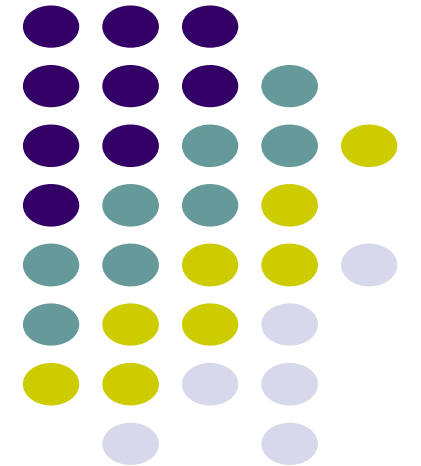
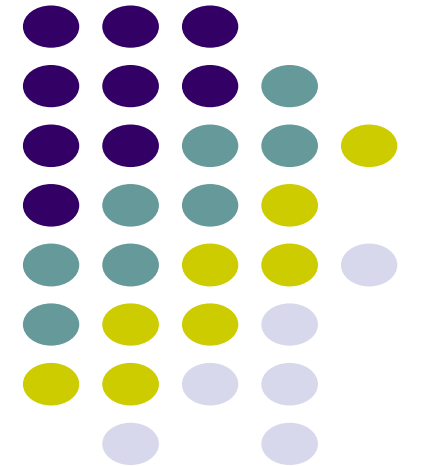


Processo para criar o Data Stage e o Data Warehouse



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_Pais.

```
CREATE TABLE D_Pais  
(  
    Id_GrupGeo int NOT NULL,  
    Sigla char(2) NOT NULL,  
    LinData date NOT NULL,  
    LinOrig varchar(50) NOT NULL  
);
```

```
create index IX_D_PaisIdGrupo on D_Pais (Id_GrupGeo);  
create index IX_D_PaisSigla on D_Pais (Sigla);
```



DW – Dimensão Geografia

- Script da tabela chamada D_Pais.

```
CREATE TABLE D_Pais(  
    Id_Pais int NOT NULL,  
    Id_GrupGeo int NOT NULL,  
    Sigla char(2) NOT NULL,  
    LinData date NOT NULL,  
    LinOrig varchar(50) NOT NULL,  
    CONSTRAINT PK_D_Pais PRIMARY KEY  
(  
        Id_Pais  
)  
);
```



DW – Dimensão Geografia

- Script da tabela chamada D_Pais.

```
CREATE INDEX IX_D_Pais ON D_Pais  
(  
    Id_grupoGeo  
);
```

```
ALTER TABLE D_Pais ADD CONSTRAINT  
FK_D_Pais_D_grupoGeografico FOREIGN KEY(Id_grupoGeo)  
REFERENCES D_grupoGeografico (Id_grupoGeo);
```



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.



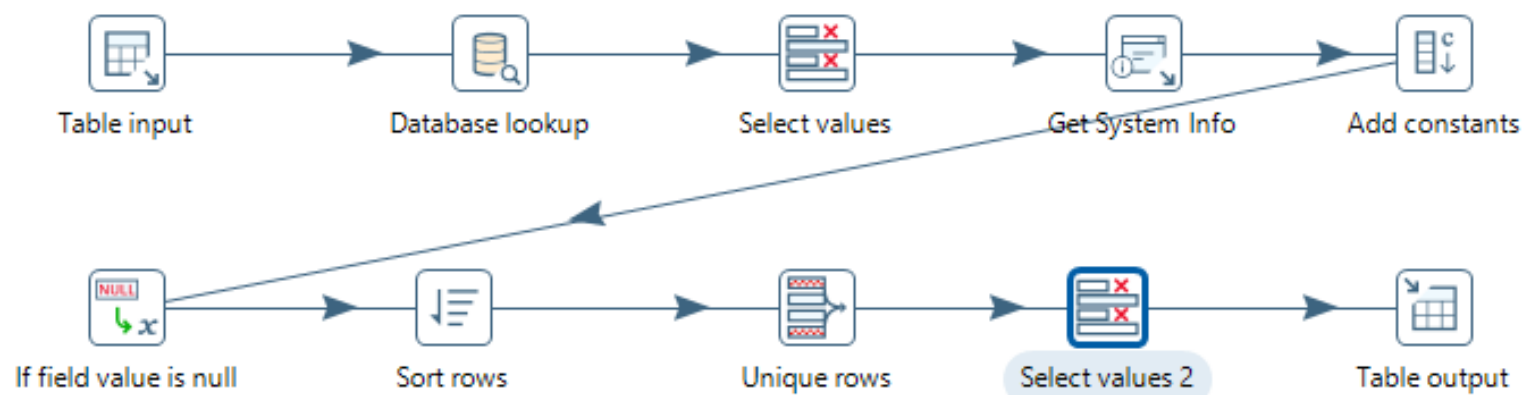
DS – Dimensão Geografia no Data Stage

- 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_Pais no Data Stage.





DS – Dimensão Geografia no Data Stage

- Algumas observações:
 - Passo Table Input da transformação D_Pais da dimensão Geografia possui os mesmos dados de entrada que os da dimensão Tempo.
 - O Passo Select Values da transformação D_Pais da dimensão Geografia é similar ao passo Select Values BI da dimensão Tempo, só que nele tem-se os atributos de Pais 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_Pais.

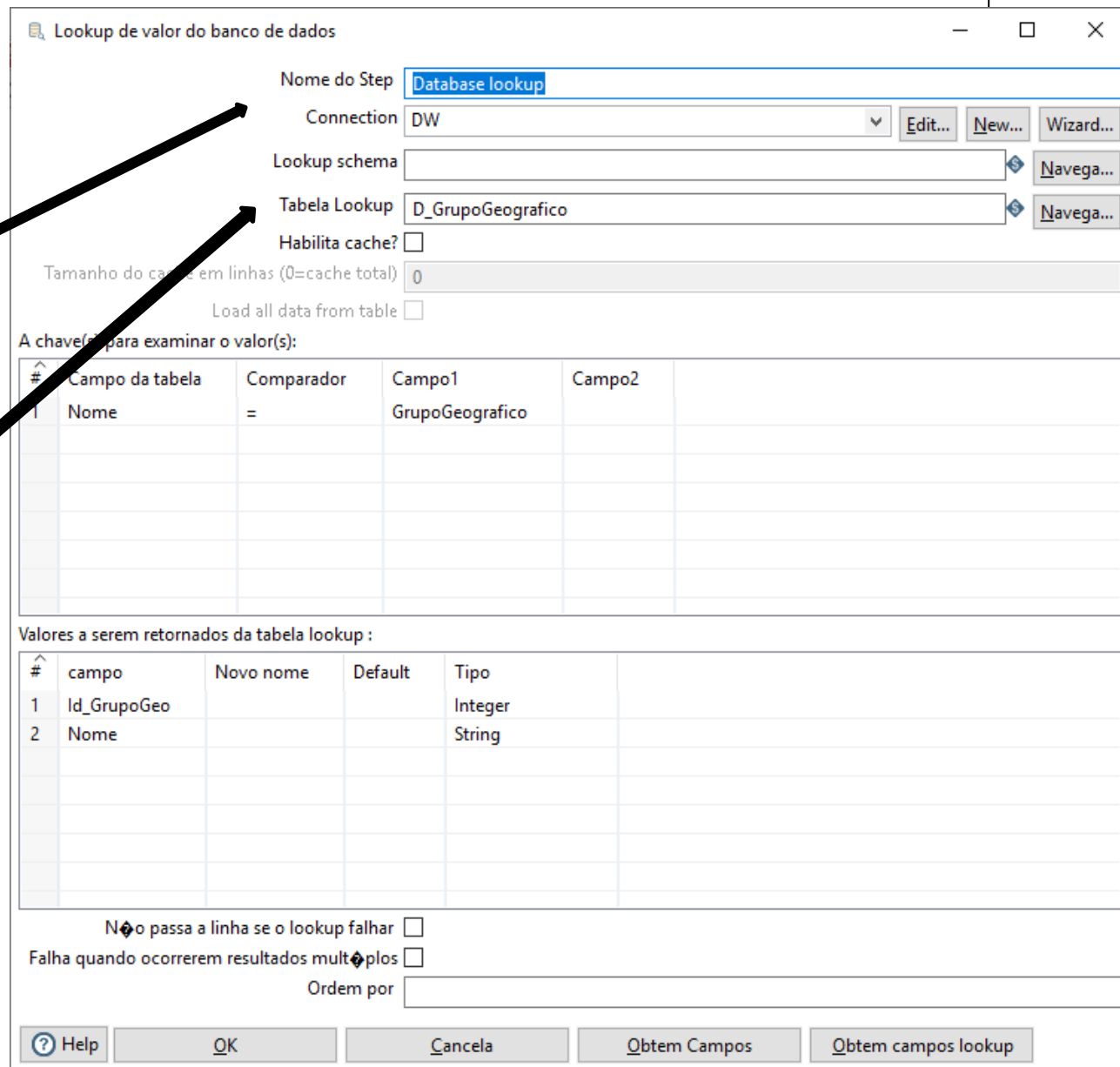


DS – Dimensão Geografia no Data Stage

- 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_GrupoGeografico que é selecionada pelo botão de navegação



Lookup de valor do banco de dados

Nome do Step: Database lookup

Connection: DW

Lookup schema:

Tabela Lookup: D_GrupoGeografico

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	Nome	=	GrupoGeografico	

Valores a serem retornados da tabela lookup:

#	campo	Novo nome	Default	Tipo
1	Id_GrupoGeo			Integer
2	Nome			String

☐ Não passa a linha se o lookup falhar

☐ Falha quando ocorrerem resultados múltiplos

Ordem por:

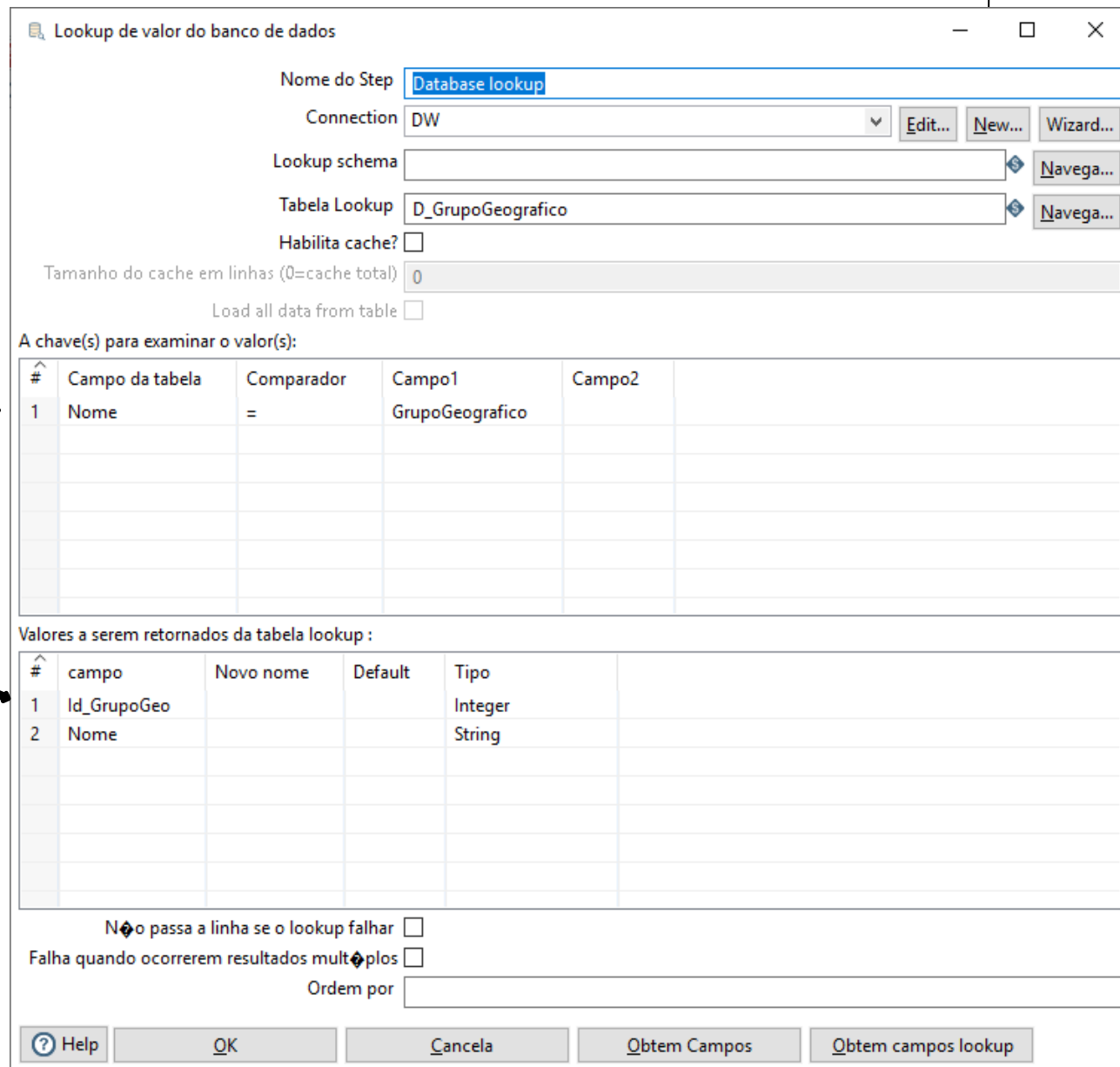
Buttons: Help, OK, Cancela, Obtem Campos, Obtem campos lookup



DS – Dimensão Geografia no Data Stage

- Passo Database lookup.

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



Lookup de valor do banco de dados

Nome do Step: Database lookup

Connection: DW

Lookup schema:

Tabela Lookup: D_GrupoGeografico

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	Nome	=	GrupoGeografico	

Valores a serem retornados da tabela lookup:

#	campo	Novo nome	Default	Tipo
1	Id_GroupGeo			Integer
2	Nome			String

Não passa a linha se o lookup falhar: ☐

Falha quando ocorrerem resultados múltiplos: ☐

Ordem por:

Buttons: Help, OK, Cancela, Obtem Campos, Obtem campos lookup



DS – Dimensão Geografia no Data Stage

- Passo Select values.

Nome do passo

Select values

Step name: Select values

Select & Alter Remove Meta-data

Fields to alter the meta-data for :

#	Fieldname	Rename to	Type	Length
1	Pais		None	50
2	GrupoGeografico		None	50
3	Id_GrupoGeo		None	9
4	Nome		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 GrupoGeografico da dimensão Geografia.



DS – Dimensão Geografia no Data Stage

- Passo If field value is null

Na parte inferior, no item Fields utilizamos o botão Obtem campos para selecionar os campos Pais, Id_GrupoGeo e LinOrig e na coluna Replace by value adicionar respectivamente ZZ, 1 e Registro padrão inserido manualmente.



If field value is null

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	Pais	ZZ		N
2	Id_GrupoGeo	1		N
3	LinOrig	Registro padrão inserido manualmente		N

< >

? Help OK Obtem 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 Pais e ordenar de forma ascendente.

Sort rows

Nome do Step:

Sort directory:

TMP-file prefix:

Sort size (rows in memory):

Free memory threshold (in %):

Compress TMP Files? ☐

Only pass unique rows? (verifies) ☐

Fields:

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

< >

Nome do
passo



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 Pais.
A função deste passo é eliminar as linhas duplicadas.

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	Pais	N

? Help OK Cancela Get

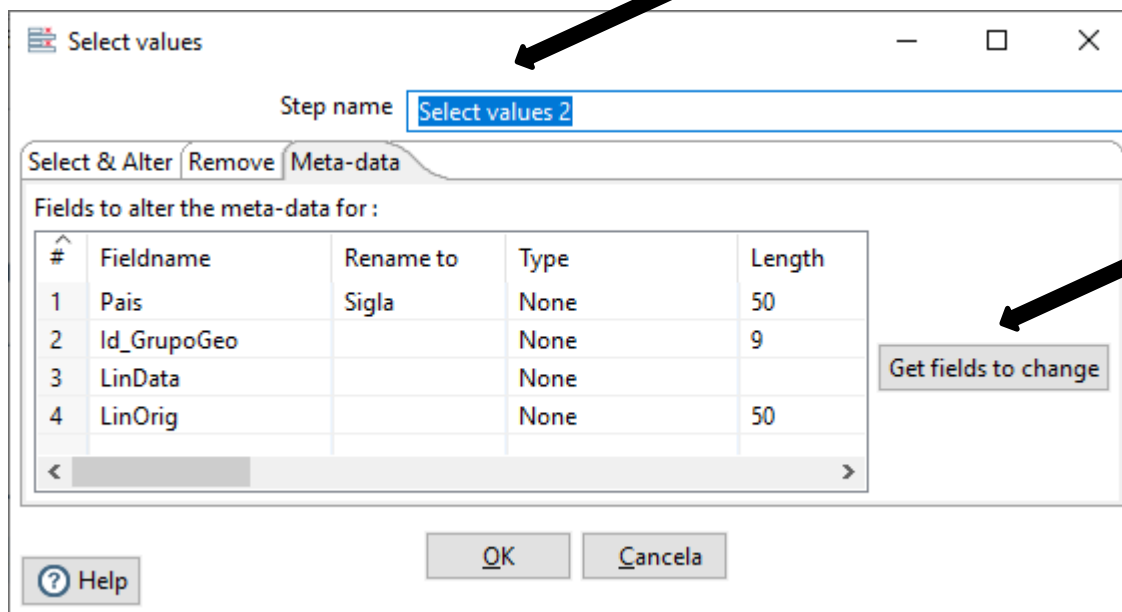
Nome do passo



DS – Dimensão Geografia no Data Stage

- Passo Select values 2.

Nome do passo



Select values

Step name: Select values 2

Select & Alter Remove Meta-data

Fields to alter the meta-data for:

#	Fieldname	Rename to	Type	Length
1	Pais	Sigla	None	50
2	Id_GrupoGeo		None	9
3	LinData		None	
4	LinOrig		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_Pais 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_Pais.

Usado para obter os campos que vem do passo anterior e associar com os campos da Tabela alvo D_Pais.

Nome do Step: Table output

Connection: DS

Target schema:

Target table: D_Pais

Commit size: 1000

Truncate table: ☒

Ignore insert errors: ☐

Specify database fields: ☒

Main options Database fields

Colunas a inserir:

#	Table field	Stream field
1	Sigla	Sigla
2	Id_GrupoGeo	Id_GrupoGeo
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_Pais 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_grupogeo
, sigla
, lindata
, linorig
FROM d_pais
```

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_pais 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_Pais que é selecionada pelo botão de navegação

Combinação lookup / update

Nome do Step: Combination lookup/update

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

Target schema: [Navega...]

Tabela de destino: D_Pais [Navega...]

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_GrupoGeo	Id_GrupoGeo
2	Sigla	Sigla
3	LinData	LinData
4	LinOrig	LinOrig

Campo chave técnica: Id_Pais

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: []

[?] Help OK Cancela Obtem Campos SQL

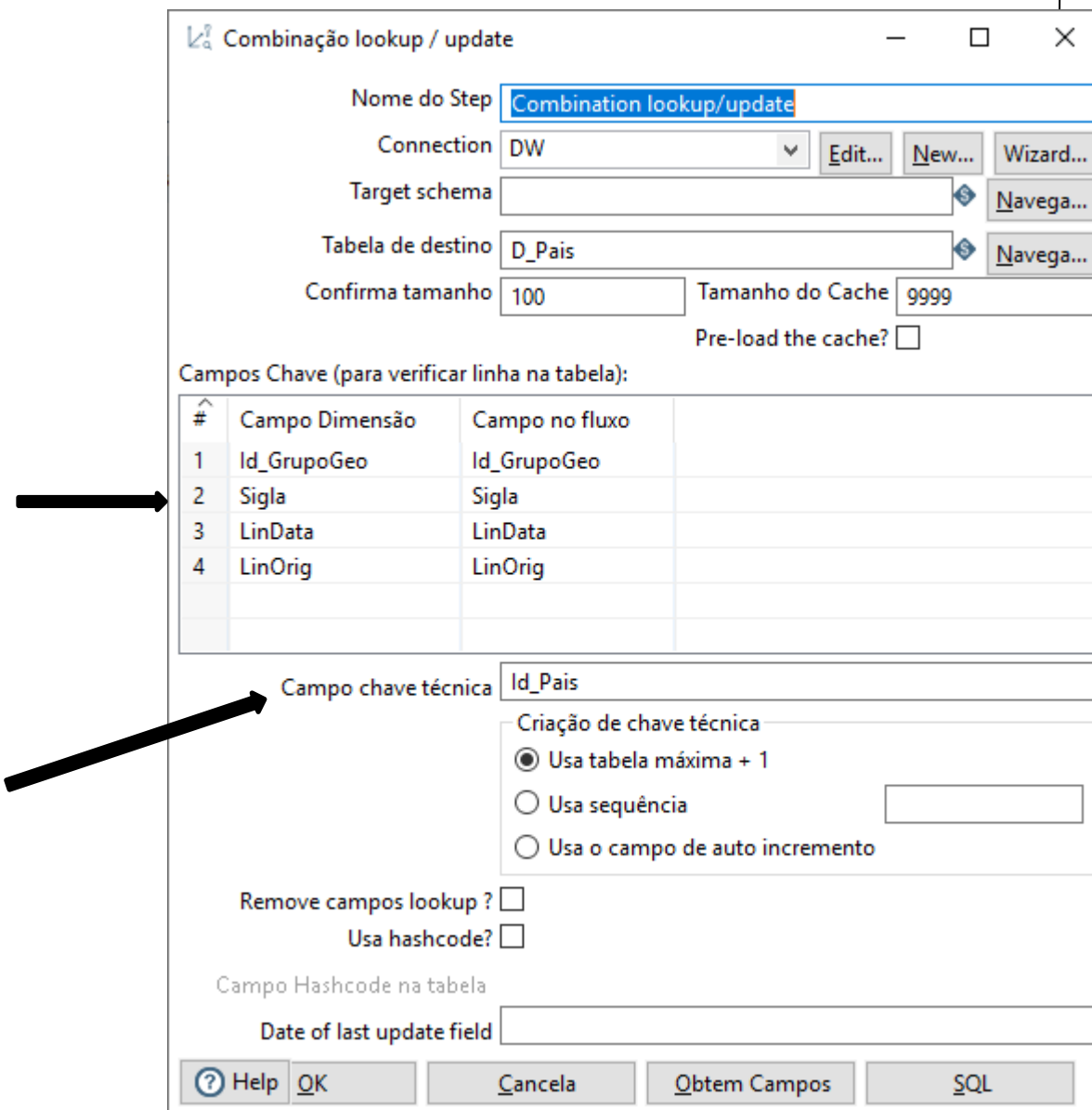


DS – Dimensão Geografia no Data Warehouse

- 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_Pais 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_Pais a partir de 1 e é incrementado de uma unidade.



#	Campo Dimensão	Campo no fluxo	
1	Id_GrupoGeo	Id_GrupoGeo	
2	Sigla	Sigla	
3	LinData	LinData	
4	LinOrig	LinOrig	



DS – Dimensão Geografia no Data Warehouse



- Algumas observações:
 - Sempre apagamos a tabela D_Pais 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_Pais 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_Pais 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. Esse passo que fazemos para preencher os dados ainda no Stage garante que teremos a tabela Filho com a devida surrogate quando a enviarmos para o DW. A indexação se faz necessária por conta das cargas do dia a dia que podem ter muitos e muitos registros, mesmo tratando-se de dimensões (que dificilmente superam as centenas de registros).