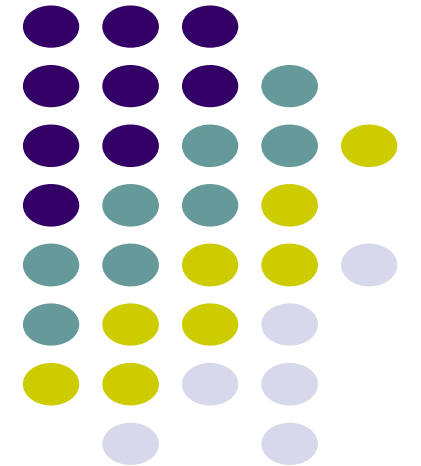


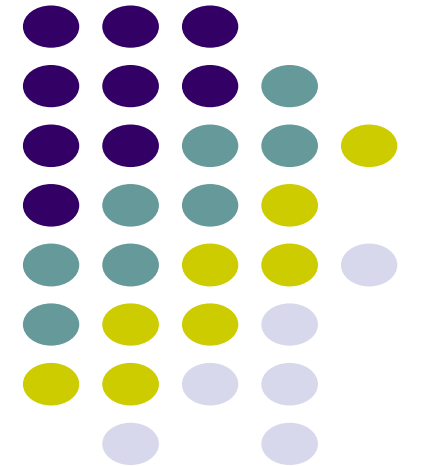
# Processo para criar o Data Stage e o Data Warehouse

---



# Processo para criar o Data Stage e o Data Warehouse

---





# DS – Dimensão Geografia no Data Stage

- Dimensão de Geografia é formada por grupo geográfico, País e região geográfica para o DW.





# DS – Dimensão Geografia no Data Stage

- Script da tabela chamada D\_RegiaoVendas.

```
CREATE TABLE D_RegiaoVendas  
(  
    Id_Pais int NOT NULL,  
    Nome varchar(20) NOT NULL,  
    LinData date NOT NULL,  
    LinOrig varchar(50) NOT NULL  
);
```

```
create index IX_D_RegiaoldPais on D_RegiaoVendas (Id_Pais);  
create index IX_D_RegiaoNome  on D_RegiaoVendas (Nome);
```

- Script da tabela chamada D\_RegiaoVendas.

```
CREATE TABLE D_RegiaoVendas(  
    Id_RegiaoVendas int NOT NULL,  
    Id_Pais int NOT NULL,  
    Nome varchar(20) NOT NULL,  
    LinData date NOT NULL,  
    LinOrig varchar(50) NOT NULL,  
    CONSTRAINT PK_D_RegiaoVendas PRIMARY KEY  
(  
        Id_RegiaoVendas  
    )  
);
```

- Script da tabela chamada D\_RegiaoVendas.

```
CREATE INDEX IX_D_RegiaoVendas ON D_RegiaoVendas  
(  
    Id_Pais  
);
```

```
ALTER TABLE D_RegiaoVendas ADD CONSTRAINT  
FK_D_RegiaoVendas_D_Pais FOREIGN KEY(Id_Pais)  
REFERENCES D_Pais (Id_Pais);
```



# DS – Dimensão Geografia no Data Stage

- As transformações no pentaho para fazer carga dos dados deverá responder pelos seguintes passos:
  - Carregar primeiro os dados da dimensão D\_GrupoGeografico;
  - Carregar após finalizada a carga da dimensão D\_GrupoGeografico os dados da dimensão D\_Pais;
  - Por fim, após finalizada a carga da dimensão D\_Pais os dados da Dimensão D\_Região\_Vendas.



# DS – Dimensão Geografia no Data Stage

- As transformações no pentaho para fazer carga dos dados deverá responder pelos seguintes critérios:
  - A carga dos dados é feita na ordem do menos granular para o mais granular, pois temos dependência dos registros. Ou seja, para inserir uma região de vendas, ela deve pertencer a um País previamente carregado.
  - Essa ordenação fará com que carreguemos o DS e o DW para cada uma das tabelas para depois seguir para a próxima, até finalizarmos.
  - Da mesma forma que a D\_Cliente, a chave de cada tabela será artificialmente criada por um autonumerador, a nossa Surrogate Key.





# DS – Dimensão Geografia no Data Stage

- As transformações no pentaho para fazer carga dos dados deverá responder pelos seguintes critérios:
  - A transformação e (quando houver) validação dos dados ocorrem nas tabelas da dimensão geografia no DS. Quando os dados forem ser inseridos no DW, já deverão estar ok.
  - A execução das mesmas transformações no Pentaho não acrescenta dados já existentes nas tabelas da dimensão geografia no Data Warehouse. Só vai ser inserido dados se os mesmos não existirem nele.
  - Teremos de colocar a fonte do dado e a data em que ele entrou para nossa base, como recurso de Lineage para cada uma das tabelas.



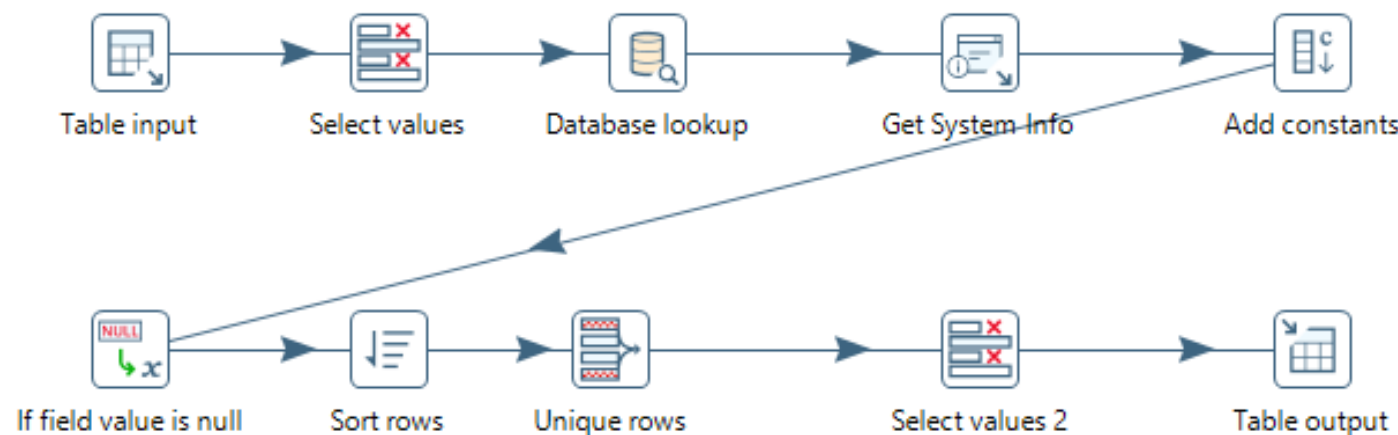
# DS – Dimensão Geografia no Data Stage

- Algumas observações:
  - Note que na definição das tabelas no DS e no DW indexamos os nomes (por onde as buscas ocorrerão) as chaves das tabelas “Pai”. Uma das regras da boa performance é que as chaves estrangeiras sejam sempre indexadas!



# DS – Dimensão Geografia no Data Stage

- A transformação seguinte mostra os passos para carga da D\_Regiao\_Vendas no Data Stage.





# DS – Dimensão Geografia no Data Stage

- Algumas observações:
  - Passo Table Input da transformação RegiãoVendas da dimensão Geografia possui os mesmos dados de entrada que os da dimensão Tempo.
  - O Passo Select Values da transformação RegiãoVendas da dimensão Geografia é similar ao passo Select Values BI da dimensão Tempo, só que nele tem-se os atributos de D\_Região\_Vendas que vem pelo fluxo.
  - Os passos Get System Info e Add constants tem a mesma informação da transformação cliente e ela é armazenada na tabela D\_Região\_Vendas.



# DS – Dimensão Geografia no Data Stage

- Passo Select values.

Nome do  
passo

Select values

Step name:

Select & Alter Remove Meta-data

Fields to alter the meta-data for :

#	Fieldname	Rename to	Type	Length
1	RegiaoVendas		None	50
2	Pais		None	50

Get fields to change

Help OK Cancela

Na aba Meta-data são selecionados 4 campos, conforme mostrado na figura. Para obter estes campos foi pressionado o botão Get fields to change e removidos os campos não pertencente a tabela D\_Regiao\_Vendas da dimensão Geografia.

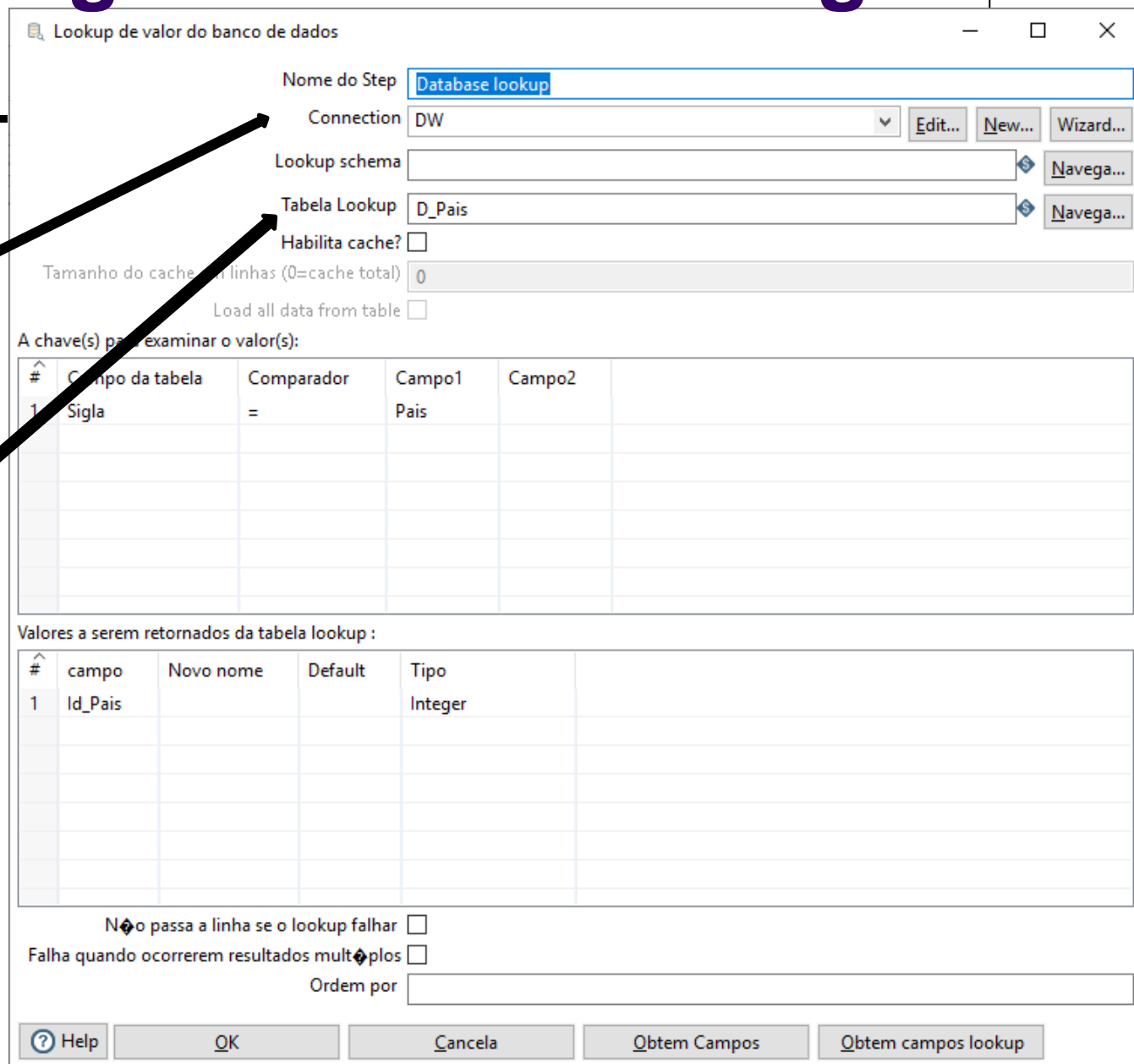


# DS – Dimensão Geografia no Data Stage

- Passo Database lookup.

Nome do passo e o nome da conexão ao banco de dados, no caso DW, que é criada através do botão New conforme já explicado para conexão ao banco de dados DS

A tabela de Lookup deste passo é a tabela D\_Pais que é selecionada pelo botão de navegação



#	Campo da tabela	Comparador	Campo1	Campo2
1	Sigla	=	Pais	

#	campo	Novo nome	Default	Tipo
1	Id_Pais			Integer



# DS – Dimensão Geografia no Data Stage

- Passo Database lookup.

Nesta área o item Campo da tabela tem atribuído o “Sigla” pertencente a tabela D\_Pais obtido pelo campo Obtem campos lookup. A coluna Campo1 obtido a partir do botão Obtem Campos que vem pelo fluxo e é Pais. Os valores deles são comparados e caso sejam iguais o Id\_Pais é recuperado e segue pelo fluxo.



Lookup de valor do banco de dados

Nome do Step: Database lookup

Connection: DW [Edit... New... Wizard...]

Lookup schema: [Navega...]

Tabela Lookup: D\_Pais [Navega...]

Habilita cache? ☐

Tamanho do cache em linhas (0=cache total): 0

Load all data from table ☐

A chave(s) para examinar o valor(s):

#	Campo da tabela	Comparador	Campo1	Campo2
1	Sigla	=	Pais	

Valores a serem retornados da tabela lookup :

#	campo	Novo nome	Default	Tipo
1	Id_Pais			Integer

Não passa a linha se o lookup falhar ☐

Falha quando ocorrerem resultados múltiplos ☐

Ordem por: [ ]

[?] Help [OK] [Cancela] [Obtem Campos] [Obtem campos lookup]



# DS – Dimensão Geografia no Data Stage

- Passo If field value is null

Na parte inferior, no item Fields utilizamos o botão **Obtêm campos** para selecionar os campos **RegiaoVendas**, **Id\_Pais** e **LinOrig** e na coluna **Replace by value** adicionar respectivamente **ZZ**, **1** e **Registro padrão inserido manualmente**.

Step name: **If field value is null**

Replace Null for all fields

Replace by value:

Set empty string? ☐

Mask (Date):

Select fields ☒

Select value type ☐

Value types

#	Type	Replace by value	Conversion mask (Date)	Set empty string?

Fields

#	Field	Replace by value	Conversion mask (Date)	Set empty string?
1	RegiaoVendas	XX		N
2	Id_Pais	1		N
3	LinOrig	Registro padrão inserido manualmente		N

Buttons: **Help**, **OK**, **Obtêm campos**, **Cancela**

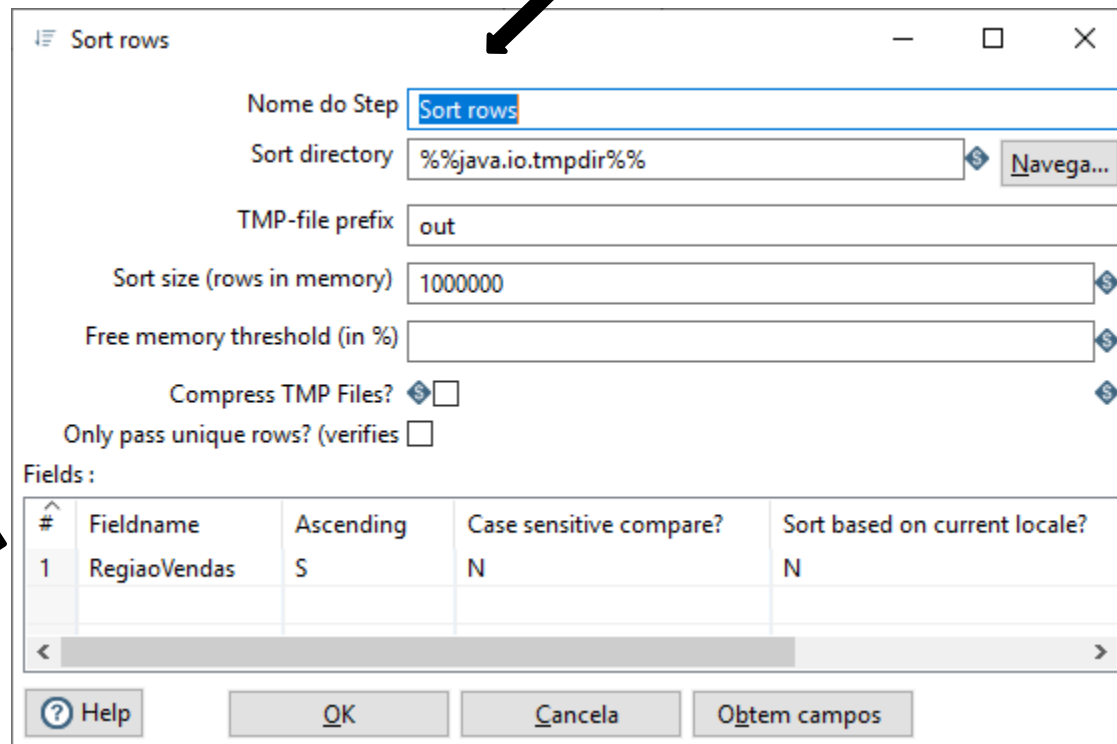




# DS – Dimensão Geografia no Data Stage

- Passo Sort rows

Na parte inferior, no item Fields utilizamos o botão Obtem campos para selecionar o campo RegiaoVendas e ordenar de forma ascendente.



Nome do passo

Nome do Step: Sort rows

Sort directory: %%java.io.tmpdir%% [Navega...](#)

TMP-file prefix: out

Sort size (rows in memory): 1000000

Free memory threshold (in %):

Compress TMP Files? ☐

Only pass unique rows? (verifies) ☐

Fields:

#	Fieldname	Ascending	Case sensitive compare?	Sort based on current locale?
1	RegiaoVendas	S	N	N

< >

Buttons: ? Help, OK, Cancela, Obtem campos



# DS – Dimensão Geografia no Data Stage

- Passo Unique rows

Na parte inferior, no item Fields to compare on utilizamos o botão Get para selecionar o campo RegiaoVendas.

A função deste passo é eliminar as linhas duplicadas.

Nome do passo

linhas únicas

Nome do Step Unique rows

Settings

Add counter to output? ☐ Counter field

Redirect duplicate row ☐ Error description

Fields to compare on (no entries means: compare complete row)

#	Fieldname	Ignore case
1	RegiaoVendas	N

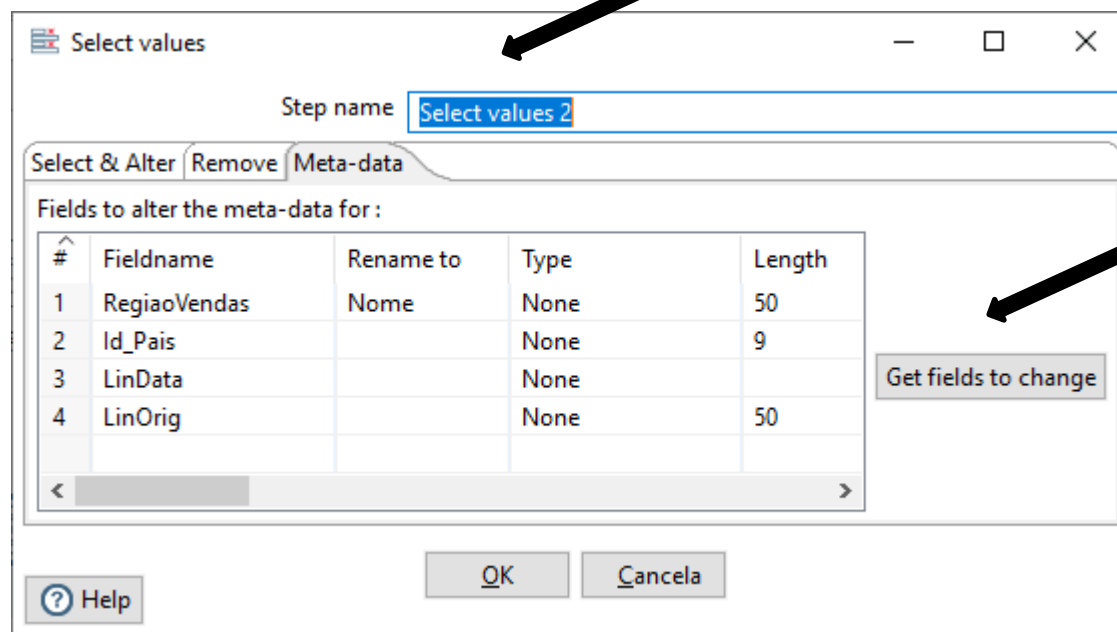
Help OK Cancela Get



# DS – Dimensão Geografia no Data Stage

- Passo Select values 2.

Nome do passo



Na aba Meta-data são selecionados 4 campos, conforme mostrado na figura. Para obter estes campos foi pressionado o botão Get fields to change e removidos os campos não pertencente a tabela D\_Regiao\_Vendas da dimensão Geografia.



# DS – Dimensão Geografia no Data Stage

- Passo Table output.

Nome do passo

Usado para criar uma conexão ao banco de dados DS para o Postgres.

Use o botão Navega para escolher a tabela alvo que no caso é D\_Regiao\_Vendas.

Usado para obter os campos que vem do passo anterior e associar com os campos da Tabela alvo D\_Regiao\_Vendas.

Nome do Step: Table output

Connection: DS

Target schema:

Target table: D\_RegiaoVendas

Commit size: 1000

Truncate table: ☒

Ignore insert errors: ☐

Specify database fields: ☒

Main options Database fields

Colunas a inserir:

#	Table field	Stream field
1	Nome	Nome
2	Id_Pais	Id_Pais
3	LinData	LinData
4	LinOrig	LinOrig

Get fields

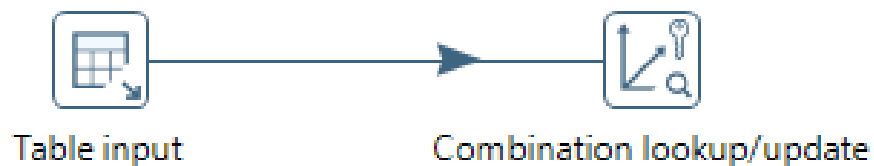
Enter field mapping

Help OK Cancela SQL



# DS – Dimensão Geografia no Data Warehouse

- A transformação seguinte mostra os passos para carga da D\_Região\_Vendas no Data Warehouse.





# DS – Dimensão Geografia no Data Warehouse

- Passo Table input.

Nome do passo

Letura de Tabela

Nome do Step:

Connection:

SQL:

```
SELECT
  Id_Pais
, Nome
, LinData
, LinOrig
FROM D_RegiaoVendas
```

Linha 1 Coluna 0

Store column info in step meta ☐

Enable lazy conversion ☐

Replace variables in script? ☐

Insert data from step

Executar para cada linha? ☐

Tamanho limite:

Obtém os nomes dos campos da tabela d\_RegiaoVendas no banco de dados DS

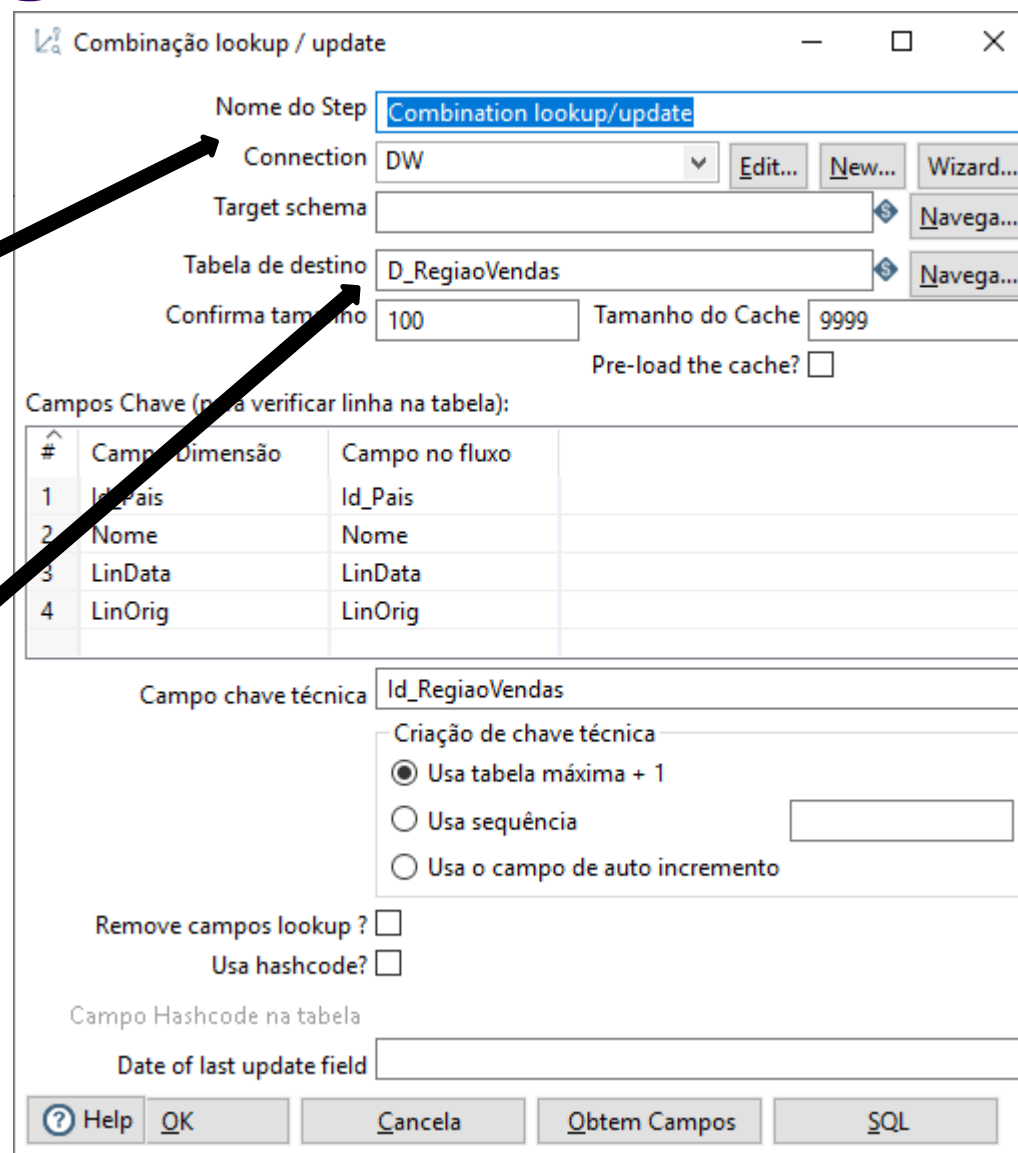


# DS – Dimensão Geografia no Data Warehouse

- Passo Combination lookup/update.

Nome do passo e o nome da conexão ao banco de dados, no caso DW, que é criada através do botão New conforme já explicado para conexão ao banco de dados DS

A tabela de destino deste passo é a tabela D\_RegiaoVendas que é selecionada pelo botão de navegação



Combinação lookup / update

Nome do Step: Combination lookup/update

Connection: DW

Target schema:

Tabela de destino: D\_RegiaoVendas

Confirma tamanho: 100

Tamanho do Cache: 9999

Pre-load the cache? ☐

Campos Chave (para verificar linha na tabela):

#	Campo Dimensão	Campo no fluxo
1	Id_Pais	Id_Pais
2	Nome	Nome
3	LinData	LinData
4	LinOrig	LinOrig

Campo chave técnica: Id\_RegiaoVendas

Criação de chave técnica

☒ Usa tabela máxima + 1

☐ Usa sequência

☐ Usa o campo de auto incremento

Remove campos lookup? ☐

Usa hashcode? ☐

Campo Hashcode na tabela

Date of last update field

Buttons: Help, OK, Cancela, Obtem Campos, SQL



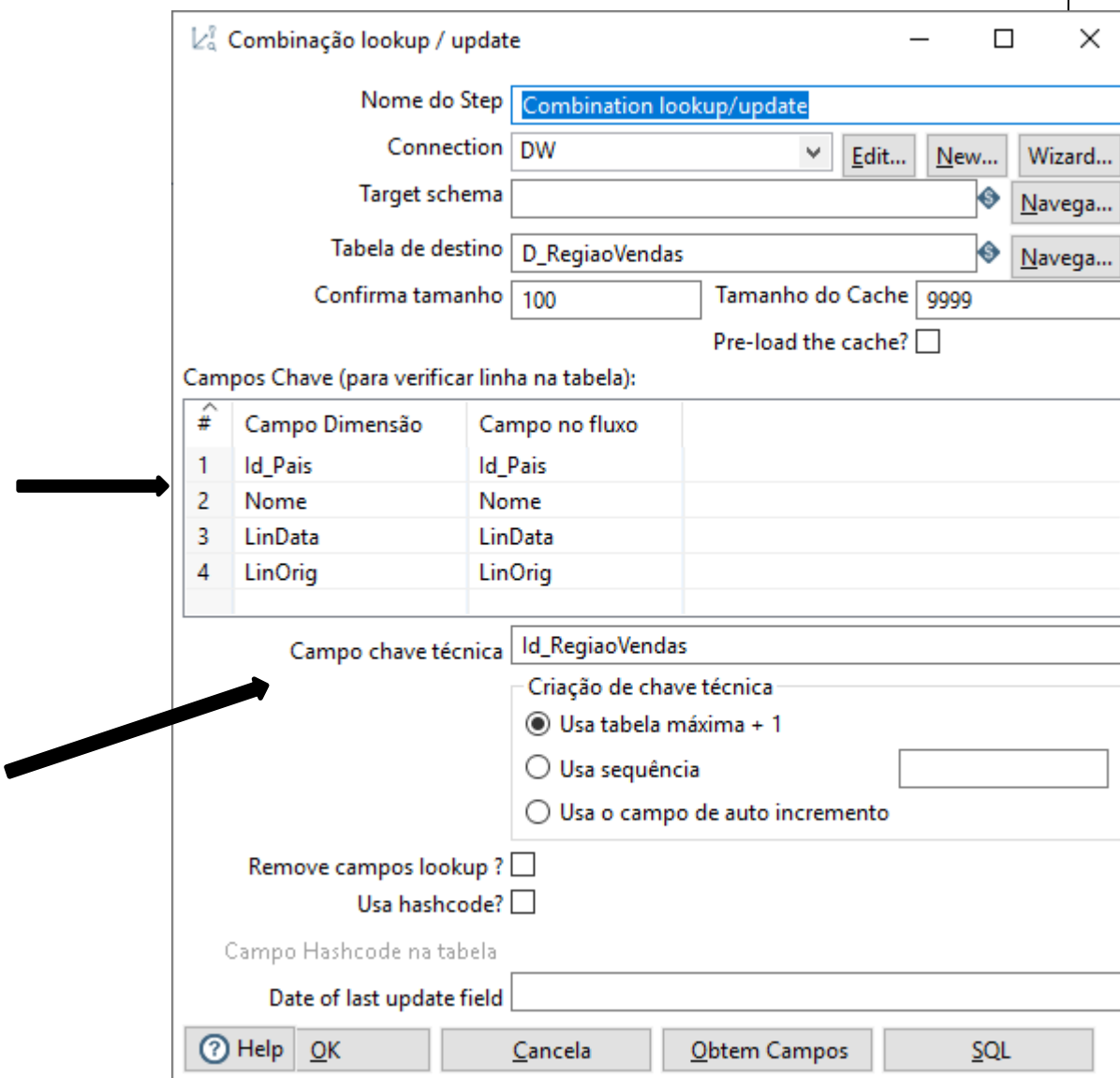
# DS – Dimensão Geografia no Data Warehouse

- Passo Combination lookup/update.

O item Campos chave permite fazer a comparação com os campos do fluxo e relação aos campos da dimensão caso os valores já existam na tabela D\_RegiaoVendas no DW nada é inserido caso contrário novos valores são inseridos.

Eles são obtidos a partir do botão Obtem Campos.

Campo chave técnica que é usada para criar os valores do campo Id\_RegiaoVendas a partir de 1 e é incrementado de uma unidade.



#	Campo Dimensão	Campo no fluxo
1	Id_Pais	Id_Pais
2	Nome	Nome
3	LinData	LinData
4	LinOrig	LinOrig





# DS – Dimensão Geografia no Data Warehouse



- Algumas observações:
  - Sempre apagamos a tabela D\_RegiaoVendas do DS para iniciar uma carga sem resquícios de cargas anteriores.
  - A transformação e (quando houver) validação dos dados ocorrem na inserção na tabela D\_RegiaoVendas no DS. Quando os dados forem ser inseridos no DW, já deverão estar ok.
  - A execução da mesma transformação no Pentaho não acrescenta dados já existentes na tabela D\_RegiaoVendas no Data Warehouse. Só vai ser inserido dados se os mesmos não existirem nele.



# DS – Dimensão Geografia no Data Stage

- Algumas observações:
  - Esse processo de carga se mostra um pouco mais complexo apenas por termos de capturar o valor da surrogate no DW da tabela Pai antes de carregarmos os dados da tabela Filho da mesma forma que na transformação País do DS.