FIAP GRADUAÇÃO

TECNOLOGIA EM DESENVOLVIMENTO DE SISTEMAS

Enterprise Analytics e *Data Warehousing*

ETL: Exercício Prático

PROFA. FERNANDA P. CAETANO proffernanda.caetano@fiap.com.br

PROF. SALVIO PADLIPSKAS salvio@fiap.com.br

FIAP

EXEMPLO DE ETL REALIZADO PELA FERRAMENTA PENTAHO DATA INTEGRATOR (PDI)

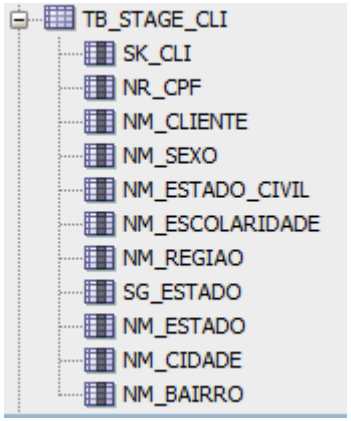
AGENDA

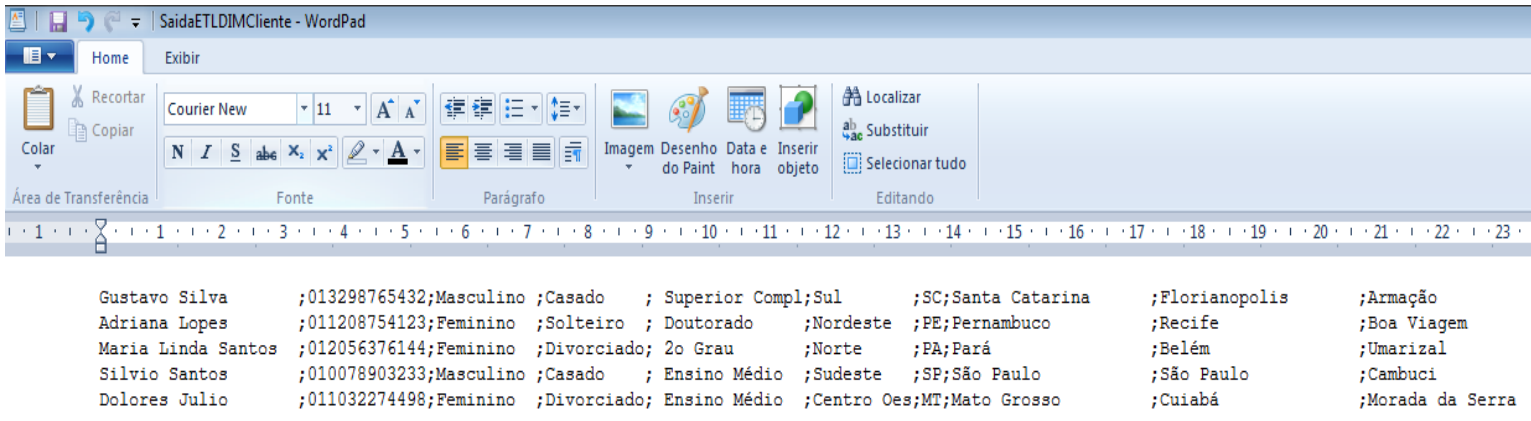
- Exemplo prático ETL com ferramenta Pentaho PDI
- ETL com origens de dados em planilha Excel e arquivo flat file
- Transformação de dados
- Geração de surrogate key
- Carga de dados

Exemplo de ETL com diversas origens

Nosso objetivo será criar um processo ETL que irá realizar a leitura e **E**xtração de origens de dados em planilha Excel e arquivo flat file, aplicar **T**ransformações (sexo e estado civil) e realizar o **L**oad em um arquivo TXT e em uma base de dados Oracle.

Abaixo temos o processo a ser construído utilizando a ferramenta Pentaho Data Integrator (PDI).

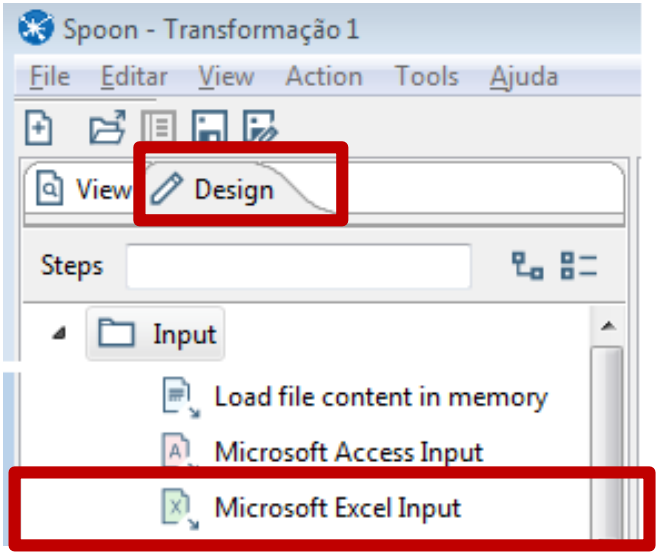




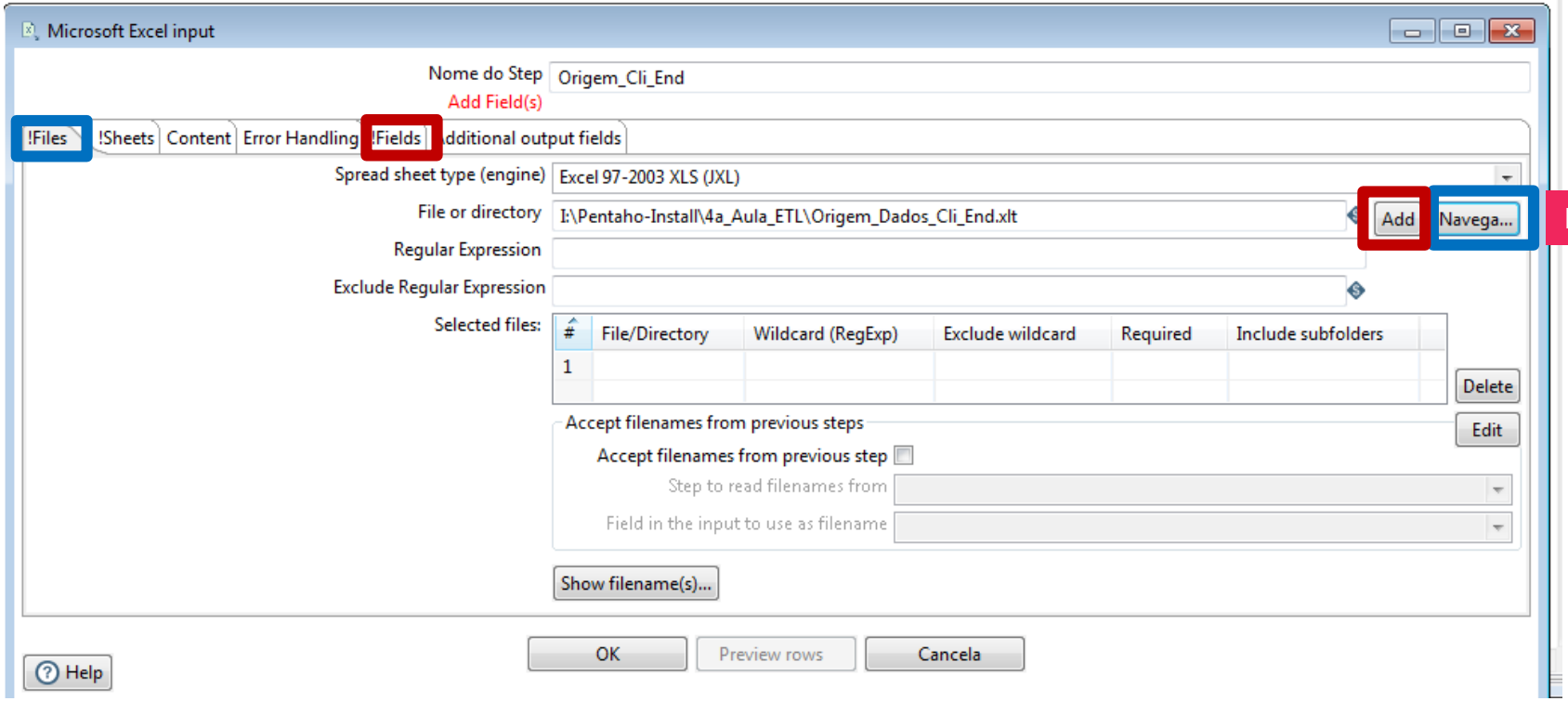
1º passo: Setup origem de dados do tipo planilha Excel

Acesso a ferramenta pelo arquivo **Spoon.bat**. Nesse exemplo utilizamos o diretório **C:\Pentaho\data-integration**

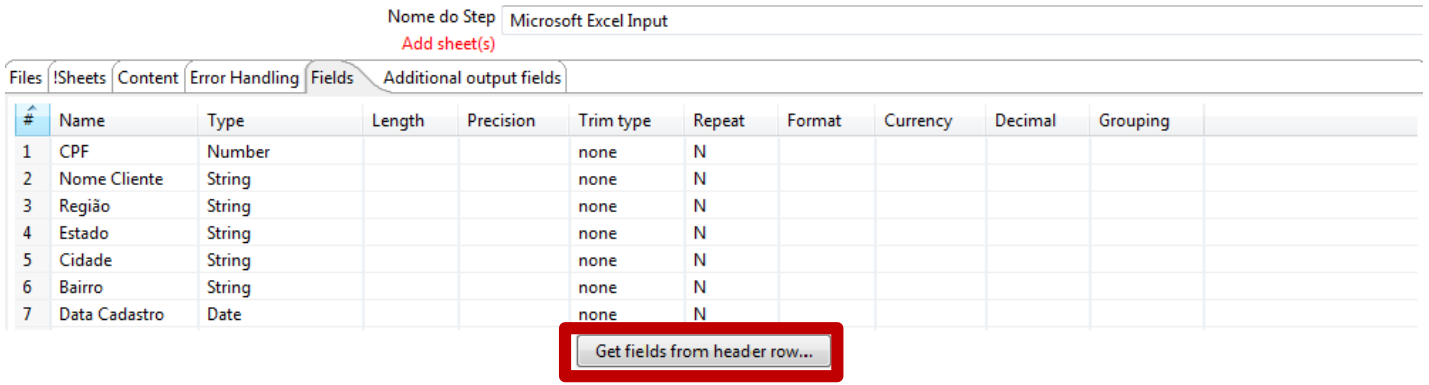
Pela ferramenta Pentaho PDI acesse a aba “Input” e selecione a opção “Microsoft Excel Input”. Essa será a nossa primeira origem de dados a ser utilizada na fase de ETL.



1º passo: Setup origem de dados do tipo planilha Excel

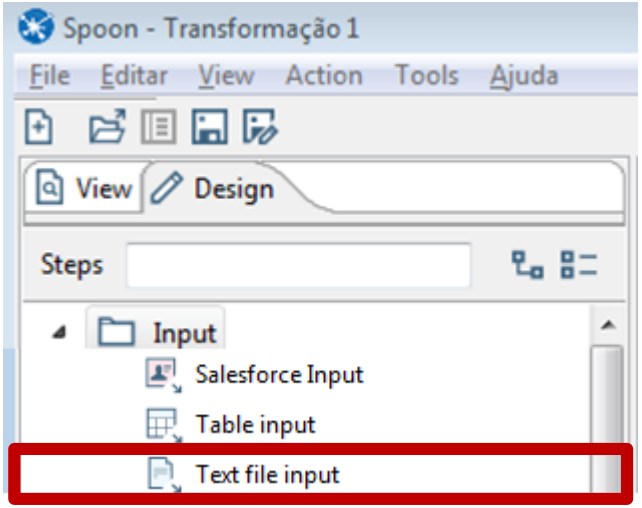


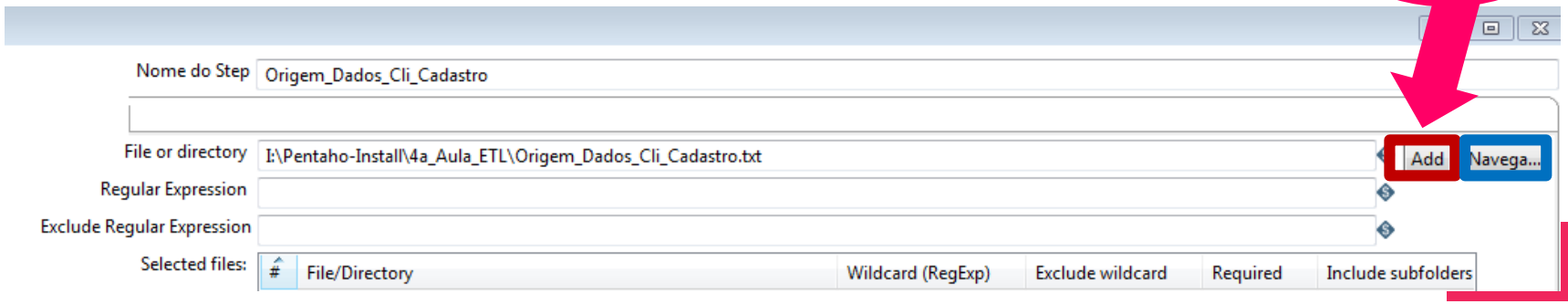
Botão Add



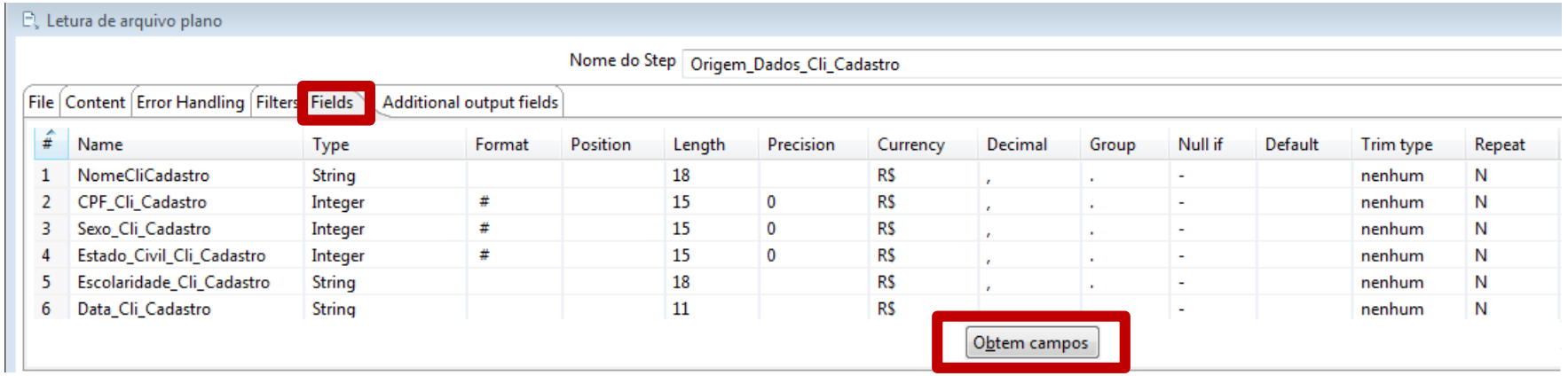
2º passo: Setup origem de dados do tipo arquivo flat file

Pela ferramenta Pentaho PDI acesse a aba “Input” e selecione a opção “Text file input”. Essa será a nossa segunda origem de dados a ser utilizada na fase de ETL





2º passo: Setup origem de dados do tipo arquivo flat file



#	Name	Type	Format	Position	Length	Precision	Currency	Decimal	Group	Null if	Default	Trim type	Repeat
1	NomeCliCadastro	String			18		R$	,	.	-		nenhum	N
2	CPF_Cli_Cadastro	Integer	#		15	0	R$	,	.	-		nenhum	N
3	Sexo_Cli_Cadastro	Integer	#		15	0	R$	,	.	-		nenhum	N
4	Estado_Civil_Cli_Cadastro	Integer	#		15	0	R$	,	.	-		nenhum	N
5	Escolaridade_Cli_Cadastro	String			18		R$	,	.	-		nenhum	N
6	Data_Cli_Cadastro	String			11		R$			-		nenhum	N

Rows of step: Origem_Dados_Cli_Cadastro (5 rows)

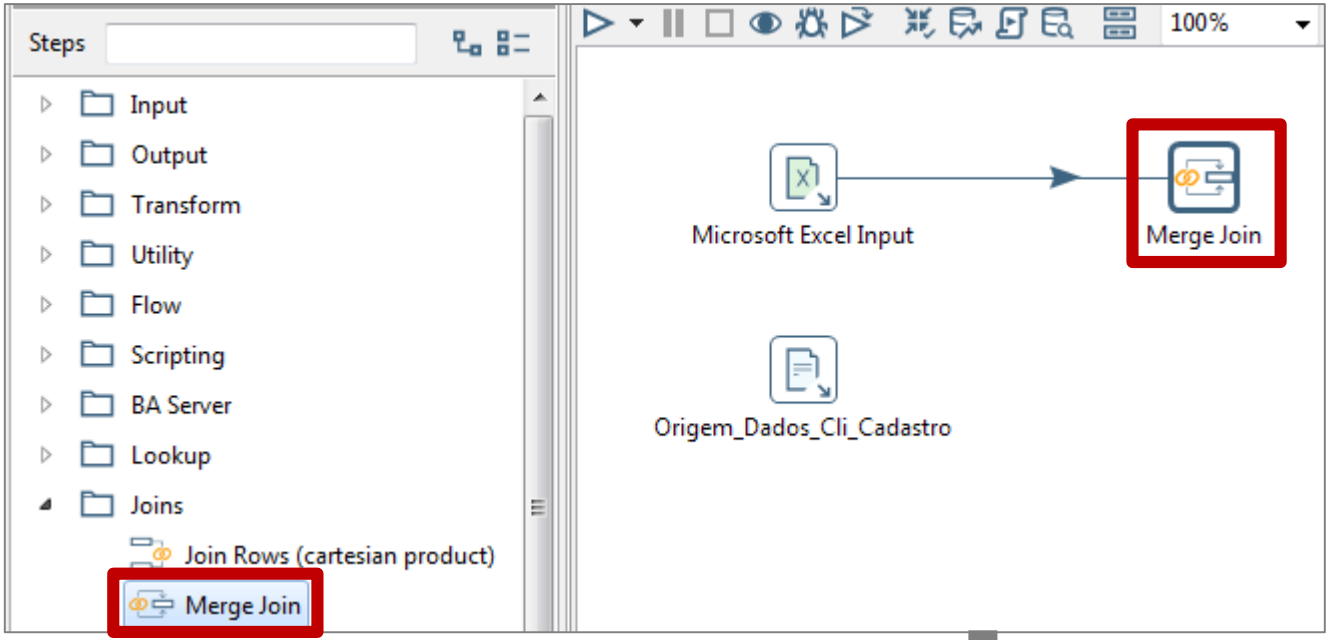
#	Nome Cli Cadastro	CPF Cli Cadastro	Sexo Cli Cadastro	Estado Civil Cli Cadastro	Escolaridade Cli Cadastro	Data Cadastro
1	Gustavo Silva	13298765432	1	2	Superior Completo	18/01/2016
2	Adriana Lopes	11208754123	2	1	Doutorado	15/03/2016
3	Maria Linda Santos	12056376144	2	3	2o Grau	31/08/2016
4	Silvio Santos	10078903233	1	2	Ensino Médio	19/09/2016
5	Dolores Julio	11032274498	2	3	Ensino Médio	13/03/2016

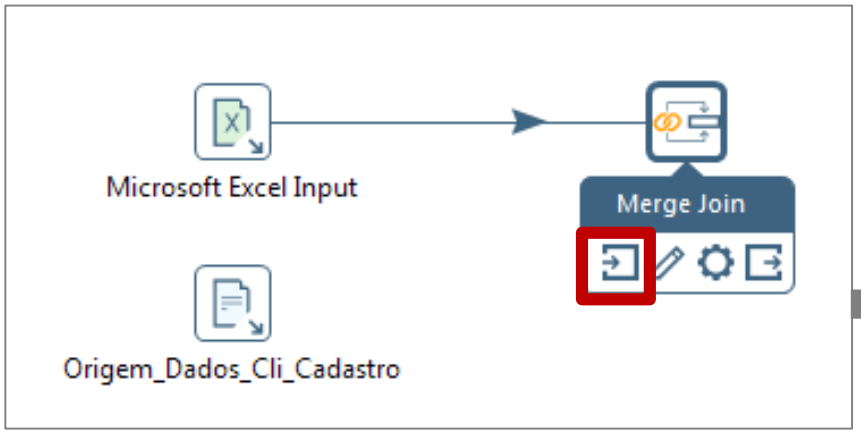
Origem_Dados_Cli_End

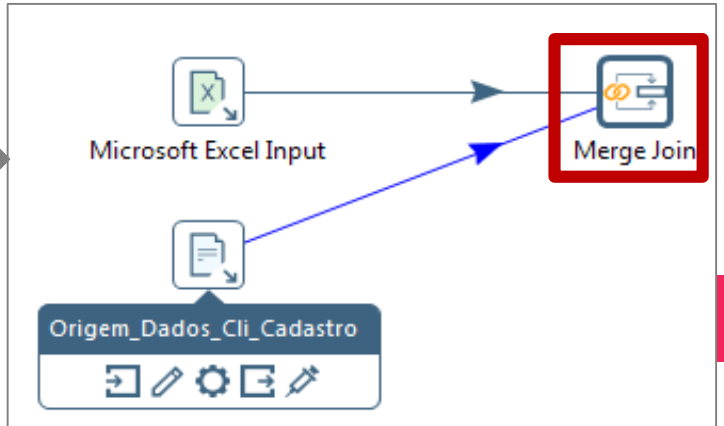
Origem_Dados_Cli_Cadastro

3º passo: Unificando os dados das 2 origens

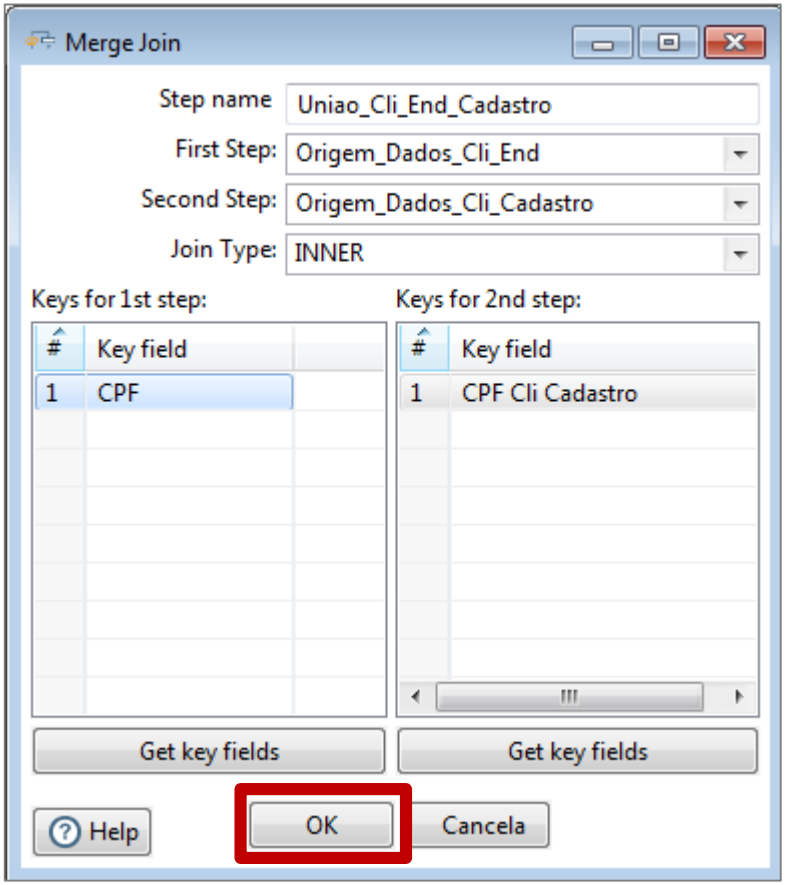
(Cliente Endereço e Cliente Cadastro)

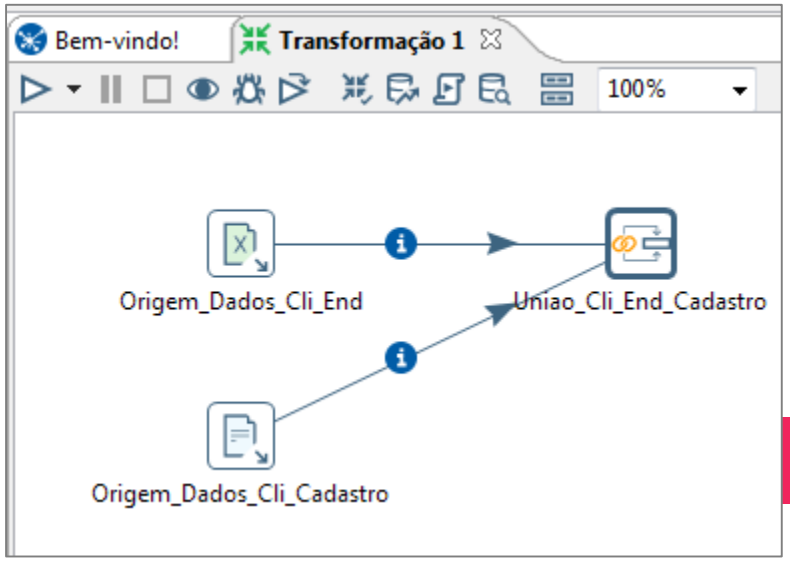






3º passo: Unificando os dados das 2 origens
(Cliente Endereço e Cliente Cadastro)

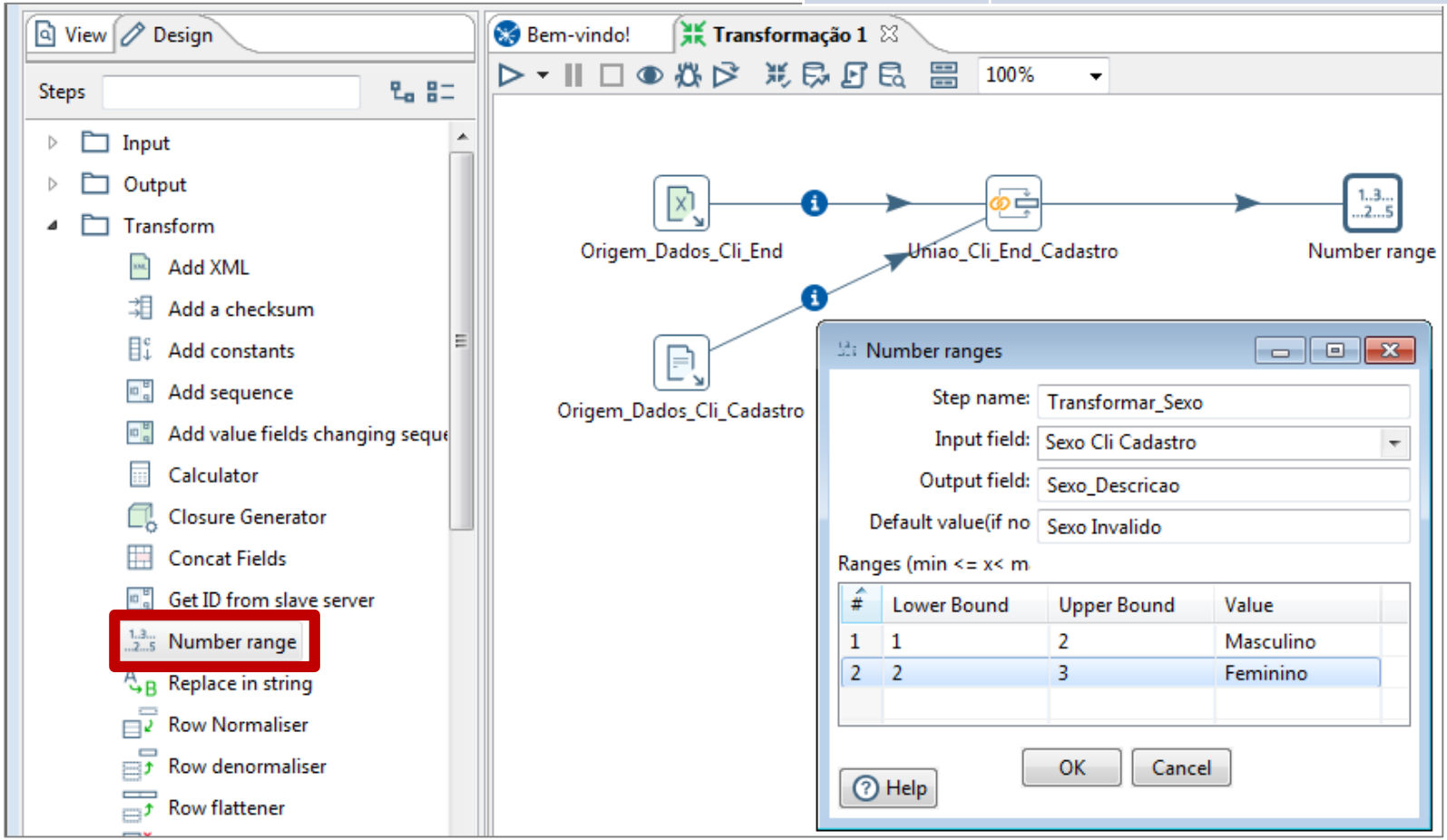




4º passo: Transformação do Sexo do Cliente

A origem de dados contém códigos para identificar o sexo do cliente. No modelo dimensional, para evitar futuros cálculos e prejudicar o desempenho, se faz necessário realizar essa importante transformação.

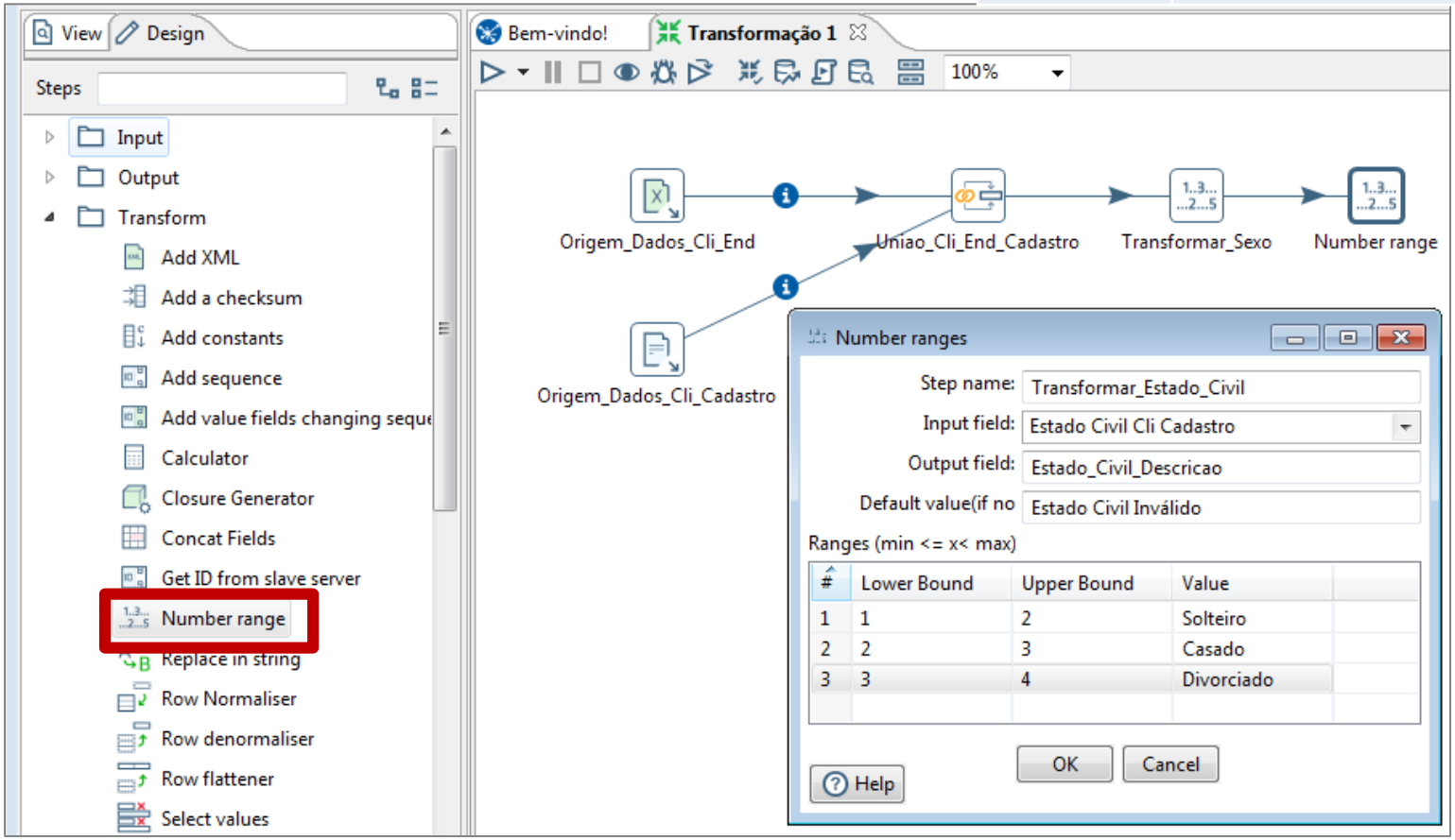
Valor original	Novo valor
1	Masculino
2	Feminino
Outro valor	Sexo Inválido



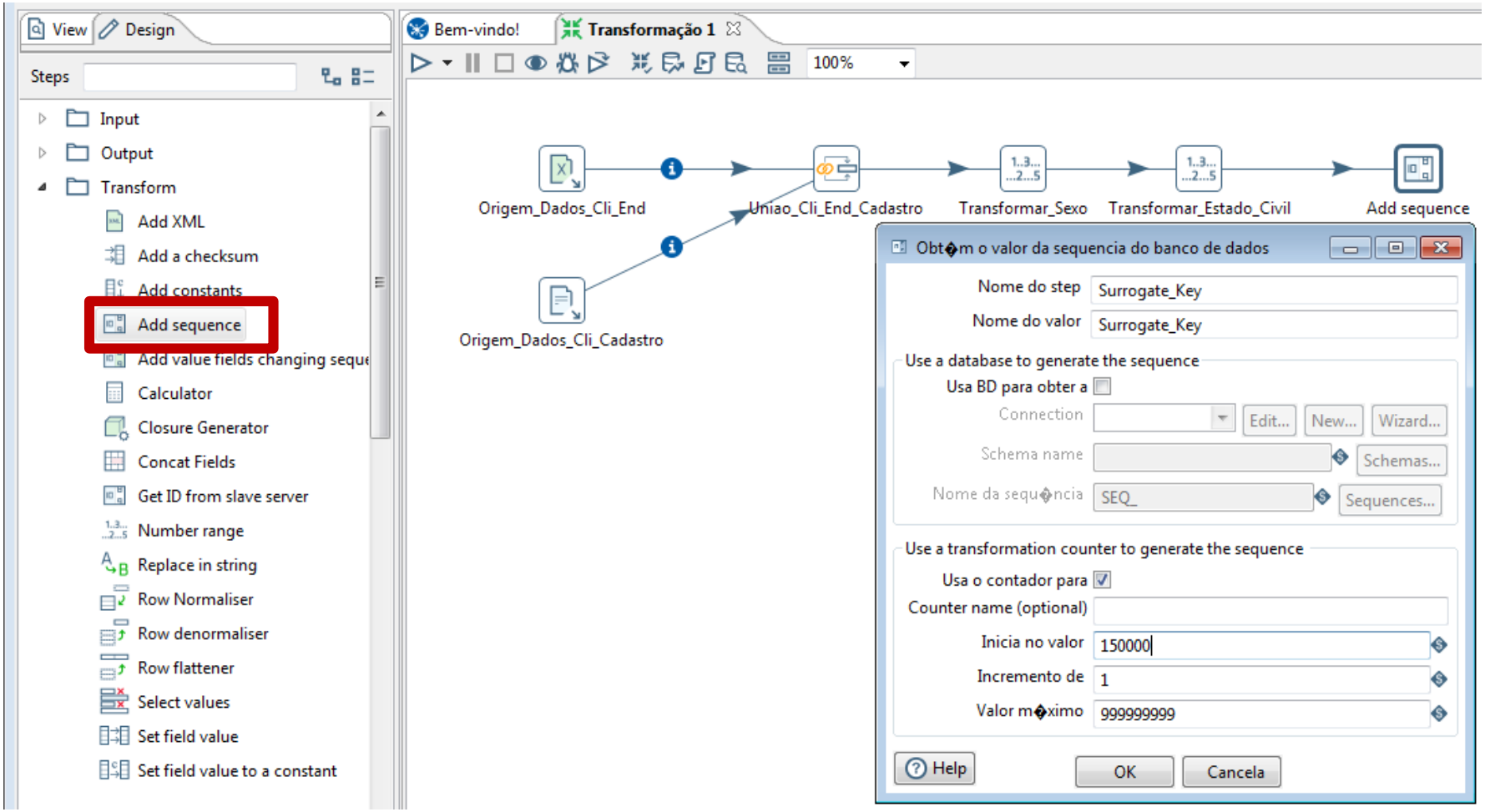
5º passo: Transformação do Estado Civil do Cliente

A origem de dados contém códigos para identificar o Estado Civil do cliente. No modelo dimensional, para evitar futuros cálculos e prejudicar o desempenho, se faz necessário realizar essa importante transformação.

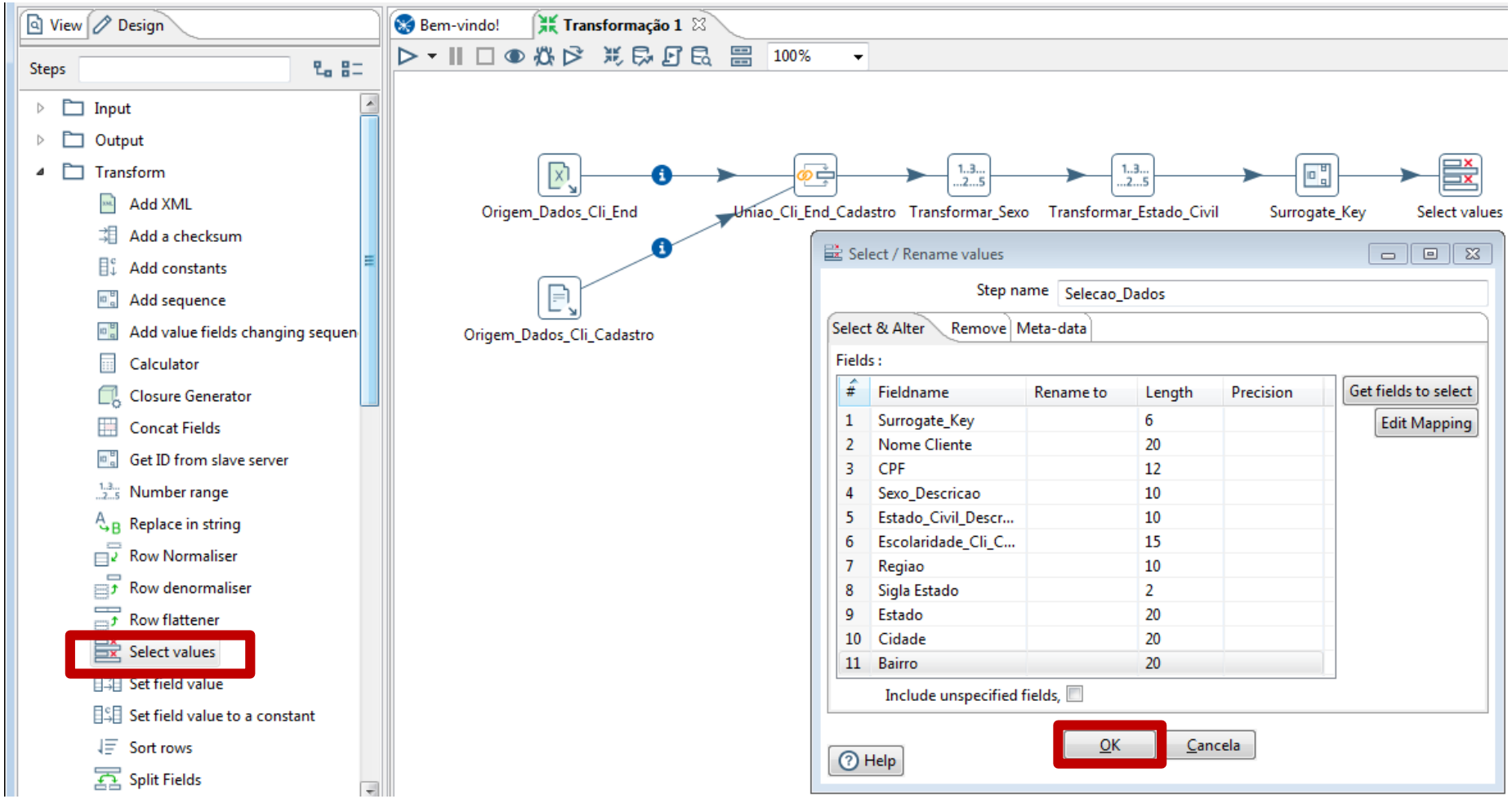
Valor original	Novo valor
1	Solteiro
2	Casado
3	Divorciado
Outro valor	Estado Civil Inválido



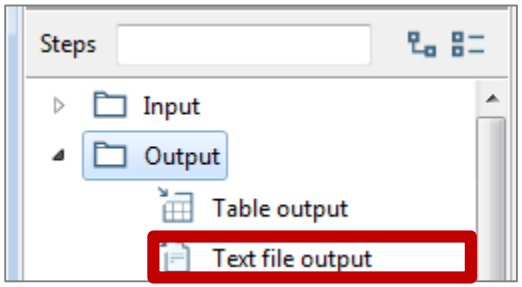
6º passo: Criar chave única (*Surrogate Key*)

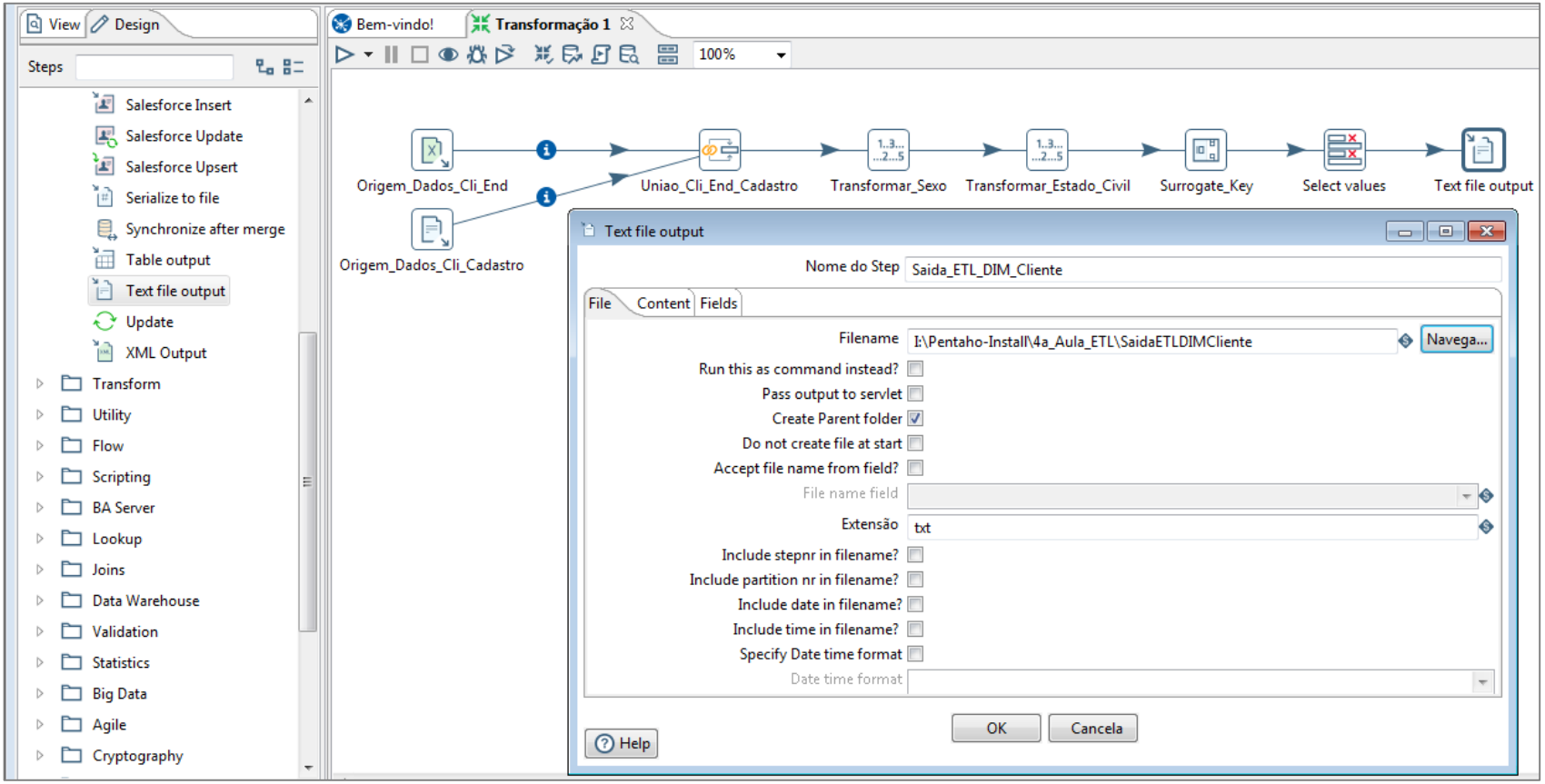


7º passo: Seleção dos dados a serem carregados

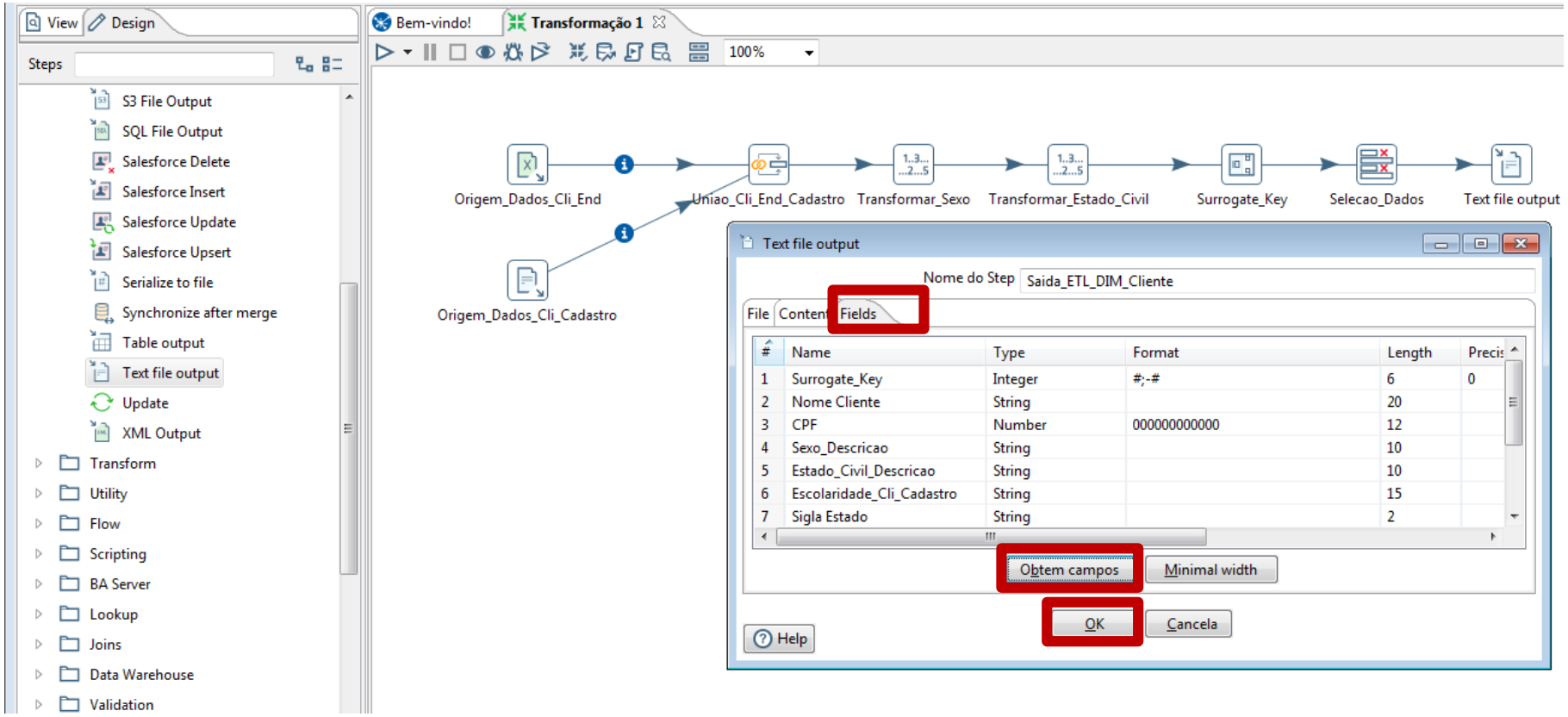


8º passo: Carga dos dados (load)

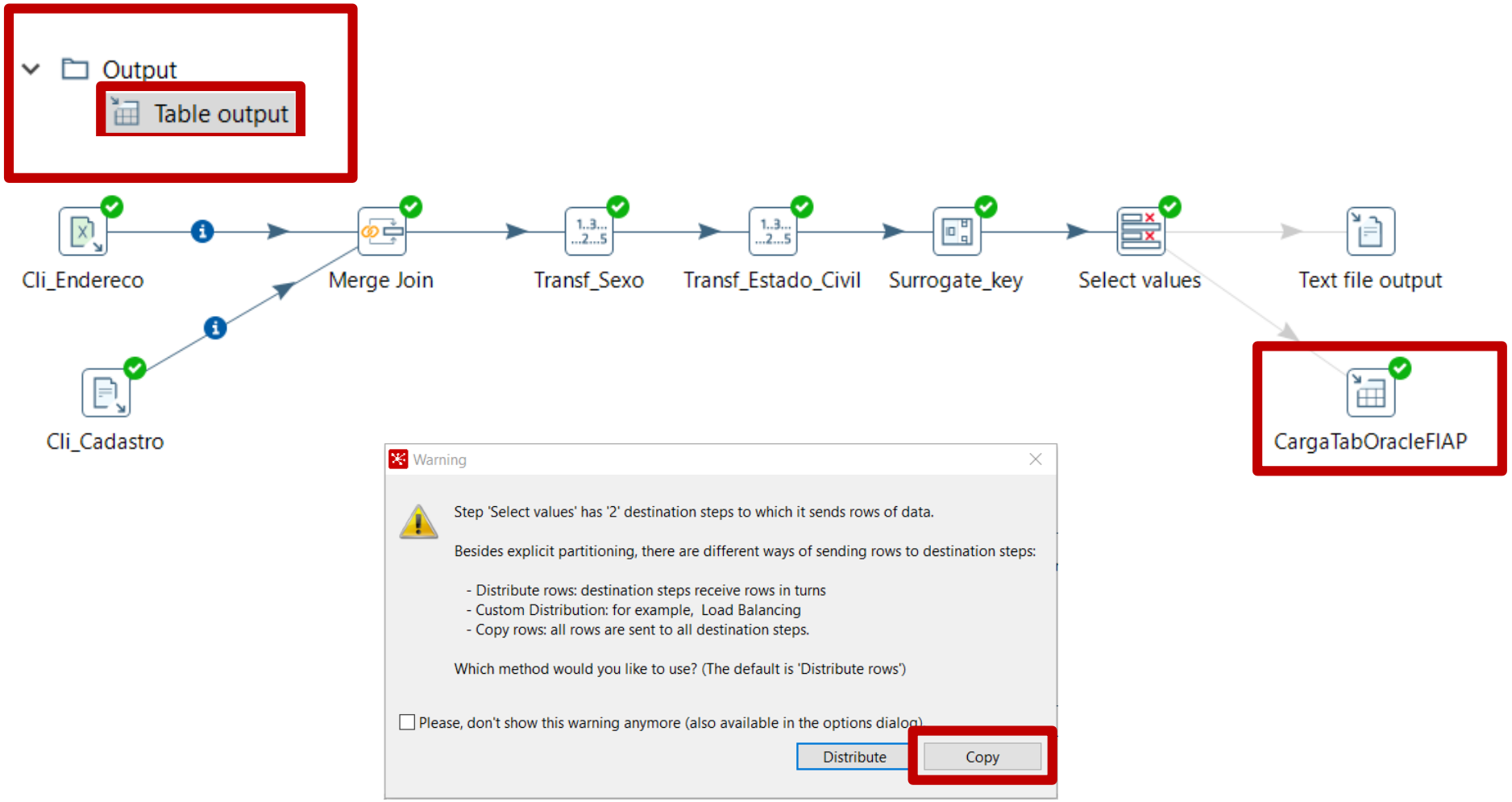




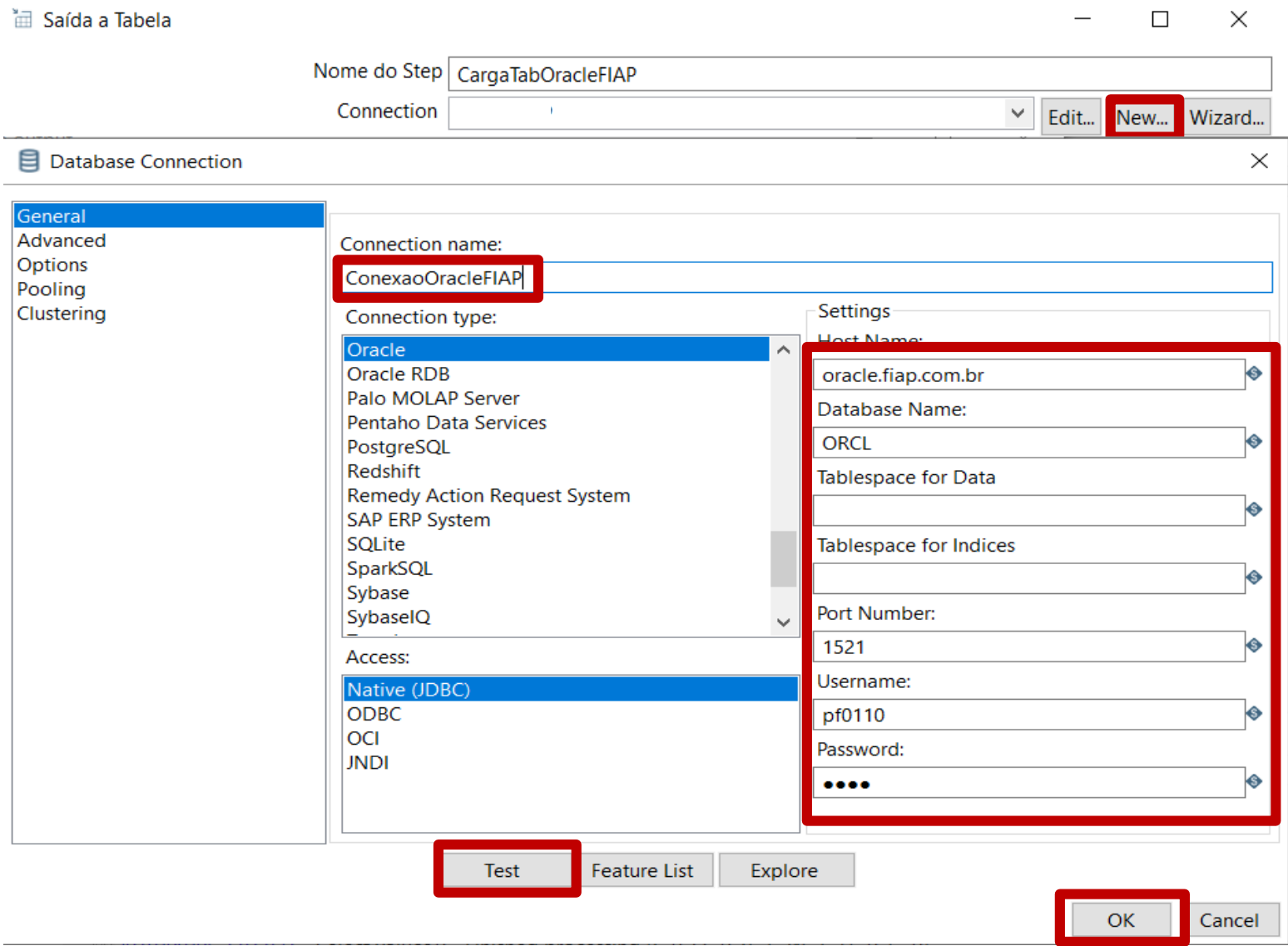
8º passo: Carga dos dados (load)



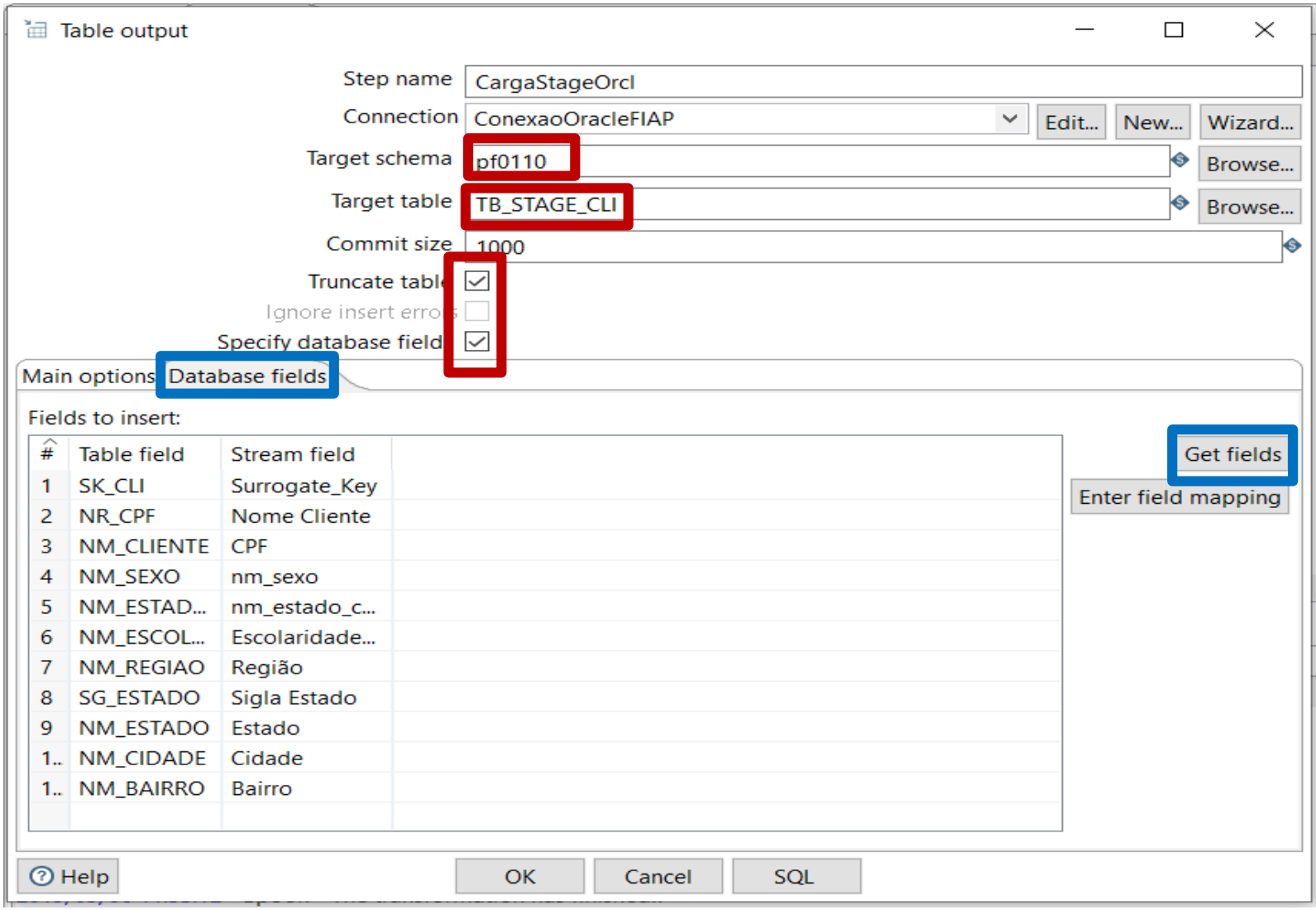
9º passo: Saída dos dados em tabela (load)



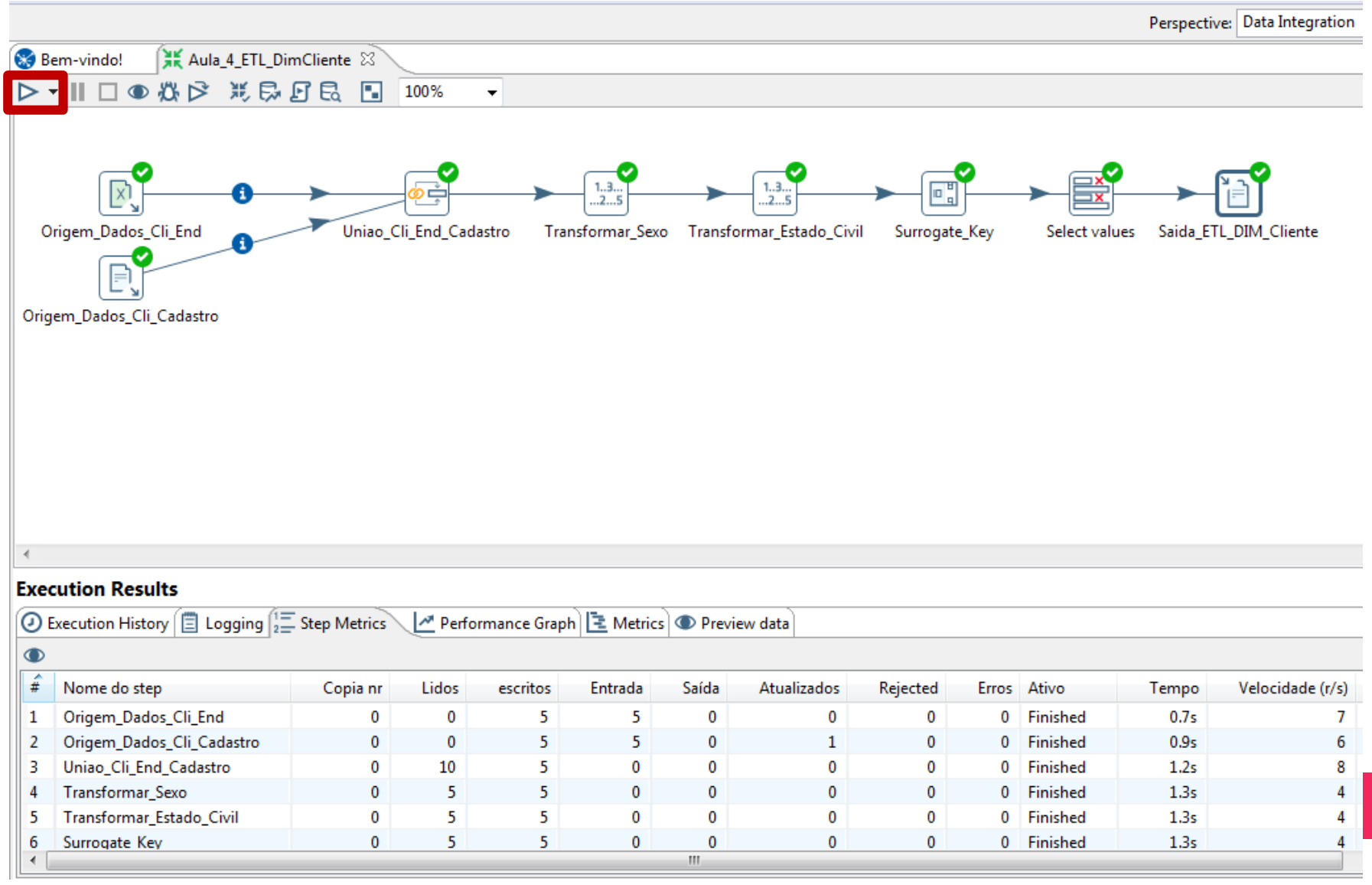
9º passo: Saída dos dados em tabela (load)



9º passo: Saída dos dados em tabela (load)



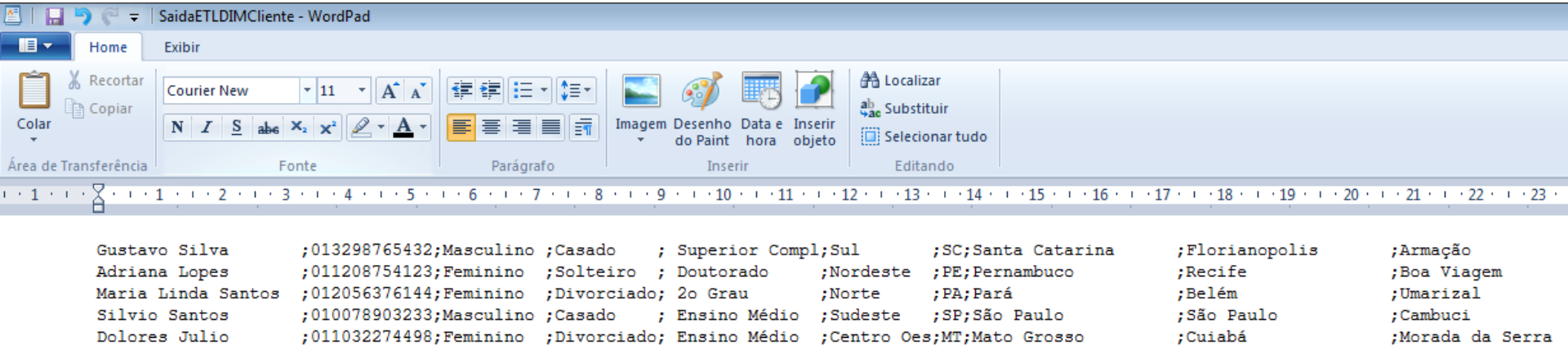
9º passo: Carga dos dados (load)



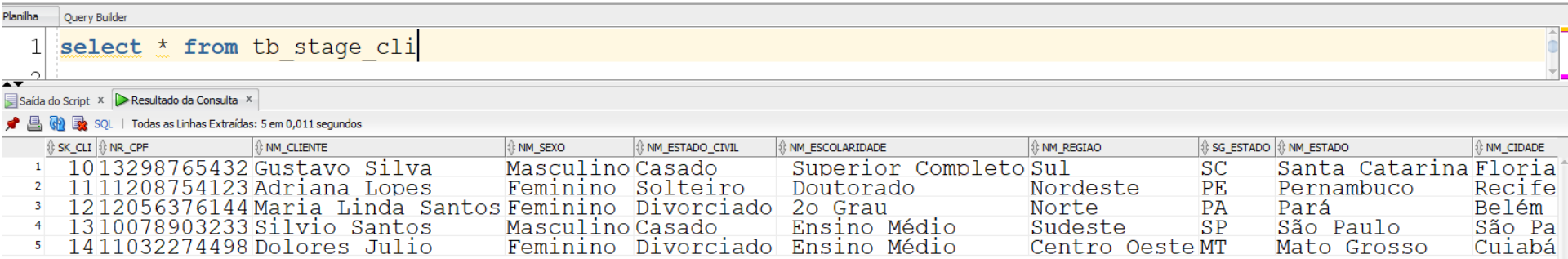
#	Nome do step	Copia nr	Lidos	escritos	Entrada	Saída	Atualizados	Rejected	Erros	Ativo	Tempo	Velocidade (r/s)
1	Origem_Dados_Cli_End	0	0	5	5	0	0	0	0	Finished	0.7s	7
2	Origem_Dados_Cli_Cadastro	0	0	5	5	0	1	0	0	Finished	0.9s	6
3	Uniao_Cli_End_Cadastro	0	10	5	0	0	0	0	0	Finished	1.2s	8
4	Transformar_Sexo	0	5	5	0	0	0	0	0	Finished	1.3s	4
5	Transformar_Estado_Civil	0	5	5	0	0	0	0	0	Finished	1.3s	4
6	Surrogate_Key	0	5	5	0	0	0	0	0	Finished	1.3s	4

Última etapa: Consultando o resultado

Arquivo TXT



Saída em um típico SGBDR Oracle



	SK_CLI	NR_CPF	NM_CLIENTE	NM_SEXO	NM_ESTADO_CIVIL	NM_ESCOLARIDADE	NM_REGIAO	SG_ESTADO	NM_ESTADO	NM_CIDADE
1	10	13298765432	Gustavo Silva	Masculino	Casado	Superior Completo	Sul	SC	Santa Catarina	Floria
2	11	11208754123	Adriana Lopes	Feminino	Solteiro	Doutorado	Nordeste	PE	Pernambuco	Recife
3	12	12056376144	Maria Linda Santos	Feminino	Divorciado	2o Grau	Norte	PA	Pará	Belém
4	13	10078903233	Silvio Santos	Masculino	Casado	Ensino Médio	Sudeste	SP	São Paulo	São Pa
5	14	11032274498	Dolores Julio	Feminino	Divorciado	Ensino Médio	Centro Oeste	MT	Mato Grosso	Cuiabá