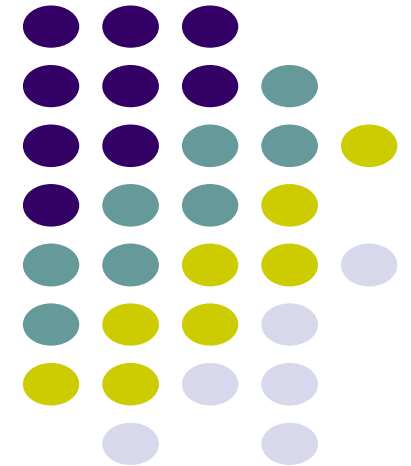


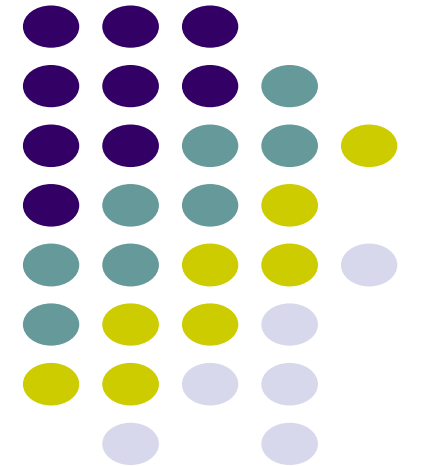


Business Intelligence: O desenho do Data Warehouse

Vitor Valerio de Souza Campos



Processo para criar o Data Stage e o Data Warehouse



- A tabela D_Cliente conterá os dados de Cliente e é composta de 5 atributos apresentados abaixo.

D_Cliente

Cod_Cliente (varchar(10))
Nome (varchar(50))
Email (varchar(50))
LinData (date)
LinOrig (varchar(50))



DS – Dimensão Cliente no Data Stage

- Script da tabela chamada D_Cliente.

```
CREATE TABLE D_Cliente(  
    Cod_Cliente varchar(10) NOT NULL,  
    Nome varchar(50) NOT NULL,  
    Email varchar(50) NOT NULL,  
    LinData date NOT NULL,  
    LinOrig varchar(50) NOT NULL  
);
```

```
create index IX_Cod_Cliente on D_Cliente (Cod_Cliente);
```



DS – Dimensão Cliente no Data Warehouse



- Script da tabela chamada D_Cliente.

```
CREATE TABLE D_Cliente(  
    Id_Cliente int NOT NULL default 0,  
    Cod_Cliente varchar(10) NOT NULL,  
    Nome varchar(50) NOT NULL,  
    Email varchar(50) NOT NULL,  
    LinData date NOT NULL,  
    LinOrig varchar(50) NOT NULL,  
    CONSTRAINT PK_D_Cliente PRIMARY KEY  
    (  
        Id_Cliente  
    ));  
CREATE INDEX IX_Cod_Cliente ON D_Cliente (Cod_Cliente);
```



DS – Dimensão Cliente no Data Stage

- As transformações no pentaho para fazer carga dos dados deverá levar em consideração que:
 - Teremos um Id_Cliente no DW que não existe no arquivo original do DS. É a chave artificial que, como vimos anteriormente, recebe o nome de Surrogate Key. Essa chave será criada de forma incremental. Note que ela não existe no DS. Por ser incremental, ela apenas existirá no destino do DW.
 - O código do Cliente, seu nome e seu e-mail serão carregados e atrelados a essa surrogate.
 - Teremos de colocar a fonte do dado e a data em que ele entrou para nossa Base, como recurso de Data Lineage.



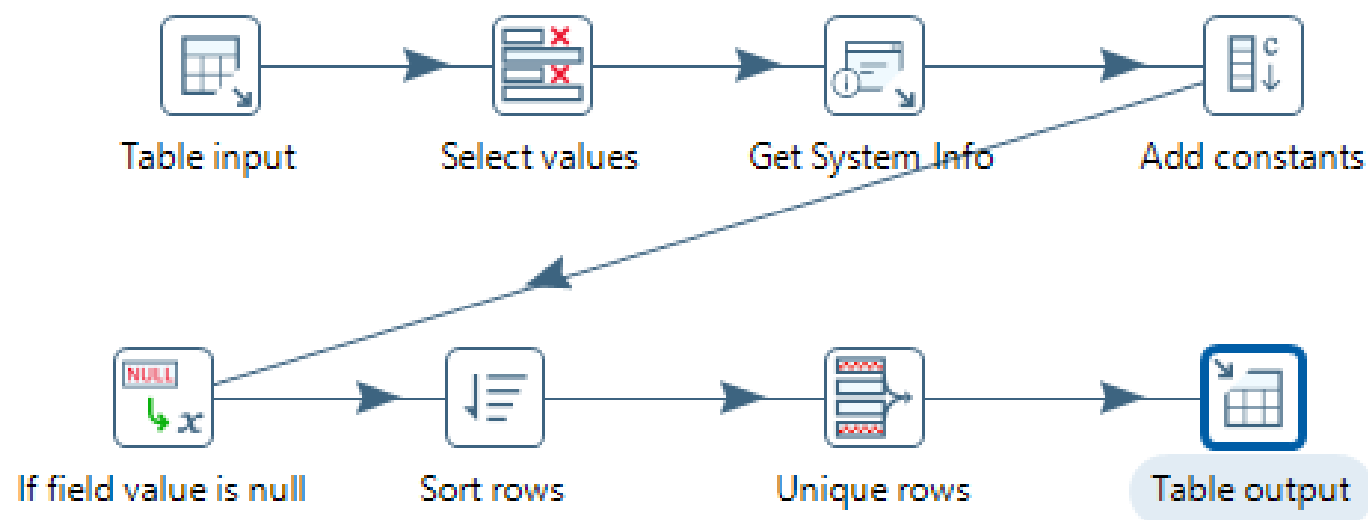
DS – Dimensão Cliente no Data Stage

- Algumas Observações:
 - Da mesma forma que em Data, é importante que o código do cliente seja indexado tanto no Data Stage quanto no Data Warehouse, e não uma Primary Key (que não existirá na tabela de Stage). Isso garantirá a performance da carga, uma vez que precisamos verificar se trata-se ou não de um registro novo justamente por esse campo.
 - Quando temos uma chave primária, temos automaticamente um índice. Quando não estamos usando uma chave primária para a pesquisa, como é o caso do código do cliente, temos de criar manualmente o índice para tornar a consulta mais rápida.



DS – Dimensão Cliente no Data Stage

- A transformação seguinte mostra os passos para carga da D_Cliente no Data Stage.





DS – Dimensão Cliente no Data Stage

- Algumas Observações:
 - Passo Table Input da dimensão Cliente possui os mesmos dados de entrada que os da dimensão Tempo.
 - O Passo Select Values da dimensão Cliente é similar ao passo Select Values BI da dimensão Tempo, só que nele tem-se os atributos de Cliente que vem pelo fluxo e neste passo eles são renomeados.



DS – Dimensão Cliente 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	CodCliente	Cod_Cliente	String	50
2	NomeCliente	Nome	String	50
3	EmailCliente	Email	String	50

Get fields to change

Help OK Cancela

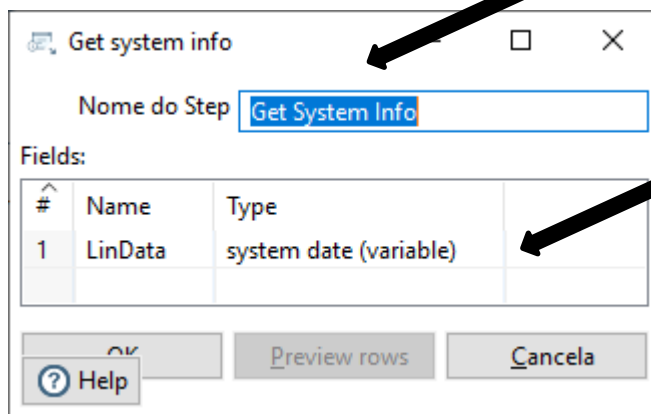
Na aba Meta-data são selecionados 3 campos, conforme mostrado na figura e os campos são renomeados. Para obter estes campos foi pressionado o botão Get fields to change e removidos os campos não pertencente a dimensão tempo.



DS – Dimensão Cliente no Data Stage

- Passo Get System Info.

Nome do passo



Get system info

Nome do Step:

Fields:

#	Name	Type
1	LinData	system date (variable)

Buttons: ? Help, Preview rows, Cancela

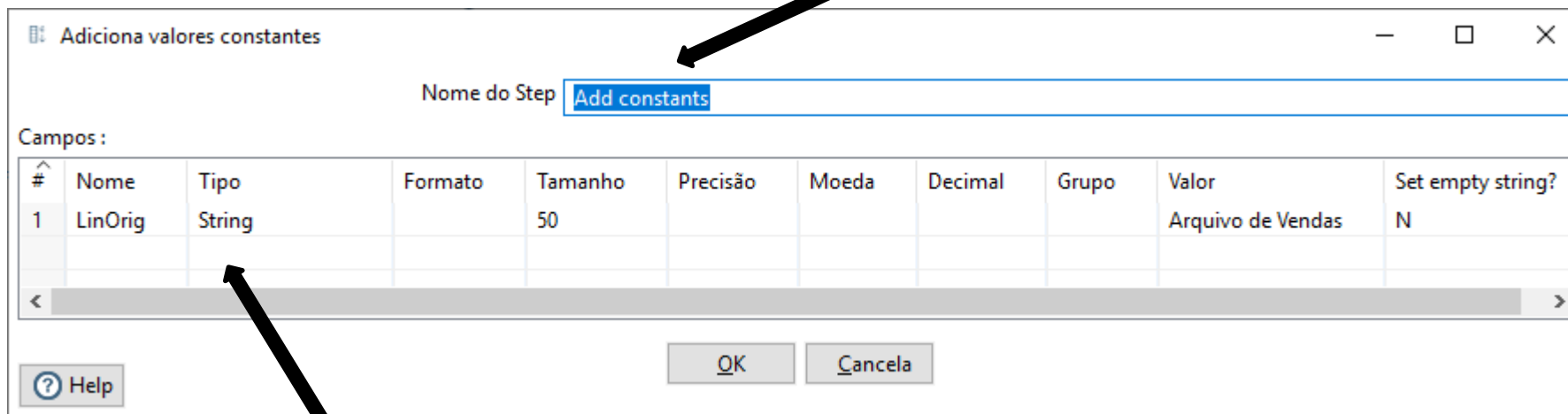
Em Fields, na coluna Name, digitamos LinData, que corresponde ao campo LinData da tabela Cliente, que deve ser usado para informar a data da carga dos dados de cliente no DW e na coluna Type selecionamos system date (variable) que insere a data do sistema no momento da execução da transformação, adicionando assim, ao fluxo de dados a data do sistema.



DS – Dimensão Cliente no Data Stage

- Passo Add constants.

Nome do passo



Adiciona valores constantes

Nome do Step: Add constants

Campos :

#	Nome	Tipo	Formato	Tamanho	Precisão	Moeda	Decimal	Grupo	Valor	Set empty string?
1	LinOrig	String		50					Arquivo de Vendas	N

OK Cancela

Help

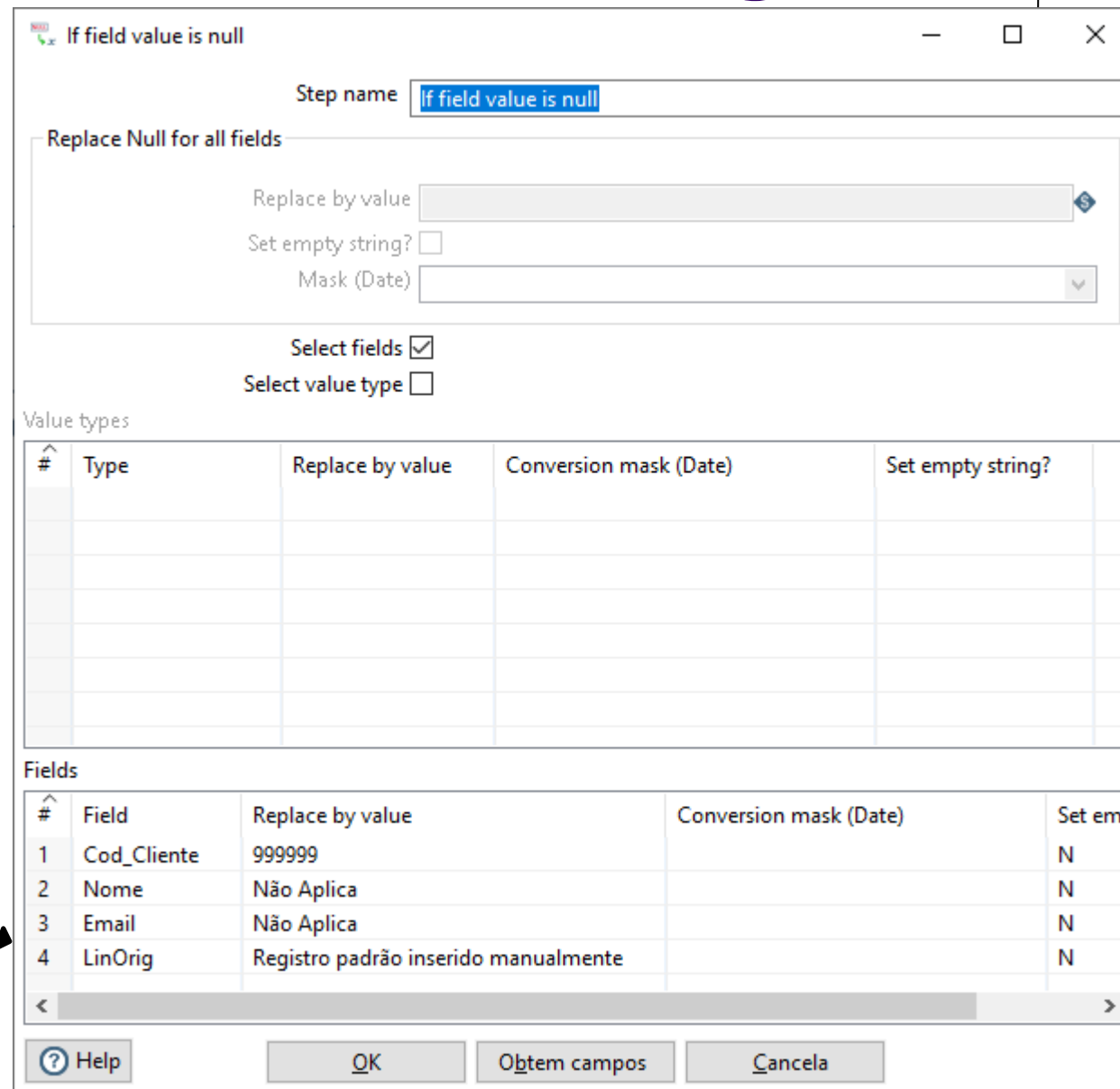
Em Campos é inserido na coluna Nome LinOrig, que corresponde ao campo LinOrig da tabela Cliente, seu tipo é definido como String e o seu tamanho como 50 e no campo Valor é inserido a string “Arquivo de Vendas”, por fim em Set empty string define-se como N.



DS – Dimensão Cliente no Data Stage

- Passo If field value is null

Na parte inferior, no item Fields utilizamos o botão Obtêm campos para selecionar os campos Cod_Cliente, Nome, Email, LinOrig e na coluna Replace by value adicionar respectivamente 999999, Não Aplica, Não Aplica e Registro padrão inserido manualmente.



If field value is null

Step name

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 em
1	Cod_Cliente	999999		N
2	Nome	Não Aplica		N
3	Email	Não Aplica		N
4	LinOrig	Registro padrão inserido manualmente		N

< >

Help OK Obtêm campos Cancela



DS – Dimensão Cliente no Data Stage

- Passo Sort rows

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

Nome do passo

Nome do Step: Sort rows

Sort directory: %%java.io.tmpdir%%

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	Cod_Cliente	S	N	N

Help OK Cancela Obtem campos



DS – Dimensão Cliente no Data Stage

- Passo Unique rows

Na parte inferior, no item Fields to compare on utilizamos o botão Get para selecionar os campos Cod_Cliente, Nome, Email, LinData e LinOrig.

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	Cod_Cliente	N
2	Nome	N
3	Email	N
4	LinData	N
5	LinOrig	N

Help OK Cancela Get



DS – Dimensão Cliente no Data Stage

- Passo Table output.

Saída a Tabela

Nome do Step: Table output

Connection: DS

Target schema:

Target table: D_Cliente

Commit size: 1000

Truncate table: ☒

Ignore insert errors: ☐

Specify database fields: ☒

Main options Database fields

Colunas a inserir:

#	Table field	Stream field
1	Cod_Cliente	Cod_Cliente
2	Nome	Nome
3	Email	Email
4	LinOrig	LinOrig
5	LinData	LinData

Buttons: Get fields, Enter field mapping, Help, OK, Cancela, SQL

Nome do passo

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

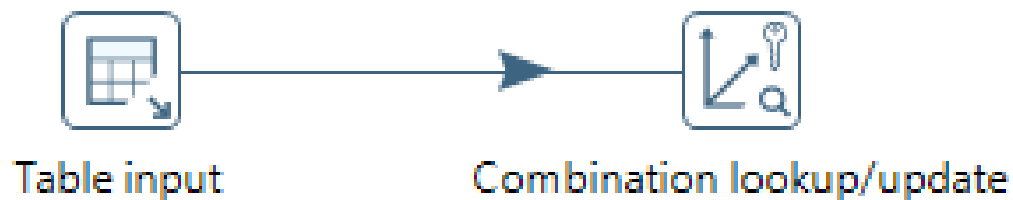
Use o botão Navega para escolher a tabela alvo que no caso é D_Cliente.

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



DS – Dimensão Cliente no Data Warehouse

- A transformação seguinte mostra os passos para carga da D_Cliente no Data Warehouse.





DS – Dimensão Cliente no Data Warehouse

- Passo Table input.

Nome do passo

Letura de Tabela

Nome do Step: Table input

Connection: DS

SQL:

```
SELECT
  cod_cliente
, nome
, email
, lindata
, linorig
FROM d_cliente
```

Linha 1 Coluna 0

Store column info in step ☐

Enable lazy conversion ☐

Replace variables in script? ☐

Insert data from step

Executar para cada linha? ☐

Tamanho limite: 0

Help OK Preview Cancela

Obtém os nomes dos campos da tabela d_cliente no banco de dados DS



DS – Dimensão Cliente 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_Cliente 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_Cliente [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	Cod_Cliente	Cod_Cliente
2	Nome	Nome
3	Email	Email
4	LinData	LinData
5	LinOrig	LinOrig

Campo chave técnica: Id_Cliente

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]



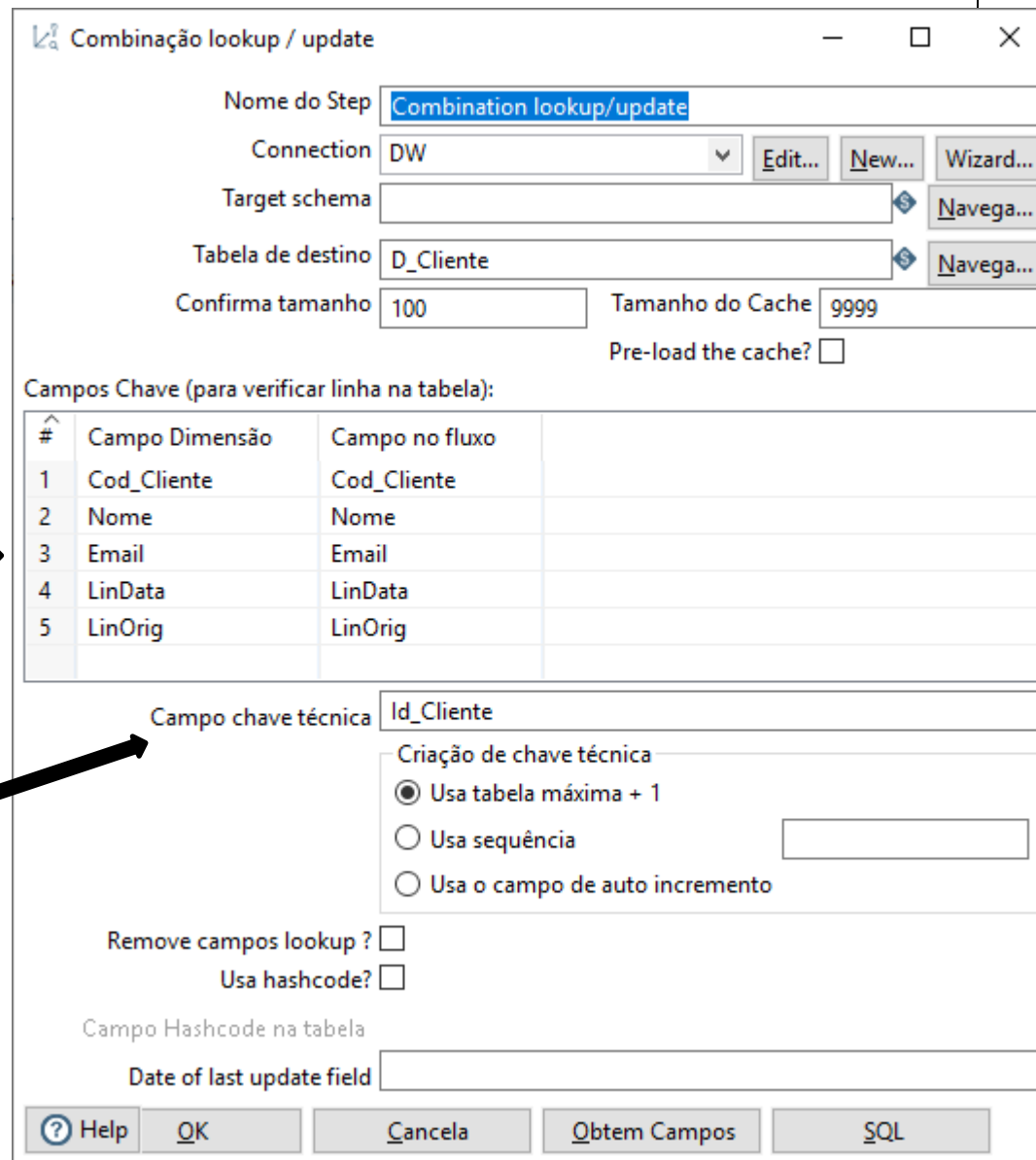
DS – Dimensão Cliente no Data Warehouse

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



Combinação lookup / update

Nome do Step: Combination lookup/update

Connection: DW

Target schema:

Tabela de destino: D_Cliente

Confirma tamanho: 100

Tamanho do Cache: 9999

Pre-load the cache? ☐

Campos Chave (para verificar linha na tabela):

#	Campo Dimensão	Campo no fluxo
1	Cod_Cliente	Cod_Cliente
2	Nome	Nome
3	Email	Email
4	LinData	LinData
5	LinOrig	LinOrig

Campo chave técnica: Id_Cliente

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 Cliente no Data Warehouse

- Algumas observações:
 - Sempre apagamos a tabela D_Cliente 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_Cliente 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_Cliente no Data Warehouse. Só vai ser inserido dados se os mesmos não existirem nele.



DS – Dimensão Cliente no Data Warehouse

- Algumas observações:
 - Também temos quando ele entrou no nosso DW e de onde ele veio para cada um dos valores carregados! Esta informação é útil quando formos confrontar os dados do BI com os apresentados pelas demais fontes em auditorias futuras.
 - Visto que temos a origem e a data, podemos facilmente acessar a pasta onde foram armazenados nossos arquivos e achar o arquivo que originou a informação pela data de carga dele.