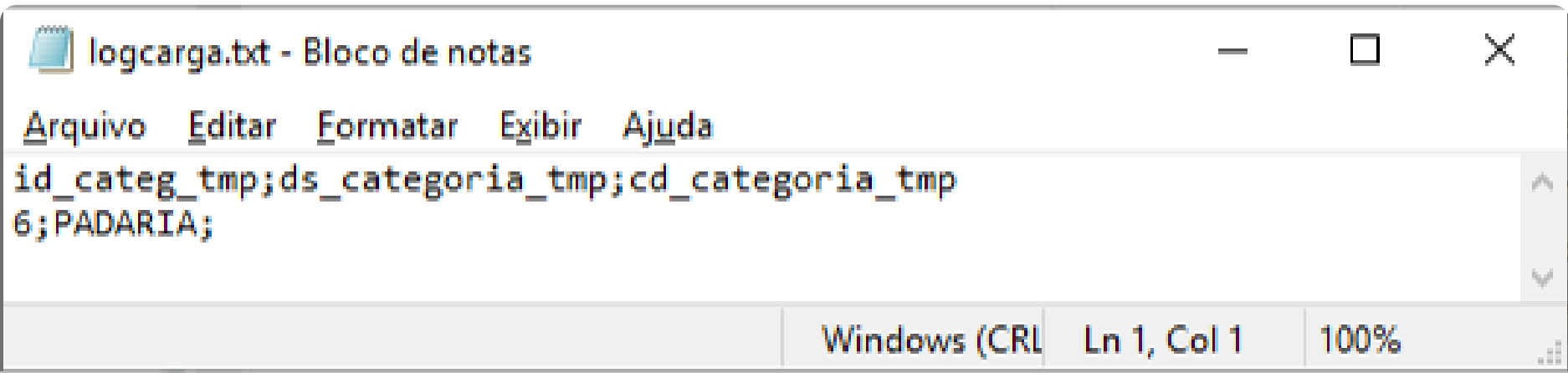
OK

Cancela

Com a transformação pronta é hora de executá-la completamente. A figura a seguir ilustra a finalização de execução da transformação Transf-Categoria. Não é raro ocorrer erros durante a execução de algum passo. Caso isso ocorra, verifique o erro apresentado na janela que será aberta, corrija o problema e execute novamente a transformação.



A imagem ilustra os registros descartadas inseridos no arquivo logcarga após a execução da transformação Transf-Categoria-Produto:



Após a execução da transformação, a Dimensão Categoria está preenchida com os dados extraídos da tabela origem Categoria. A Figura ilustra os registros inseridos na Dimensão Categoria após a execução da transformação Transf-Categoria-Produto. Observe que o elemento Padaria não foi inserido e que o elemento de código 220, que não possuía descrição preenchida, foi tratado e apresenta elemento registrado.

Quando o sistema origem preencher o código do elemento Padaria, ele será inserido na Dimensão Categoria e quando o elemento 220 tiver sua descrição corrigida, ele será atualizado na dimensão.

Edit Data - postgres (localhost:5432) - DWSuper - dim_categoria

FileEditViewToolsHelp

No limit

	sk_categoria [PK] serial	cd_categoria integer	ds_categoria character varying(45)
1	1	0	NÃO INFORMADO
2	2	0	NÃO SE APLICA
3	4	100	HIGIENE
4	5	110	LIMPEZA
5	6	90	GRÃO
6	7	60	BEBIDA
7	8	220	DESCRIÇÃO EM BRANCO
*			

7 rows.

Vamos praticar?

Agora, como exercício, complete as transformações das dimensões Produto e Cliente com os passos de tratamento e carga dos dados.

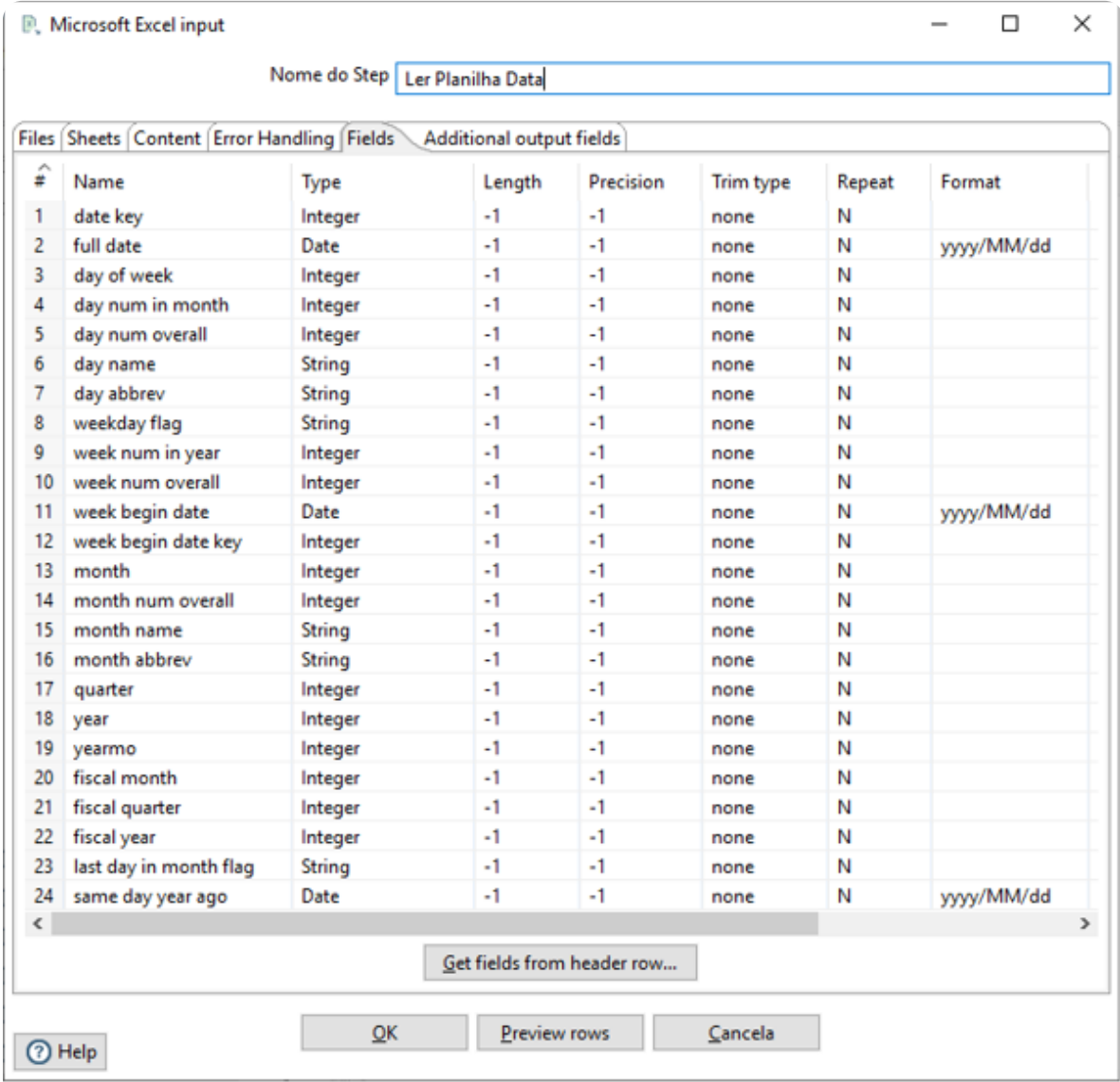
Atenção! Aqui existe uma videoaula, acesso pelo conteúdo online

📄 Dimensão Data

👉 Clique no botão acima.

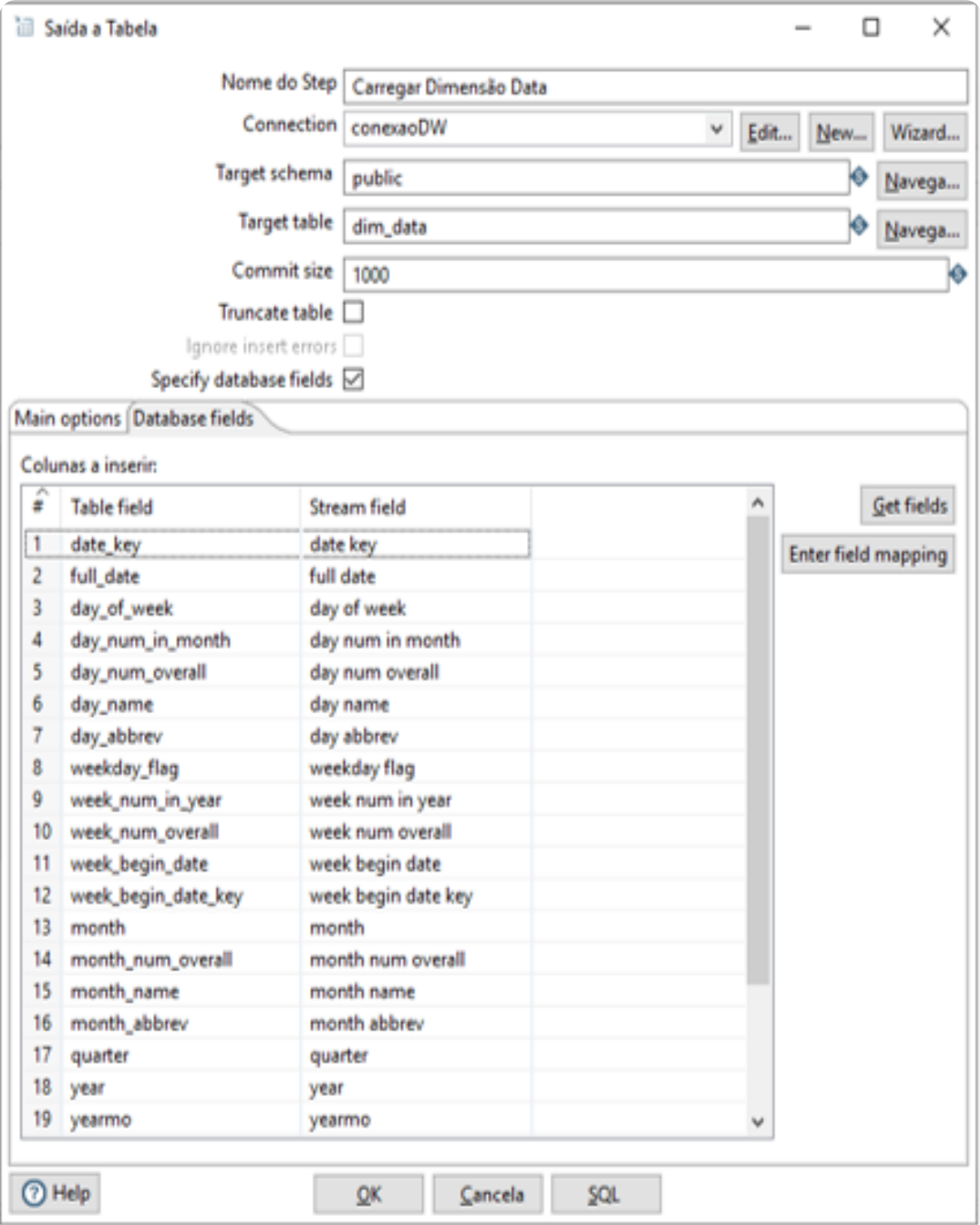
A Dimensão Data pode ser carregada por meio da planilha disponibilizada no site Kimball Group (<https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/books/data-warehouse-dw-lifecycle-toolkit/>). A planilha Excel que contém as colunas e as linhas para a carga da Dimensão Data pode ser alterada conforme a necessidade do projeto. Caso queira iniciar em determinada data, altere a data inicial no endereço de célula Coluna Full Date.

Utilize o step Microsoft Excel Input para ler os dados planilha, na aba Files escolha o arquivo Excel de datas baixado e configure conforme ilustrado.



📷 Leitura da Planilha Data. Fonte: Kimball Group (2020).

Adicione o step Table Output para inserir os dados na Dimensão Data, ligue ao step Microsoft Excel Input e configure. Note que uma boa prática seria alterar os nomes das colunas seleccionadas para nomes em Português, pois serão usados como cabeçalhos de relatórios e consultas analíticas.



 Carga da Dimensão Data. Fonte: O autor

Nesse momento, as dimensões do DW Supermercado estão preenchidas com os dados tratados extraídos das bases de origem, aguardando a carga das tabelas fatos do modelo de dados dimensional que será visto na próxima aula.

Nesta aula estudamos o processo de transformação e carga dos dados das dimensões no modelo de dados dimensional, utilizando a ferramenta de ETL Pentaho Data Integrator (PDI).

Agora, é hora de fixar o entendimento!

Atenção! Aqui existe uma videoaula, acesso pelo conteúdo online

Atividade

O processo de ETL é composto por três etapas: extração dos dados, transformação e carga dos dados. Sobre o processo ETL é correto afirmar que:

a) A extração dos dados lê e copia os dados da base de dados origem, a transformação dos dados realiza a limpeza deles e aplica as regras de negócio estabelecidas e a carga de dados carrega os dados tratados na base de dados destino.

- b) A extração dos dados lê e aplica a transformação dos dados na base de dados origem e a carga de dados carrega os dados tratados na base de dados destino.
- c) A extração dos dados lê e copia os dados da base de dados origem, a transformação carrega os dados tratados na base de dados destino.
- d) A extração dos dados lê e copia os dados da base de dados origem, a transformação dos dados realiza a limpeza dos dados e aplica as regras de negócio estabelecidas e a carga de dados informa as bases de dados origem se os dados estão corretos ou se há problemas de integridade.
- e) A extração dos dados verifica se os dados existem na base de dados origem, a transformação dos dados copia os dados para as tabelas temporárias e a carga de dados carrega os dados tratados na base de dados destino.

A etapa de Transformação dos dados aplica vários tipos de tratamentos aos dados. Um deles é a divisão que:

- a) Seleciona ou exclui colunas referentes aos dados coletados das fontes origens.
- b) Adequa a representação do conteúdo identificando o significado do dado.
- c) Constrói novos indicadores a partir de outros indicadores.
- d) Transforma linhas em colunas e vice-versa.
- e) Cria outras colunas de dados a partir de uma coluna de dado.

A ferramenta de ETL PDI possui steps que realizam tarefas de leitura de dados, transformações, carga de dados, entre outros. Podemos citar como exemplo de steps:

- a) Table Input, Add Sequence e ETL.
- b) Data Mart, Filter Rows e Modified JavaScript value.
- c) Insert/Update, Modified JavaScript value e PDI.
- d) Table Input, Filter Rows e Modified JavaScript value.
- e) Modified JavaScript value, Table Output e Spoon.

Notas

Título modal ¹

Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos. Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos. Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos.

Título modal ¹

Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos. Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos. Lorem Ipsum é simplesmente uma simulação de texto da indústria tipográfica e de impressos.

Referências

KIMBALL, M. R. R. **The Data Warehouse Toolkit - The Definitive Guide to Dimensional Modeling**. 3. ed. Indianapolis: John Wiley Sons, 2013.

VANTARA, H. Pentaho Documentation (Pentaho Data Integration). Hitachi – Inspire the Next. Disponível em: https://help.pentaho.com/Documentation/8.2/Products/Data_Integration#Kettle. São Paulo: Prentice Hall, 2005.

HORNGREN, Charles T.; DATAR, Srikant M.; FOSTER, George. [Contabilidade de Custo](#). São Paulo: Prentice Hall, 2005.

Próxima aula

- Processo ETL – A carga das Tabelas Fato, Agregada e a Consolidação;
- O processo de expurgo de dados;
- A construção do JOB na ferramenta PDI.

Explore mais

- Conheça mais sobre o PDI e aprofunde os conhecimentos sobre os steps no site Hitachi Vantara. Link: <https://www.hitachivantara.com/pt-br/company.html>.