

FIAP GRADUAÇÃO

# TECNOLOGIA EM DESENVOLVIMENTO DE SISTEMAS

*Enterprise Analytics e Data Warehousing*

ETL: Exercício Prático

PROFA. FERNANDA P. CAETANO [proffernanda.caetano@fiap.com.br](mailto:proffernanda.caetano@fiap.com.br)

PROF. SALVIO PADLIPSKAS [salvio@fiap.com.br](mailto:salvio@fiap.com.br)



# 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).



Visualização da estrutura de tabelas no Pentaho Data Integrator:

- TB\_STAGE\_CLI
  - SK\_CLI
  - NR\_CPF
  - NM\_CLIENTE
  - NM\_SEXO
  - NM\_ESTADO\_CIVIL
  - NM\_ESCOLARIDADE
  - NM\_REGIAO
  - SG\_ESTADO
  - NM\_ESTADO
  - NM\_CIDADE
  - NM\_BAIRRO

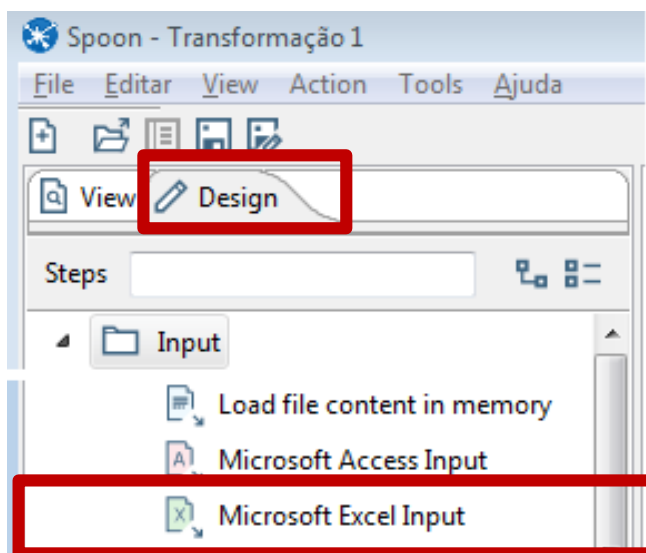
Visualização dos dados extraídos e transformados no WordPad:

Gustavo Silva	;013298765432;	Masculino	;Casado	; Superior Compl;	Sul	;SC;Santa Catarina	;Florianopolis	;Armação
Adriana Lopes	;011208754123;	Feminino	;Solteiro	; Doutorado	;Nordeste	;PE;Pernambuco	;Recife	;Boa Viagem
Maria Linda Santos	;012056376144;	Feminino	;Divorciado;	2o Grau	;Norte	;PA;Pará	;Belém	;Umarizal
Silvio Santos	;010078903233;	Masculino	;Casado	; Ensino Médio	;Sudeste	;SP;São Paulo	;São Paulo	;Cambuci
Dolores Julio	;011032274498;	Feminino	;Divorciado;	Ensino Médio	;Centro Oes;	MT;Mato Grosso	;Cuiabá	;Morada da Serra

## 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



Microsoft Excel input

Nome do Step: Origem\_Cli\_End  
Add Field(s)

!Files !Sheets Content Error Handling **Fields** Additional output fields

Spread sheet type (engine): Excel 97-2003 XLS (JXL)

File or directory: I:\Pentaho-Install\4a\_Aula\_ETL\Origem\_Dados\_Cli\_End.xlt **Add** Navega...

Regular Expression

Exclude Regular Expression

Selected files:

#	File/Directory	Wildcard (RegExp)	Exclude wildcard	Required	Include subfolders
1					

Delete

Accept filenames from previous steps

Accept filenames from previous step ☐

Step to read filenames from

Field in the input to use as filename

Show filename(s)...

OK Preview rows Cancela

Help

Botão Add

Nome do Step: Microsoft Excel Input  
Add sheet(s)

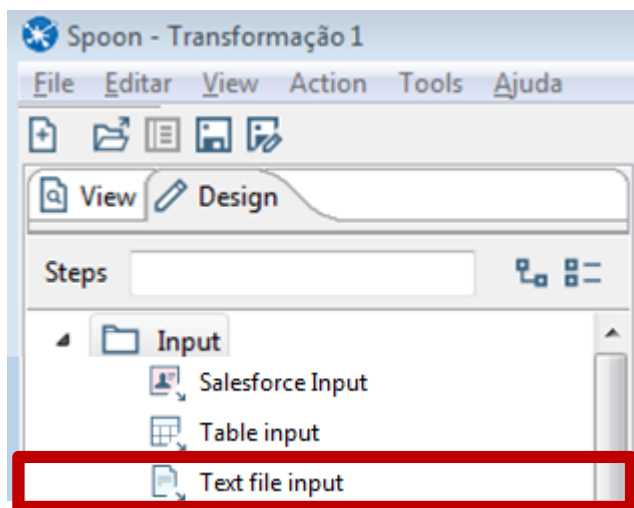
Files !Sheets Content Error Handling **Fields** Additional output fields

#	Name	Type	Length	Precision	Trim type	Repeat	Format	Currency	Decimal	Grouping
1	CPF	Number			none	N				
2	Nome Cliente	String			none	N				
3	Região	String			none	N				
4	Estado	String			none	N				
5	Cidade	String			none	N				
6	Bairro	String			none	N				
7	Data Cadastro	Date			none	N				

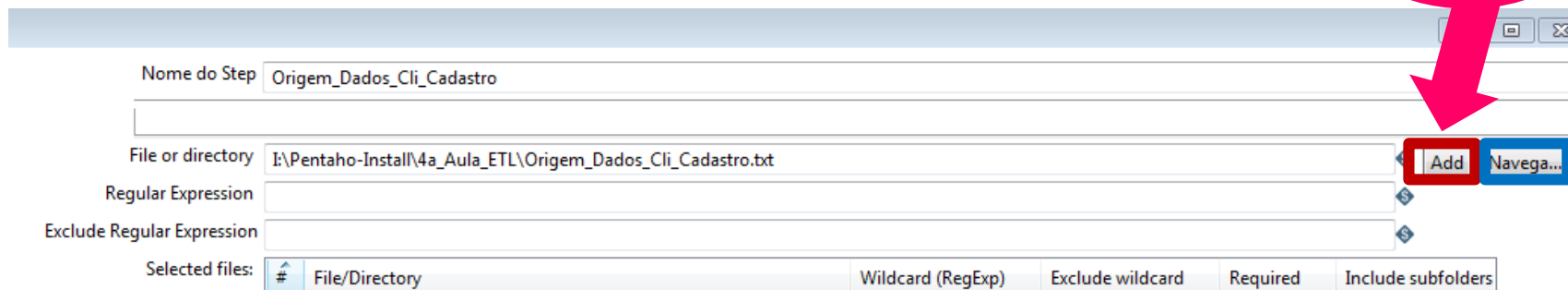
Get fields from header row...

## 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



Pressione o  
botão Add





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

Letura de arquivo plano

Nome do Step: Origem\_Dados\_Cli\_Cadastro

File Content Error Handling Filters **Fields** Additional output fields

#	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

Obtem campos

Preview rows

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

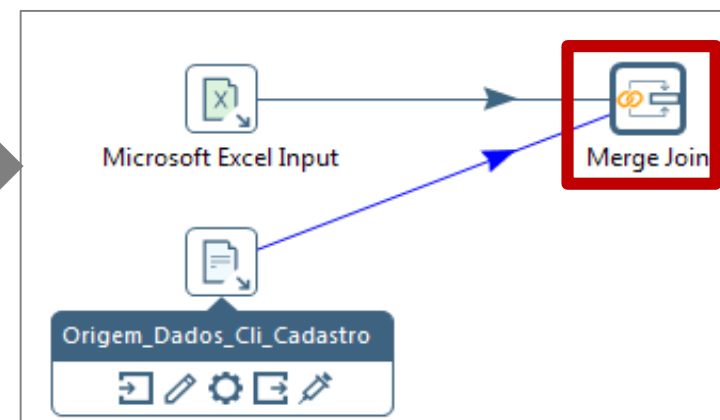
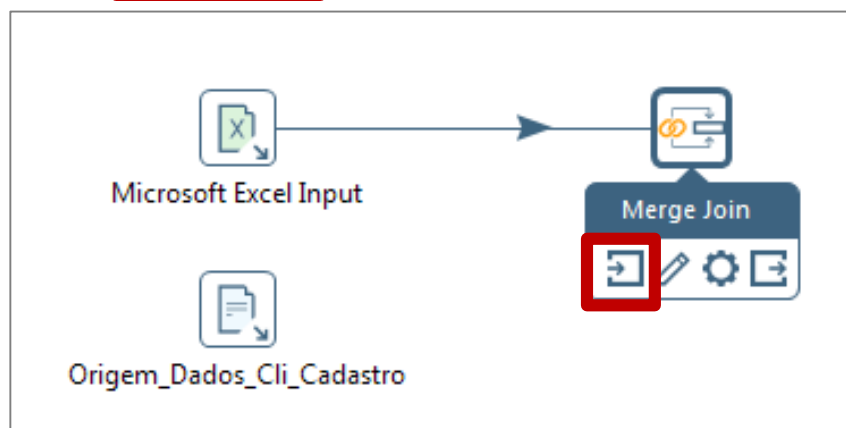
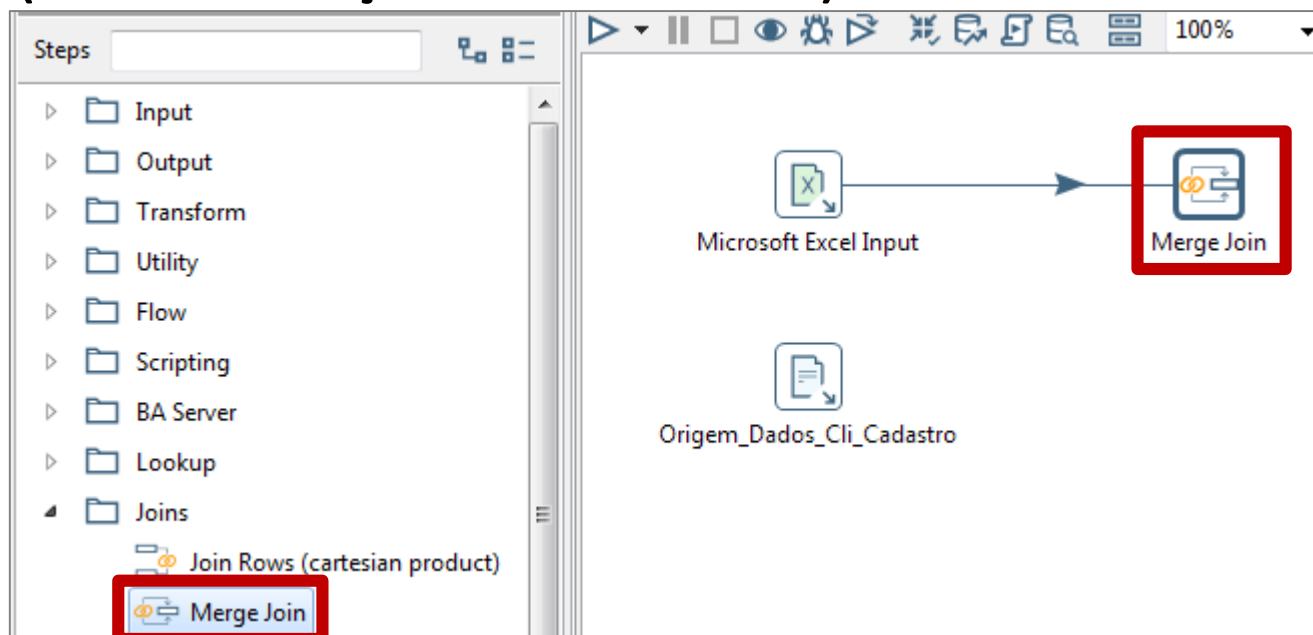


Origem\_Dados\_Cli\_End



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)

Merge Join

Step name: Uniao\_Cli\_End\_Cadastro

First Step: Origem\_Dados\_Cli\_End

Second Step: Origem\_Dados\_Cli\_Cadastro

Join Type: INNER

Keys for 1st step:

#	Key field
1	CPF

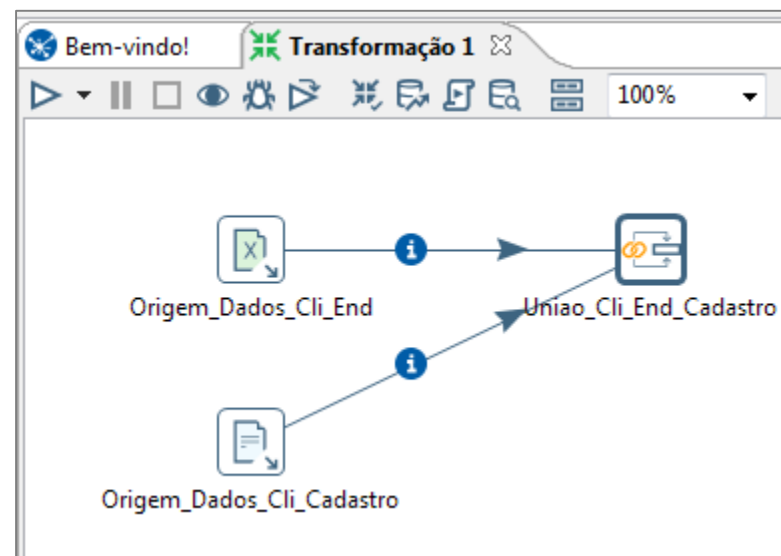
Keys for 2nd step:

#	Key field
1	CPF Cli Cadastro

Get key fields

Get key fields

Help OK Cancela



## 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

The screenshot shows the Alteryx Designer interface. On the left, the 'Steps' pane lists various transformation tools. The 'Number range' tool is highlighted with a red box. The main workspace displays a workflow diagram with the following components:

- Origem\_Dados\_Cli\_End**: A data source icon.
- Origem\_Dados\_Cli\_Cadastro**: A data source icon.
- Uniao\_Cli\_End\_Cadastro**: A join tool connecting the two data sources.
- Number range**: A tool that applies the transformation rules to the joined data.

The 'Number ranges' dialog box is open, showing the configuration for the 'Transformar\_Sexo' step:

- Step name:** Transformar\_Sexo
- Input field:** Sexo Cli Cadastro
- Output field:** Sexo\_Descricao
- Default value(if no match):** Sexo Invalido
- Ranges (min <= x < m):**

#	Lower Bound	Upper Bound	Value
1	1	2	Masculino
2	2	3	Feminino

## 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

The screenshot shows the Alteryx Designer interface with a workflow named 'Transformação 1'. The workflow includes the following steps: 'Origem\_Dados\_Cli\_End', 'Uniao\_Cli\_End\_Cadastro', 'Transformar\_Sexo', and 'Number range'. A 'Number ranges' dialog box is open, showing the configuration for the 'Transformar\_Estado\_Civil' step. The dialog box has the following fields: 'Step name: Transformar\_Estado\_Civil', 'Input field: Estado Civil Cli Cadastro', 'Output field: Estado\_Civil\_Descricao', and 'Default value(if no): Estado Civil Inválido'. The 'Ranges (min <= x < max)' table is also displayed:

#	Lower Bound	Upper Bound	Value
1	1	2	Solteiro
2	2	3	Casado
3	3	4	Divorciado

The 'Number range' tool in the 'Transform' section of the left sidebar is highlighted with a red box.

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

The screenshot displays the Alteryx Designer interface during the configuration of a data transformation workflow. On the left, the 'Steps' pane shows the 'Add sequence' step highlighted with a red rectangle. The main workspace shows a workflow diagram with the following steps: 'Origem\_Dados\_Cli\_End', 'Uniao\_Cli\_End\_Cadastro', 'Transformar\_Sexo', 'Transformar\_Estado\_Civil', and 'Add sequence'. The 'Add sequence' step is currently selected, and its configuration dialog is open.

The configuration dialog for the 'Add sequence' step is titled 'Obter o valor da sequência do banco de dados'. It contains the following fields and options:

- Nome do step:** Surrogate\_Key
- Nome do valor:** Surrogate\_Key
- Use a database to generate the sequence:**
  - ☐ Usa BD para obter a
  - Connection: [Dropdown menu]
  - Schema name: [Text field]
  - Nome da sequência: SEQ\_
- Use a transformation counter to generate the sequence:**
  - ☒ Usa o contador para
  - Counter name (optional): [Text field]
  - Inicia no valor: 150000
  - Incremento de: 1
  - Valor máximo: 999999999

At the bottom of the dialog are buttons for 'Help', 'OK', and 'Cancela'.

## 7º passo: Seleção dos dados a serem carregados

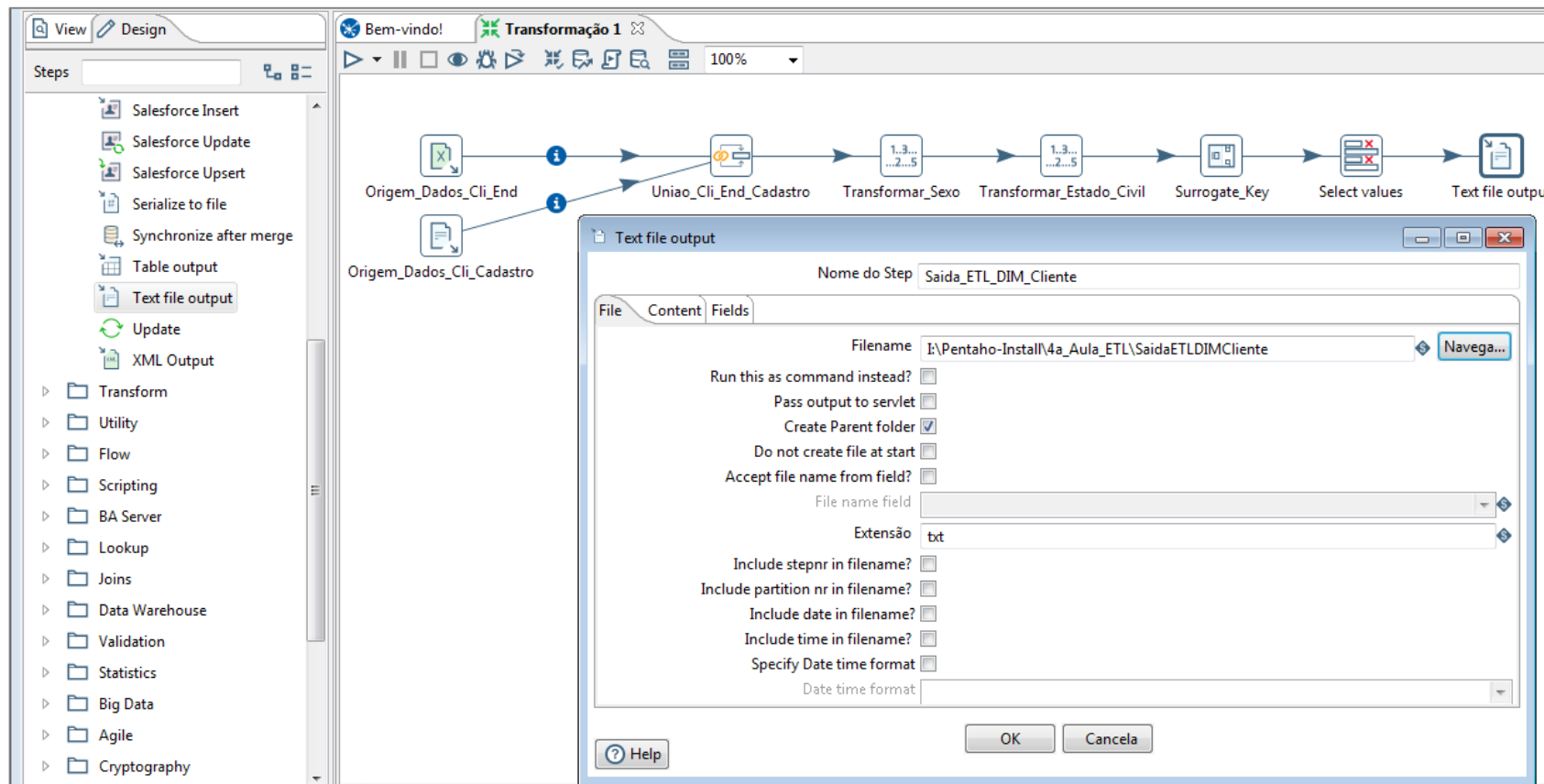
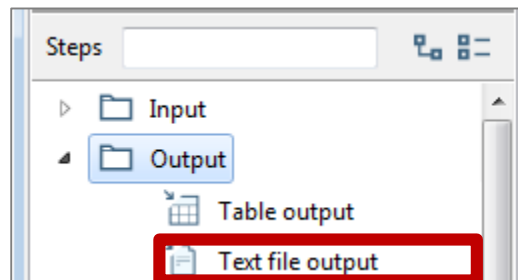
The screenshot displays the Apache NiFi 'Transform' step configuration. The left sidebar lists various transformation options, with 'Select values' highlighted. The main canvas shows a data flow diagram with the following steps: 'Origem\_Dados\_Cli\_End', 'Uniao\_Cli\_End\_Cadastro', 'Transformar\_Sexo', 'Transformar\_Estado\_Civil', 'Surrogate\_Key', and 'Select values'. A 'Select / Rename values' dialog box is open, showing a table of fields to be selected.

**Select / Rename values dialog box:**

#	Fieldname	Rename to	Length	Precision
1	Surrogate_Key		6	
2	Nome Cliente		20	
3	CPF		12	
4	Sexo_Descricao		10	
5	Estado_Civil_Descr...		10	
6	Escolaridade_Cli_C...		15	
7	Regiao		10	
8	Sigla Estado		2	
9	Estado		20	
10	Cidade		20	
11	Bairro		20	

The dialog box also includes a 'Get fields to select' button, an 'Edit Mapping' button, and an 'Include unspecified fields' checkbox. The 'OK' button is highlighted.

## 8º passo: Carga dos dados (load)





## 8º passo: Carga dos dados (load)

The screenshot shows a data transformation tool interface. On the left is a 'Steps' panel with a list of available components: S3 File Output, SQL File Output, Salesforce Delete, Salesforce Insert, Salesforce Update, Salesforce Upsert, Serialize to file, Synchronize after merge, Table output, Text file output, Update, and XML Output. Below this is a folder tree containing Transform, Utility, Flow, Scripting, BA Server, Lookup, Joins, Data Warehouse, and Validation.

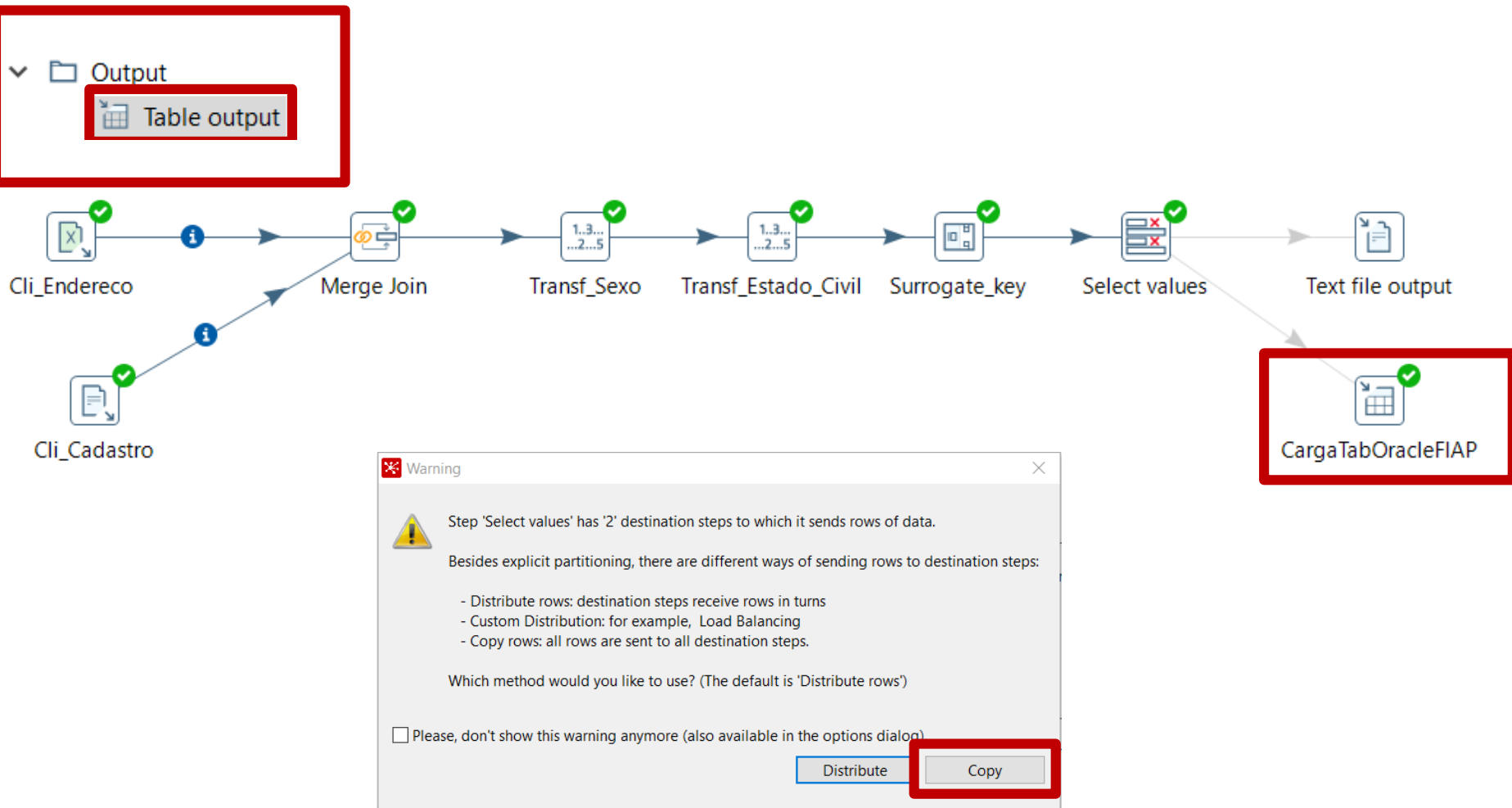
The main workspace displays a workflow diagram titled 'Transformação 1'. The workflow consists of the following steps: 'Origem\_Dados\_Cli\_End' (Excel icon), 'Uniao\_Cli\_End\_Cadastro' (Join icon), 'Transformar\_Sexo' (Filter icon), 'Transformar\_Estado\_Civil' (Filter icon), 'Surrogate\_Key' (Table icon), 'Selecao\_Dados' (Filter icon), and 'Text file output' (File icon). Arrows indicate the flow from left to right. There are also two input sources: 'Origem\_Dados\_Cli\_Cadastro' (File icon) and 'Origem\_Dados\_Cli\_End' (Excel icon), both pointing into the 'Uniao\_Cli\_End\_Cadastro' step.

Overlaid on the workflow is a 'Text file output' dialog box. The 'Nome do Step' field is set to 'Saida\_ETL\_DIM\_Cliente'. The 'Fields' tab is selected, showing a table of fields to be output:

#	Name	Type	Format	Length	Precis
1	Surrogate_Key	Integer	#;-#	6	0
2	Nome Cliente	String		20	
3	CPF	Number	000000000000	12	
4	Sexo_Descricao	String		10	
5	Estado_Civil_Descricao	String		10	
6	Escolaridade_Cli_Cadastro	String		15	
7	Sigla Estado	String		2	

At the bottom of the dialog, there are buttons for 'Obtem campos' (highlighted with a red box), 'Minimal width', 'OK' (highlighted with a red box), and 'Cancela'. A 'Help' button is also present in the bottom left corner.

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



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

Saída a Tabela

Nome do Step: CargaTabOracleFIAP

Connection: [v] Edit... **New...** Wizard...

Database Connection

General  
Advanced  
Options  
Pooling  
Clustering

Connection name:  
**ConexaoOracleFIAP**

Connection type:  
Oracle  
Oracle RDB  
Palo MOLAP Server  
Pentaho Data Services  
PostgreSQL  
Redshift  
Remedy Action Request System  
SAP ERP System  
SQLite  
SparkSQL  
Sybase  
SybaseIQ

Access:  
**Native (JDBC)**  
ODBC  
OCI  
JNDI

Settings

Host Name:  
oracle.fiap.com.br

Database Name:  
ORCL

Tablespace for Data  
[ ]

Tablespace for Indices  
[ ]

Port Number:  
1521

Username:  
pf0110

Password:  
[ ]

Test Feature List Explore

**OK** Cancel

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

Table output

Step name: CargaStageOrcl

Connection: ConexaoOracleFIAP

Target schema: pf0110

Target table: TB\_STAGE\_CLI

Commit size: 1000

Truncate table: ☒

Ignore insert errors: ☐

Specify database field: ☒

Main options: Database fields

Fields to insert:

#	Table field	Stream field
1	SK_CLI	Surrogate_Key
2	NR_CPF	Nome Cliente
3	NM_CLIENTE	CPF
4	NM_SEXO	nm_sexo
5	NM_ESTAD...	nm_estado_c...
6	NM_ESCOL...	Escolaridade...
7	NM_REGIAO	Região
8	SG_ESTADO	Sigla Estado
9	NM_ESTADO	Estado
1..	NM_CIDADE	Cidade
1..	NM_BAIRRO	Bairro

Get fields

Enter field mapping

Help OK Cancel SQL

## 9º passo: Carga dos dados (load)

Perspective: Data Integration

Bem-vindo! Aula\_4\_ETL\_DimCliente

100%

```

graph LR
    Origem_Dados_Cli_End[Origem_Dados_Cli_End] --> Uniao_Cli_End_Cadastro[Uniao_Cli_End_Cadastro]
    Origem_Dados_Cli_Cadastro[Origem_Dados_Cli_Cadastro] --> Uniao_Cli_End_Cadastro
    Uniao_Cli_End_Cadastro --> Transformar_Sexo[Transformar_Sexo]
    Transformar_Sexo --> Transformar_Estado_Civil[Transformar_Estado_Civil]
    Transformar_Estado_Civil --> Surrogate_Key[Surrogate_Key]
    Surrogate_Key --> Select_values[Select values]
    Select_values --> Saida_ETL_DIM_Cliente[Saida_ETL_DIM_Cliente]
  
```

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

#	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

SaidaETLDimCliente - WordPad

Home Exibir

Recortar Copiar Colar

Courier New 11

N I S abc x x

Parágrafo

Imagem Desenho do Paint Data e hora Inserir objeto

Localizar Substituir Selecionar tudo

Área de Transferência

Fonte

Parágrafo

Inserir

Editando

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

Gustavo Silva	;013298765432;Masculino	;Casado	; Superior Compl;	Sul	;SC;Santa Catarina	;Florianopolis	;Armação
Adriana Lopes	;011208754123;Feminino	;Solteiro	; Doutorado	;Nordeste	;PE;Pernambuco	;Recife	;Boa Viagem
Maria Linda Santos	;012056376144;Feminino	;Divorciado;	2o Grau	;Norte	;PA;Pará	;Belém	;Umarizal
Silvio Santos	;010078903233;Masculino	;Casado	; Ensino Médio	;Sudeste	;SP;São Paulo	;São Paulo	;Cambuci
Dolores Julio	;011032274498;Feminino	;Divorciado;	Ensino Médio	;Centro Oes;	MT;Mato Grosso	;Cuiabá	;Morada da Serra

## Saída em um típico SGBDR Oracle

Planilha Query Builder

1 select \* from tb\_stage\_cli

2

Saída do Script x Resultado da Consulta x

Todas as Linhas Extraídas: 5 em 0,011 segundos

	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	Flória
2	11	1208754123	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

Copyright © 2019 Prof. Salvio Padlipskas e Profa. Fernanda Caetano

Todos direitos reservados. Reprodução ou divulgação total ou parcial deste documento é expressamente proibido sem o consentimento formal, por escrito, do Professor (autor).