

# Aula 8: Processo ETL – Transformação e Carga dos Dados de Tabelas Fatos

## Apresentação

---

Nesta aula serão apresentadas as etapas de transformação e carga das tabelas Fato transacional, agregada e consolidada; o conceito de expurgo de dados, a criação do JOB na ferramenta PDI e o conceito de gerenciamento de processos.

## 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;
- Identificar a etapa de carga dos dados nas tabelas fato definitivas.

## Processo ETL – Carga das Tabelas Fato Transacional, Agregada e Consolidada

---

Na aula anterior foi descrito o processo ETL e apresentada a ferramenta de ETL, que auxilia o desenvolvimento do processo de extração, transformação e carga dos dados. Nesta aula vamos dar continuidade aos conceitos aprendidos e seguir com a transformação da tabela Fato de transações, dos dados agregados e dos dados consolidados.

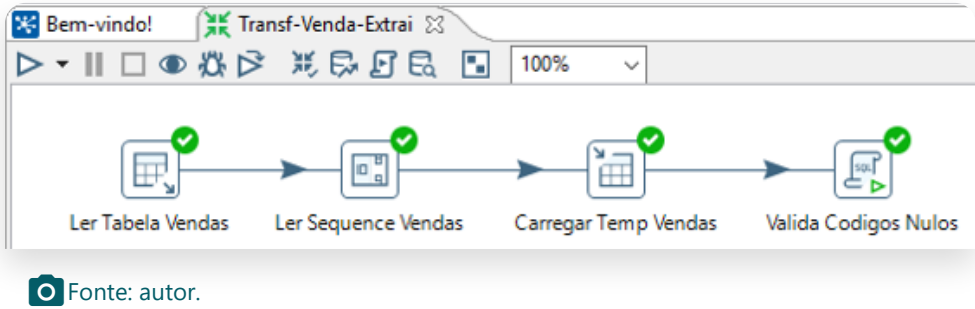
### Carga da tabela Fato

A tabela Fato armazena os dados mensuráveis do negócio, como quantidades e valores. O fato ocorrido é descrito pelas visões que compõem o DW/DM. Assim, além dos dados mensuráveis, a tabela armazena a relação com as dimensões por meio da chave primária (SK) de cada uma das tabelas.

Para que a carga da tabela Fato seja realizada com sucesso é necessário que as chaves das dimensões estejam validadas. Isso significa que a chave SK da dimensão a ser inserida na tabela Fato deve existir na dimensão. Para isso, as chaves devem ser verificadas e adicionadas à tabela temporária da tabela Fato, para que os registros validados sejam selecionados e nela inseridos.

A figura a seguir ilustra os passos realizados para a extração dos dados da tabela da base de dados de origem para a tabela Fato temporária que registra as vendas.

Extração e carga da tabela fato temporária de vendas



O step Ler Tabela Vendas (Table Input) acessa a base de dados origem e seleciona os dados contidos na tabela de vendas do sistema operacional. No nosso exemplo, podemos usar apenas uma *query* simples que irá retornar os dados que precisamos:

```
SELECT * FROM vendaproductos;
```

No entanto, em um projeto real, as regras de negócio devem ser observadas para que somente sejam extraídos os registros referentes às vendas realizadas no período desejado (Dia, Mês, entre outros).

O *step* Ler Sequence Vendas (AddSequence) gera a chave SK para a tabela temporária; o *step* Carregar Temp Vendas (Table Output) insere os registros extraídos do sistema origem na tabela Fato temporária e o *step* Valida Codigos Nulos atribui o valor 0 para as chaves nulas dos elementos que serão validados nas dimensões, conforme o comando a seguir:

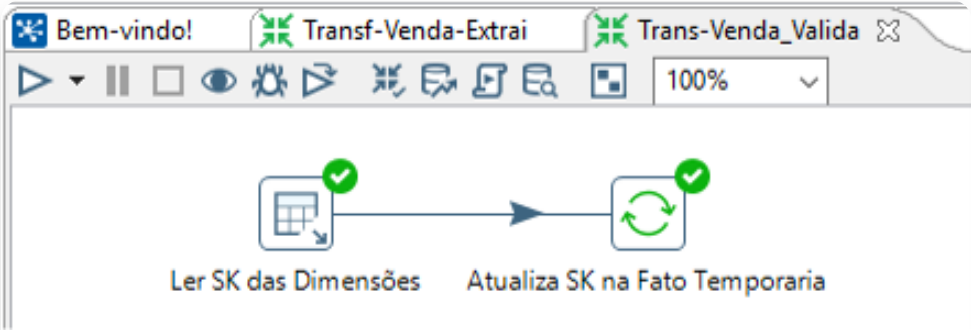
```
UPDATE tmp_ft_vendas SET cd_cliente_vendas_tmp = 0
WHERE id_ft_venda_tmp IN (SELECT id_ft_venda_tmp
                          FROM tmp_ft_vendasft
                          WHERE ft.cd_cliente_vendas_tmp ISNULL);
```

Assim, os registros de vendas em que o código do cliente é igual a nulo, receberão o valor 0 (zero) e, ao final dessa etapa, os registros estão carregados na tabela Fato temporária e prontos para as validações de chaves primárias das dimensões.

Veja os *steps* de validação das chaves SKs das dimensões.

**Atenção!** Aqui existe uma videoaula, acesso pelo conteúdo online

Transformação dos dados da Tabela de Fato Vendas.



Fonte: O autor.

O *step* Ler SK das Dimensões (Table Output) seleciona as chaves das dimensões por meio do código do elemento no sistema origem.

Após a validação e o preenchimento dos códigos das dimensões, as chaves das dimensões devem ser validadas e a atualizadas na tabela temporária da tabela Fato de vendas.

A figura abaixo ilustra a *query* contida no *step* SelecionarSKs Dimensões (Table Input), que relaciona a tabela temporária da tabela Fato de vendas com as dimensões por meio dos códigos dos elementos. Observe que, como somente a tabela temporária do Produto possui a chave da Categoria e ela está ligada diretamente à tabela Fato então, a chave sk foi recuperada utilizando a tabela tmp\_produto.

Nome do Step: Ler SK das Dimensões

Connection: conexaoDWLer

SQL:

```
SELECT ft.id_ft_venda_tmp as id_venda, prod.sk_produto, cli.sk_cliente,
       dat.date_key as sk_date_key, cat.sk_categoria
FROM   tmp_ft_vendas ft, dim_produto prod, dim_cliente cli,
       dim_data dat, tmp_produto tprod, dim_categoria cat
WHERE  ft.cd_produto_vendas_tmp = prod.cd_produto
AND    ft.cd_cliente_vendas_tmp = cli.cd_cliente
AND    ft.dt_data_venda_tmp = dat.full_date
AND    ft.cd_produto_vendas_tmp = tprod.cd_produto
AND    tprod.cd_categoria = cat.cd_categoria;
```

Linha 1 Coluna 0

Enable lazy conversion: ☐

Replace variables in script: ☐

Insert data from step:

Executar para cada linha: ☐

Tamanho limite: 0

Fonte: O autor.

Validação das chaves SKs das dimensões.

Após selecionar as SKs, elas devem ser atualizadas na tabela temporária.

Atualização das chaves SKs na tabela Fato temporária Vendas.

Step name: Atualiza SK na Fato Temporaria

Connection: conexaoDWLer

Target schema: public

Target table: tmp\_ft\_vendas

Commit size: 100

Use batch updates?: ☒

Skip lookup: ☒

Ignore lookup failure?: ☐ Flag field (key found):

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

#	Table field	Comparator	Stream field1	Stream field2
1	id_ft_venda_tmp	=	id_venda	

Update fields:

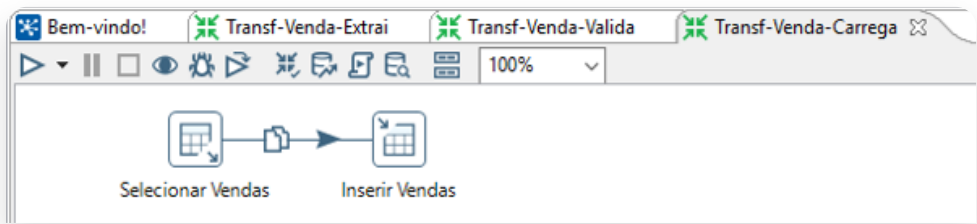
#	Table field	Stream field
1	sk_produto_ft_venda_tmp	sk_produto
2	sk_cliente_ft_venda_tmp	sk_cliente
3	sk_data_ft_venda_tmp	sk_date_key
4	sk_categoria_vendas_tmp	sk_categoria

Fonte: O autor.



Outras validações e transformações podem ser aplicadas nessa etapa conforme a necessidade dos dados que estão sendo tratados. Com os dados prontos para serem inseridos na tabela Fato definitiva, eles devem ser selecionados e inseridos na tabela Fato.

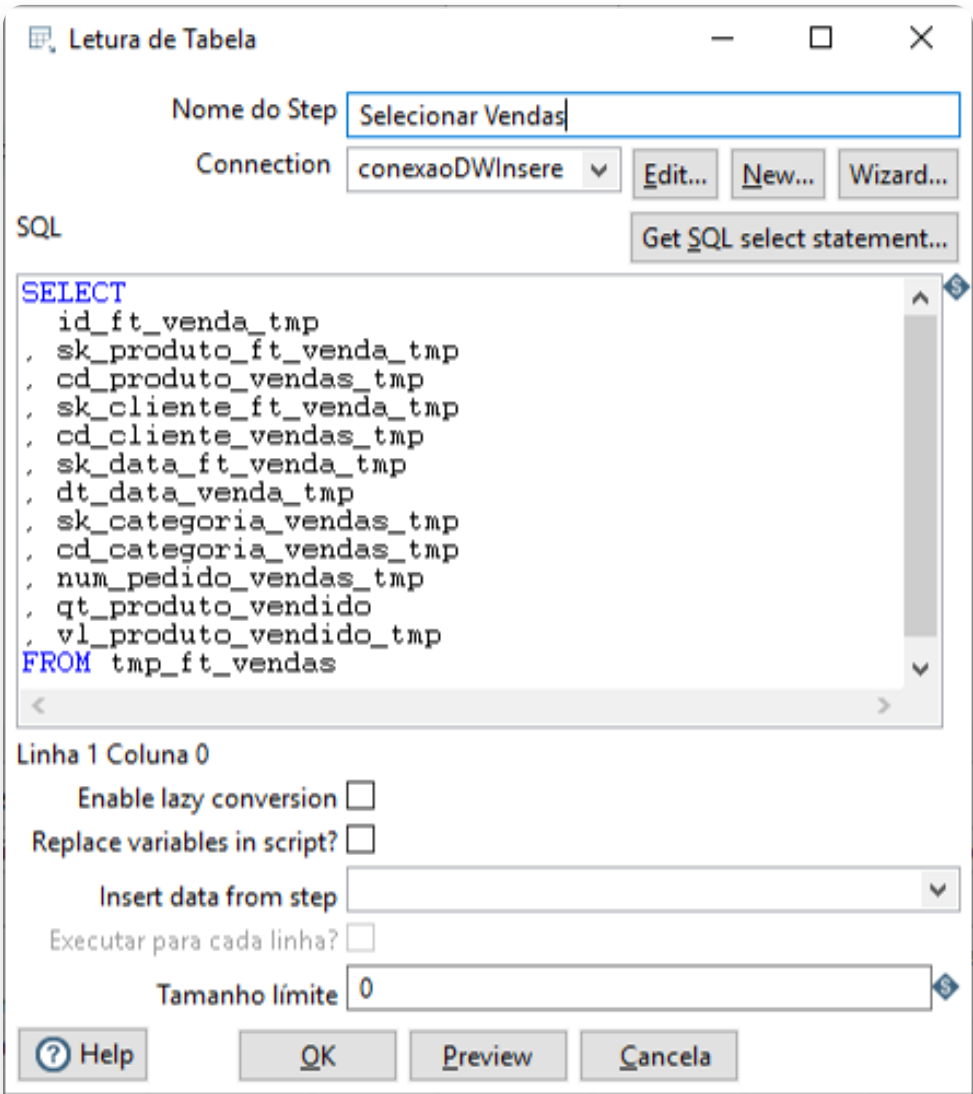
Veja o *step* Selecionar Vendas (Table Input), que seleciona os registros a serem inseridos na tabela Fato. Caso a tabela temporária contenha uma coluna indicando que o registro está liberado para ser carregado na tabela Fato, a condição deve ser respeitada na seleção das linhas, assim como outras formas de restrições na validação dos dados. No nosso exemplo, todas as linhas serão carregadas.



Fonte: O autor.

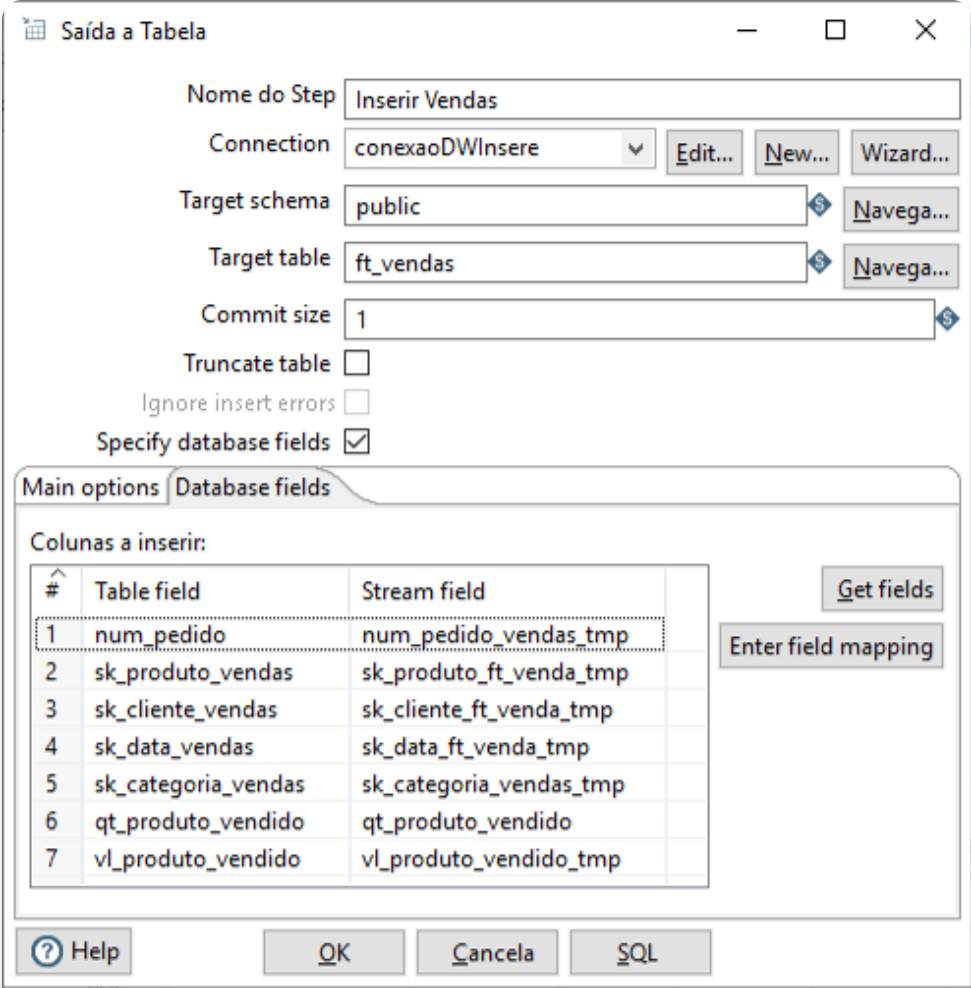
Carga da Tabela de Fato Vendas.

A seguir, vemos a seleção dos campos contidos na tabela Fato temporária. O comando a seguir foi gerado pelo botão Get SQL statement. No entanto, aqui podem ser selecionados apenas os campos a serem inseridos na tabela Fato definitiva.



Fonte: O autor.

Por fim, o *step* Inserir Vendas (Table Output) insere os registros na tabela Fato de vendas. Observe que somente os campos a serem inseridos na tabela Fato são mapeados nesse *step*.



Fonte: O autor.

Seleção dos dados para a Tabela de Fato Vendas.

Veja a tabela ft\_vendas carregada com os registros de vendas do supermercado.

	sk_produto_vendas [PK] integer	sk_cliente_vendas [PK] integer	sk_data_vendas [PK] character(8)	sk_categoria_vendas [PK] integer	num_pedido [PK] integer	qt_produto_vendido integer	vl_produto_vendido numeric(10,2)
1	10	1	20200620	1	98563	5	3.00
2	10	6	20200619	1	95687	12	3.00
3	10	7	20200620	1	95687	24	3.00
4	10	8	20200620	1	74512	6	3.00
5	11	5	20200520	1	5487	1	6.00
6	14	1	20200625	6	98563	2	10.00
7	14	4	20200518	6	56958	2	12.80
8	15	5	20200520	6	5487	10	10.00
9	16	5	20200520	6	5487	2	3.00
10	17	1	20200625	4	98563	2	6.00
11	18	5	20200520	5	5487	3	4.00
12	19	1	20200625	5	98563	1	15.00
*							

Fonte: O autor.

Tabela de Fato Vendas.

**Atenção!** Aqui existe uma videoaula, acesso pelo conteúdo online

Vamos praticar?

Agora, para exercitar os passos verificados, construa a transformação para a tabela Fato de estoque.

Carga da Tabela Fato Agregada Vendas

Em aulas anteriores você aprendeu sobre a agregação de dados, em que a tabela Fato agregada armazena informações pré-calculadas de acordo com o nível de granularidade desejado às análises para as quais está sendo construída a tabela Fato agregada, que é mais alto do que o da tabela Fato transacional.

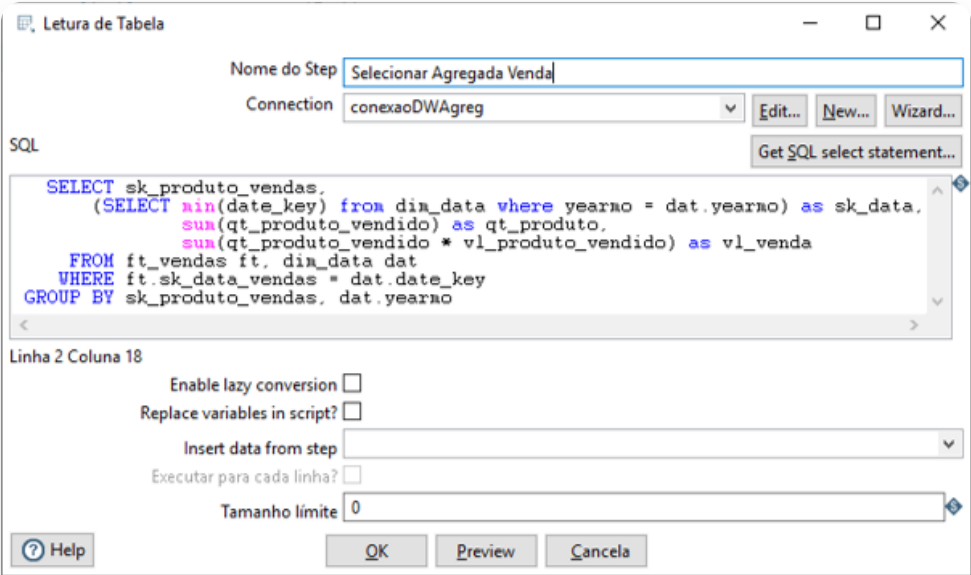
Atenção

De acordo com esse conceito, ao preparar os dados para a tabela Fato agregada os dados devem ser sumarizados e o nível de granularidade diminuído, como é o caso do exemplo a seguir.

A tabela Fato Agregada Venda é destinada a apoiar as análises referentes aos produtos vendidos no grão mês. A dimensão Data possui todos os dias do mês, mas, como vamos carregar a tabela Fato agregada no grão mês, devemos escolher um único registro da dimensão Data que represente o mês. Para o exemplo, será o primeiro dia do mês. A medida quantidade de produtos vendida deve ser sumarizada e a valor de venda deve ser calculado com base nas medidas quantidade de produtos e valor do produto vendido.

Veja a seguir o exemplo ilustrado a seguir.

Consulta de agregação da Tabela de Fato Vendas.



Fonte: O autor.

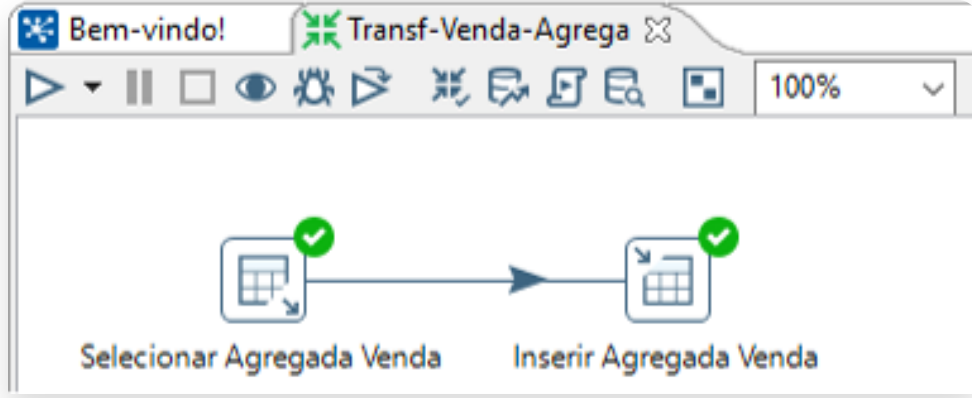
Para examinar o resultado da *query* de agregação, clique no botão preview e será exibido o resultado ilustrado na figura abaixo. Observe que o resultado apresenta o somatório da quantidade do produto e o valor total da venda por produto e por mês.

Resultado da consulta de agregação.

Examine preview data				
Rows of step: Selecionar Agregada Venda (9 rows)				
#	sk_produto_vendas	sk_data	qt_produto	vl_venda
1	14	20200501	2	25.6
2	15	20200501	10	100.0
3	16	20200501	2	6.0
4	17	20200601	2	12.0
5	10	20200601	47	141.0
6	11	20200501	1	6.0
7	18	20200501	3	12.0
8	14	20200601	2	20.0
9	19	20200601	1	15.0

Fonte: O autor.

Para completar a transformação, deve ser adicionado o *step* para inserção dos dados (Table Output) na tabela Fato agregada. No step Inserir Agregada Venda relacione os campos da query com os campos da tabela Fato agregada (agr\_vendas\_produto).



Fonte: O autor.

Carga da tabela fato agregada.

Após a execução da transformação, a tabela Fato agregada de vendas está carregada.

## Carga da Tabela Fato Consolidada

A tabela Fato consolidada agrega os dados unindo dados contidos em mais de uma tabela Fato, por exemplo, a tabela Fato de vendas e a tabela Fato de estoque do cenário Supermercado.

No cenário Supermercado há uma análise que necessita que os dados sejam consolidados, para responder à pergunta: quais são os fabricantes dos produtos que oferecem maior lucro na comercialização dos seus itens?

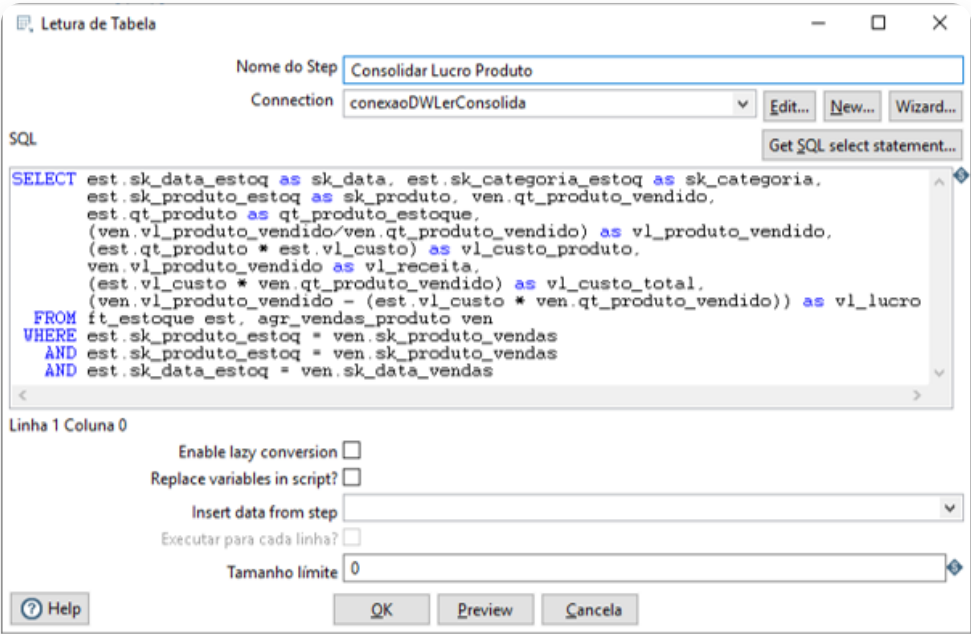
Dica



Para responder a essa questão é necessário relacionar os dados do produto comprado do fabricante e os dados do produto vendido aos clientes. As métricas desejadas devem ser calculadas utilizando-se as métricas contidas na tabela Fato estoque e na tabela Fato venda.

Como temos uma agregada de vendas no grão mês, ela será utilizada na consolidação.

Comando para consolidação dos dados Lucro Produto.



Fonte: O autor.

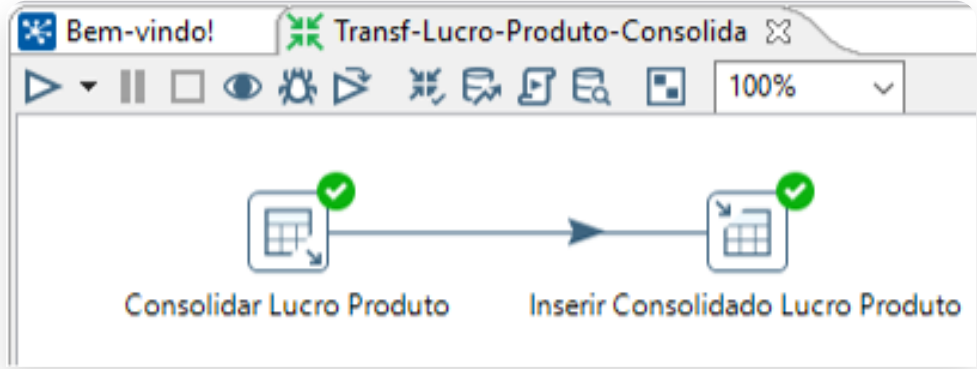
A imagem a seguir ilustra o resultado da consolidação dos dados e que será inserido na tabela Fato definitiva.

Dados consolidados Lucro Produto.

Examine preview data										
Rows of step: Table input (9 rows)										
#	sk_data	sk_categoria	sk_produto	qt_produto_vendido	qt_produto_estoque	vl_produto_vendido	vl_custo_produto	vl_receita	vl_custo_total	vl_lucro
1	20200601	1	10	47	80	3.0	80.0	141.0	47.0	94.0
2	20200501	1	11	1	50	6.0	125.0	6.0	2.5	3.5
3	20200501	6	14	2	60	12.8	240.0	25.6	8.0	17.6
4	20200601	6	14	2	40	10.0	180.0	20.0	9.0	11.0
5	20200501	6	15	10	60	10.0	180.0	100.0	30.0	70.0
6	20200501	6	16	2	60	3.0	78.0	6.0	2.6	3.4
7	20200601	4	17	2	30	6.0	96.0	12.0	6.4	5.6
8	20200501	5	18	3	30	4.0	24.0	12.0	2.4	9.6
9	20200601	5	19	1	20	15.0	80.0	15.0	4.0	11.0

Fonte: O autor.

Para completar a transformação, deve ser adicionado o *step* para inserção dos dados (Table Output) na tabela Fato agregada lucro produto. No *step* Inserir Consolidado Lucro Produto relacione os campos da query com os campos da tabela Fato agregada (agr\_lucro\_produto).



Fonte: O autor.

Dados consolidados Lucro Produto.


Nesse momento, as transformações para a carga dos dados no DW Supermercado estão prontas. Agora, um novo conceito será apresentado para o entendimento sobre como são arquivados dados com baixa ou nenhuma frequência de utilização.

## Expurgo de Dados

O expurgo de dados consiste em retirar da base de dados do DW/DM os dados que não são mais acessados ou raramente são acessados. Geralmente, é definido um tempo em que os dados ficam armazenados e, a partir dessa data, os dados são recolhidos e armazenados em outros repositórios ou mídias que só serão utilizados se houver necessidade de realizar o backup das informações.

Se houver a necessidade de realizar o expurgo de dados, deve ser criada uma transformação que aponte quais dados devem ser arquivados e qual o período que deve ser expurgado.

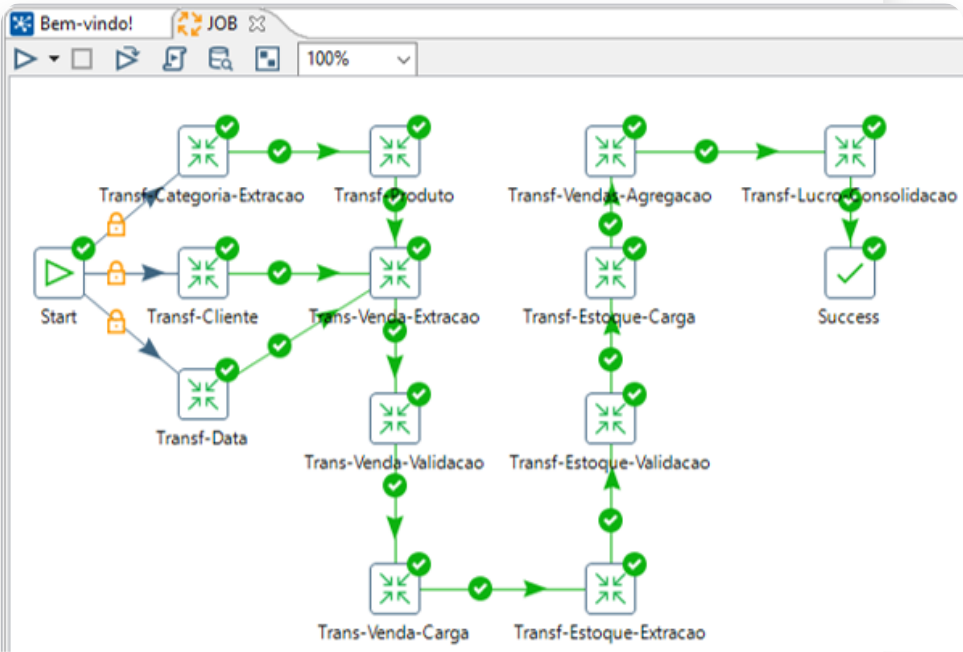
 JOB

 Clique no botão acima.

Após a criação das transformações, é hora de criar o *JOB* que encadeará as tarefas a serem executadas e que permite *scheduler* o processo para que ele seja executado em dias e horários determinados.

O Job deve começar sempre com o *step* Start. Nele, é possível definir quais são os dias e horários de execução do processo. Veja.

Processo de ETL do DW Supermercado.

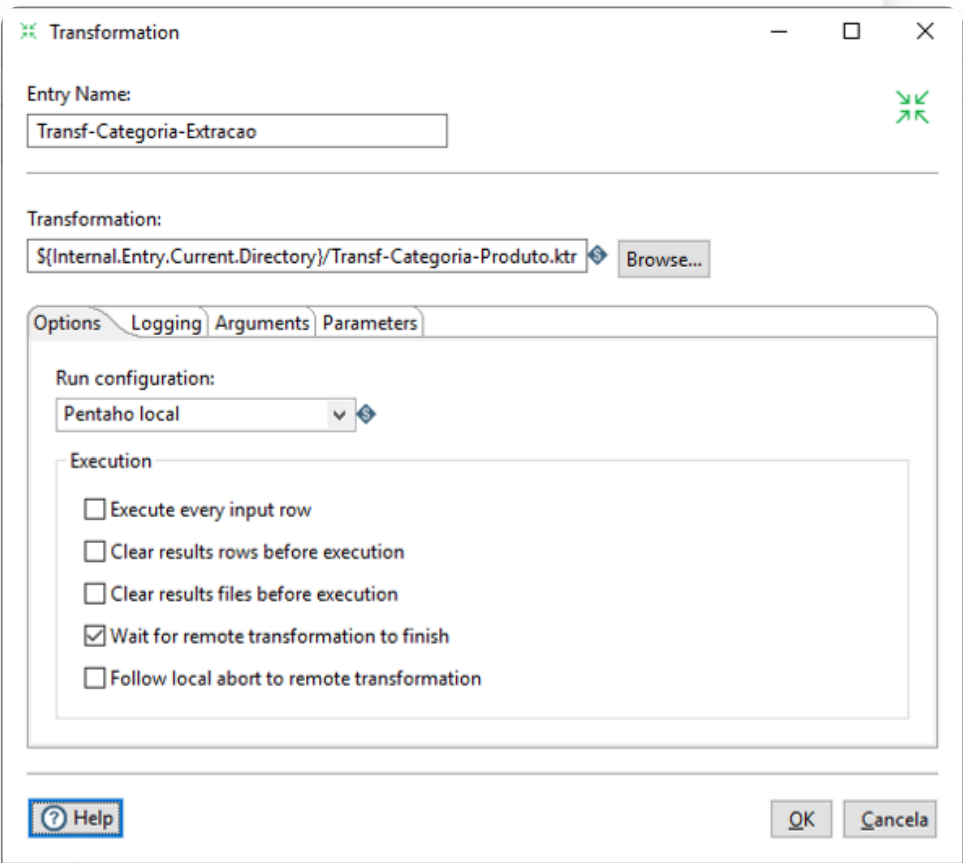


Fonte: O autor.

Em seguida, devem ser adicionados os *steps* Transformation, que apontam cada transformação criada. Essas devem ser colocadas na ordem de prioridade de execução, principalmente quando o dado contido em uma tabela precisa ser validado em outra tabela, por exemplo, a carga da tabela Fato precisa validar as SKs das dimensões que compõem a venda. Transformações que não dependem da execução de outra transformação podem ser colocadas em paralelo, como é o caso das transformações Categoria, Cliente e Data.

A Figura ilustra o exemplo de um *step* transformação que aponta para o arquivo da transformação que deverá ser executada.

Step Transformação Categoria.



Fonte: O autor.

Após a finalização do *JOB*, ele pode ser executado para processar toda a cadeia de processos de extração, transformação e carga dos dados.

Kimball (2013) descreve em seu livro que um projeto de DW/DM precisa oferecer confiabilidade, disponibilidade e gerenciabilidade: a confiabilidade garante que os processos serão executados de forma consistente; a disponibilidade deve garantir que o ambiente esteja pronto para uso quando preciso e a gerenciabilidade, no que diz respeito ao crescimento do ambiente em conformidade, suporta a confiabilidade do ambiente e sua disponibilidade.

Saiba mais

Conforme falado anteriormente, o processo de ETL precisa ser orquestrado por um *scheduler* que definirá o momento em que o processo irá iniciar e que inicie cada próxima tarefa obedecendo a execução do antecessor. Caso a ordem das tarefas não seja respeitada, possíveis erros podem ser apresentados e o objetivo principal de todo o projeto, que é a disponibilidade do ambiente analítico, não será alcançado.

Nesta aula verificamos as etapas de extração, tratamento e carga das tabelas Fato transacional, agregada e consolidada; vimos para que serve o expurgo de dados e falamos sobre gerenciamento de processos.

Agora, com base nos conceitos aplicados, vamos fixar o entendimento!

## Atividades

---

A carga dos dados em uma tabela Fato:

- a) Deve considerar as descrições de todos os elementos das dimensões para garantir a integridade dos dados.
  - b) Deve acontecer antes das dimensões para não haver problemas de integridade referencial.
  - c) Insere as chaves SKs antes da carga das medidas para garantir a integridade referencial.
  - d) Acontece após a carga das dimensões e valida as chaves SKs de cada uma das dimensões para que não haja problemas de integridade referencial.
  - e) Não possui validações, pois todos os dados foram validados na etapa de extração dos dados.
- 

O expurgo de dados do Data Warehouse (DW):

- a) Deve ser realizado ao completar um ano de informações para consolidar e arquivar os dados, mantendo sempre o último ano disponível para consultas.
  - b) Não deve contemplar dados históricos com menos de 10 anos de armazenamento.
  - c) Não se aplica a qualquer assunto de um DW.
  - d) Não deve ocorrer, porque pode prejudicar as análises realizadas no DW.
  - e) O período de remoção dos dados deve ser determinado pela organização e os dados removidos devem ser armazenados em mídia que permita recuperar os dados caso necessário.
- 

3. O processo de gerenciamento de processos definido por Kimball (2013) é composto por três pontos a serem verificados:

- a) Confiabilidade, disponibilidade e gerenciabilidade.
  - b) Processamento, disponibilidade e gerenciabilidade.
  - c) Confiabilidade, Validação e gerenciabilidade.
  - d) Confiabilidade, disponibilidade e Checagem.
  - e) Padronização, disponibilidade e gerenciabilidade.
- 

## Notas

## Título modal <sup>1</sup>

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 <sup>1</sup>

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.

## Próxima aula

- As operações de análise de dados;
- As ferramentas de OLAP;
- A construção das análises.

## Explore mais

- Conheça mais sobre o PDI e aprofunde os conhecimentos sobre os steps no site Hitachi Vantara.