



**UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO**

JESSICA PROFETA DA SILVEIRA

Tutorial: noções básicas de ETL com o PDI Pentaho

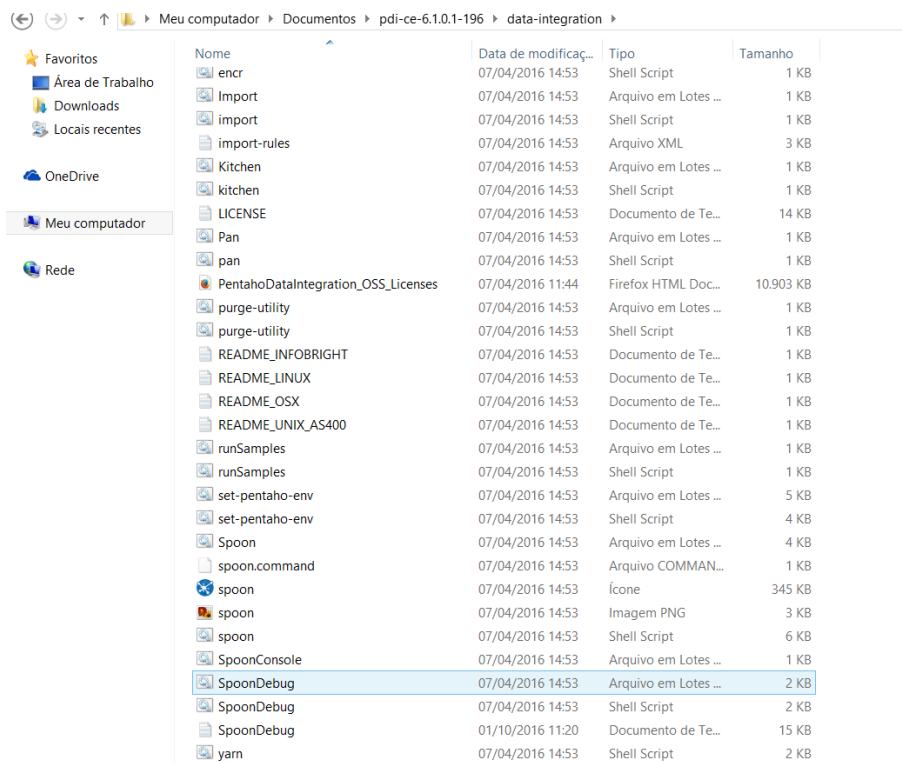
**SÃO CRISTOVÃO
2016**

TUTORIAL

Este tutorial tem como objetivo apresentar um processo de ETL através do PDI Pentaho em um banco de dados relacional. Como exemplo iremos utilizar três arquivos no formato .xls para dar carga no processo de ETL com o PDI Pentaho. Para manipular esses arquivos iremos utilizar processos de transformação para tratar os dados de carga. O resultado deste tutorial serão três arquivos: xml, txt e script MySQL.

Instalando o PDI Pentaho

1. Para baixar PDI Pentaho acesse:
<https://sourceforge.net/projects/pentaho/files/Data%20Integration/6.1/pdi-ce-6.1.0.1-196.zip/download>
2. Após realizar o download descompacte o arquivo em uma pasta de sua preferência.
3. Clicar em Spoon.bat para executar PDI Pentaho.



Nome	Data de modificaç...	Tipo	Tamanho
encr	07/04/2016 14:53	Shell Script	1 KB
Import	07/04/2016 14:53	Arquivo em Lotes ...	1 KB
import	07/04/2016 14:53	Shell Script	1 KB
import-rules	07/04/2016 14:53	Arquivo XML	3 KB
Kitchen	07/04/2016 14:53	Arquivo em Lotes ...	1 KB
kitchen	07/04/2016 14:53	Shell Script	1 KB
LICENSE	07/04/2016 14:53	Documento de Te...	14 KB
Pan	07/04/2016 14:53	Arquivo em Lotes ...	1 KB
pan	07/04/2016 14:53	Shell Script	1 KB
PentahoDataIntegration_OSS_Licenses	07/04/2016 11:44	Firefox HTML Doc...	10.903 KB
purge-utility	07/04/2016 14:53	Arquivo em Lotes ...	1 KB
purge-utility	07/04/2016 14:53	Shell Script	1 KB
README_INFOBRIGHT	07/04/2016 14:53	Documento de Te...	1 KB
README_LINUX	07/04/2016 14:53	Documento de Te...	1 KB
README_OSX	07/04/2016 14:53	Documento de Te...	1 KB
README_UNIX_AS400	07/04/2016 14:53	Documento de Te...	1 KB
runSamples	07/04/2016 14:53	Arquivo em Lotes ...	1 KB
runSamples	07/04/2016 14:53	Shell Script	1 KB
set-pentaho-env	07/04/2016 14:53	Arquivo em Lotes ...	5 KB
set-pentaho-env	07/04/2016 14:53	Shell Script	4 KB
Spoon	07/04/2016 14:53	Arquivo em Lotes ...	4 KB
spoon.command	07/04/2016 14:53	Arquivo COMMAN...	1 KB
spoon	07/04/2016 14:53	Ícone	345 KB
spoon	07/04/2016 14:53	Imagem PNG	3 KB
spoon	07/04/2016 14:53	Shell Script	6 KB
SpoonConsole	07/04/2016 14:53	Arquivo em Lotes ...	1 KB
SpoonDebug	07/04/2016 14:53	Arquivo em Lotes ...	2 KB
SpoonDebug	07/04/2016 14:53	Shell Script	2 KB
SpoonDebug	01/10/2016 11:20	Documento de Te...	15 KB
yarn	07/04/2016 14:53	Shell Script	2 KB

Conceitos Importantes

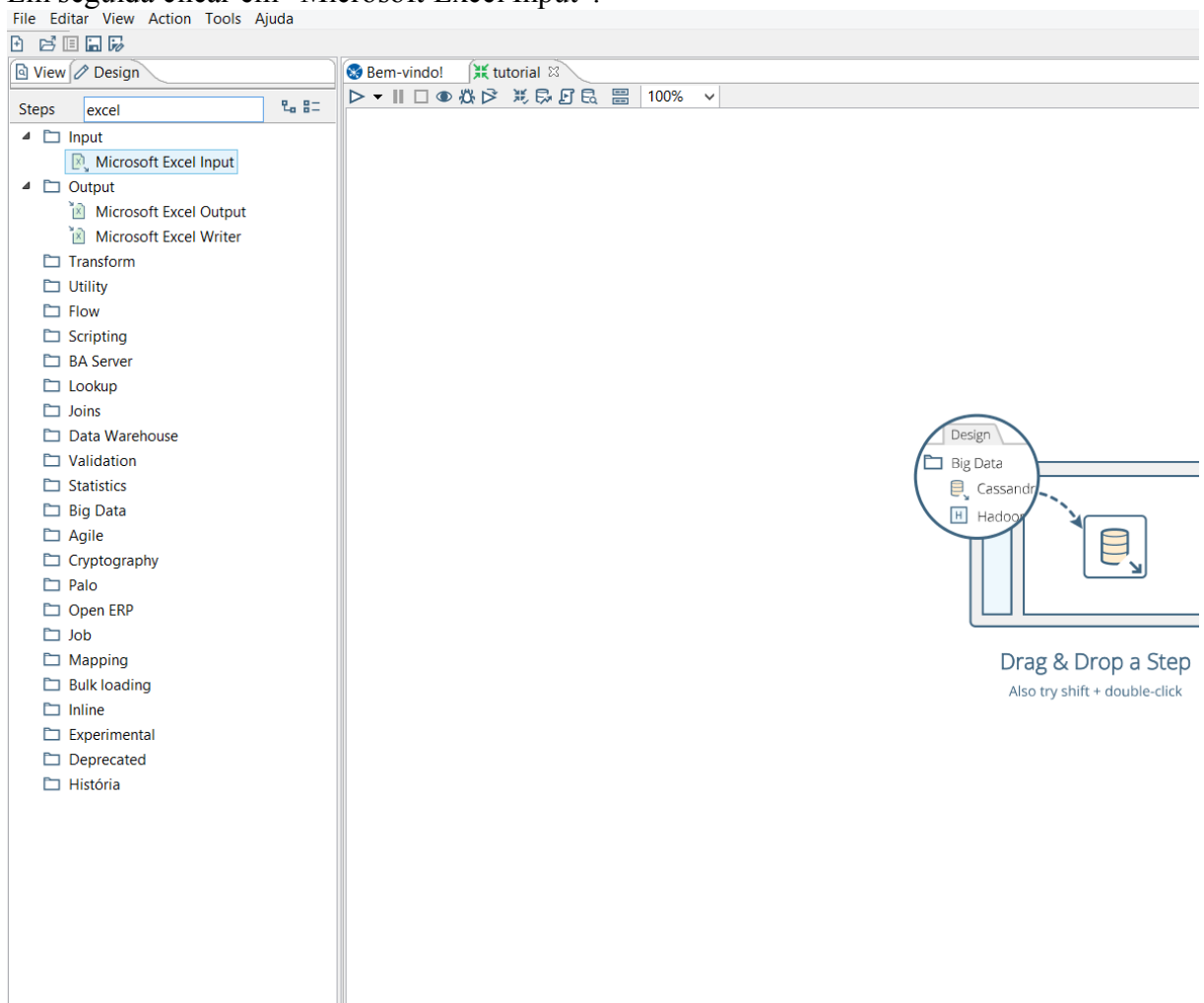
Um **Step** é a unidade mínima em uma transformação. Existe uma grande variedade de Steps que são agrupados em categorias como Input e Output. Cada Step é definido para executar uma tarefa específica como a leitura de um parâmetro ou normalizar um conjunto de dados.

Um **Hop** é uma representação gráfica de um fluxo de dados entre dois Steps, com a origem e o destino. Um Hop só possui uma origem e um destino, mas mais de um Hop pode sair de um Step. Quando isso acontece, a saída de dados pode ser copiada ou distribuída em todo lugar. Da mesma forma, mais de um Hop pode chegar em um Step. Nesse caso, o Step precisa ter a capacidade de fundir as entradas de diferentes Steps, a fim de criar uma saída.

Transformação de três arquivos .xls para um Banco de dados MySQL

Para esse exemplo utilizaremos três planilhas: alimentos1.xls, alimentos2.xls e alimentos3.xls.

1. Com o Spon aberto clicar em File > Novo > Transformação
2. Em “steps” digitar excel.
3. Em seguida clicar em “Microsoft Excel Input”.



4. Edite o step “Microsoft Excel Input Excel” input com os seguintes parâmetros:
 - Aba Files: localize o arquivo .xls com o botão Navegar.

Nome do Step:

[Add Field\(s\)](#)

Files | Sheets | Content | Error Handling | Fields | Additional output fields

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

File or directory: Add Navega...

Regular Expression:

Exclude Regular Expression:

Selected files:

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

Delete Edit

- Em seguida, clique em Add para adicionar o arquivo ao grid. No nosso caso adicionamos a planilha “alimentos1.xls”.

[Add Field\(s\)](#)

Files | Sheets | Content | Error Handling | Fields | Additional output fields

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

File or directory: Add Navega...

Regular Expression:

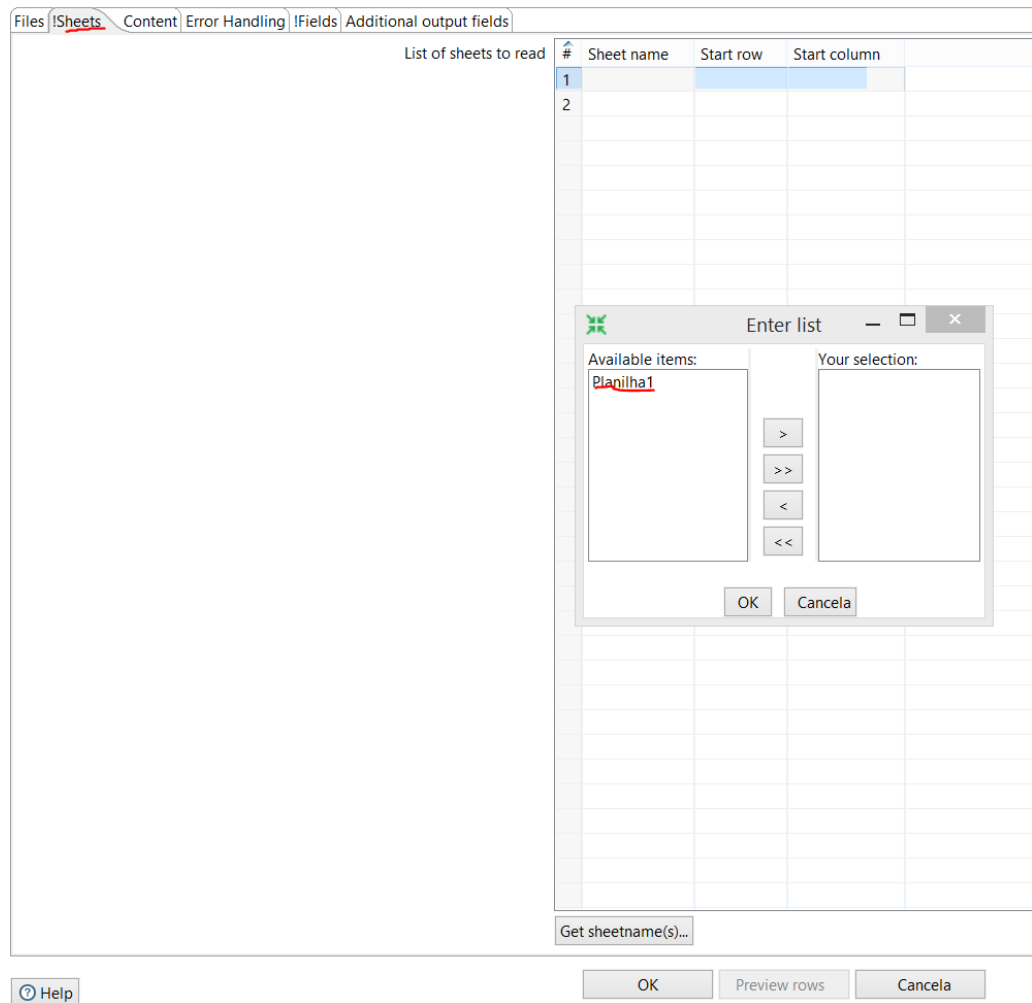
Exclude Regular Expression:

Selected files:

#	File/Directory	Wildcard (RegExp)	Exclude wildcard	Required	Include subfolders
1	C:\Users\jessi_000\Documents\bases\bases\alimentos1.xls			N	N

Delete Edit

- Para ter certeza que o arquivo foi localizado, clique no botão Show filename(s).
- Aba Sheets: Clique no botão Get sheetname e escolha a planilha desejada. Se o nome da planilha não aparecer na lista, reveja os parâmetros da aba Files.



- Aba Content: Certifique-se que o campo Header esteja marcado (vamos precisar dele na próxima aba).

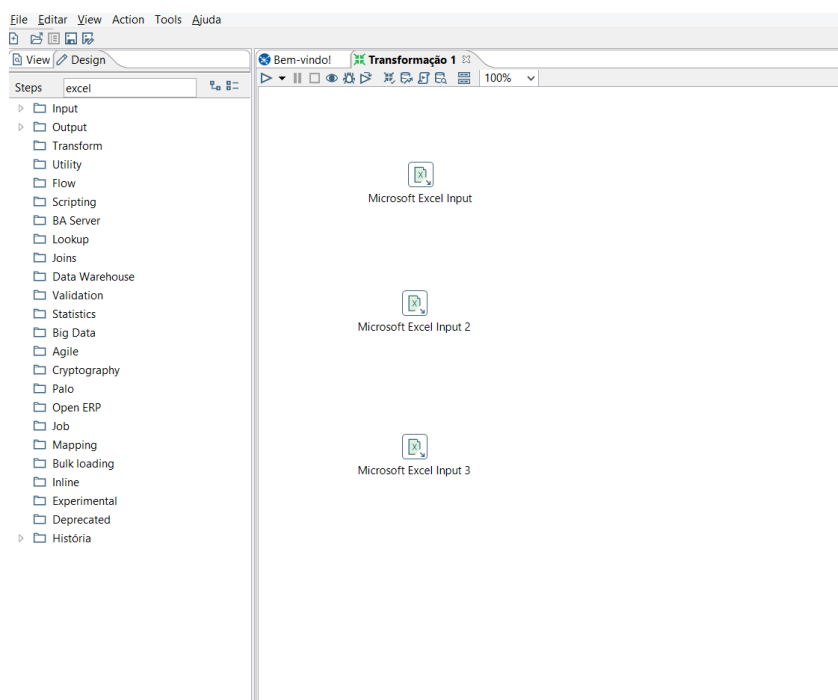
Header	<input checked="" type="checkbox"/>
No empty rows	<input checked="" type="checkbox"/>
Stop on empty row	<input type="checkbox"/>
Limit	<input type="text" value="0"/>
Encoding	<input type="text"/>

Add filenames to result ☒

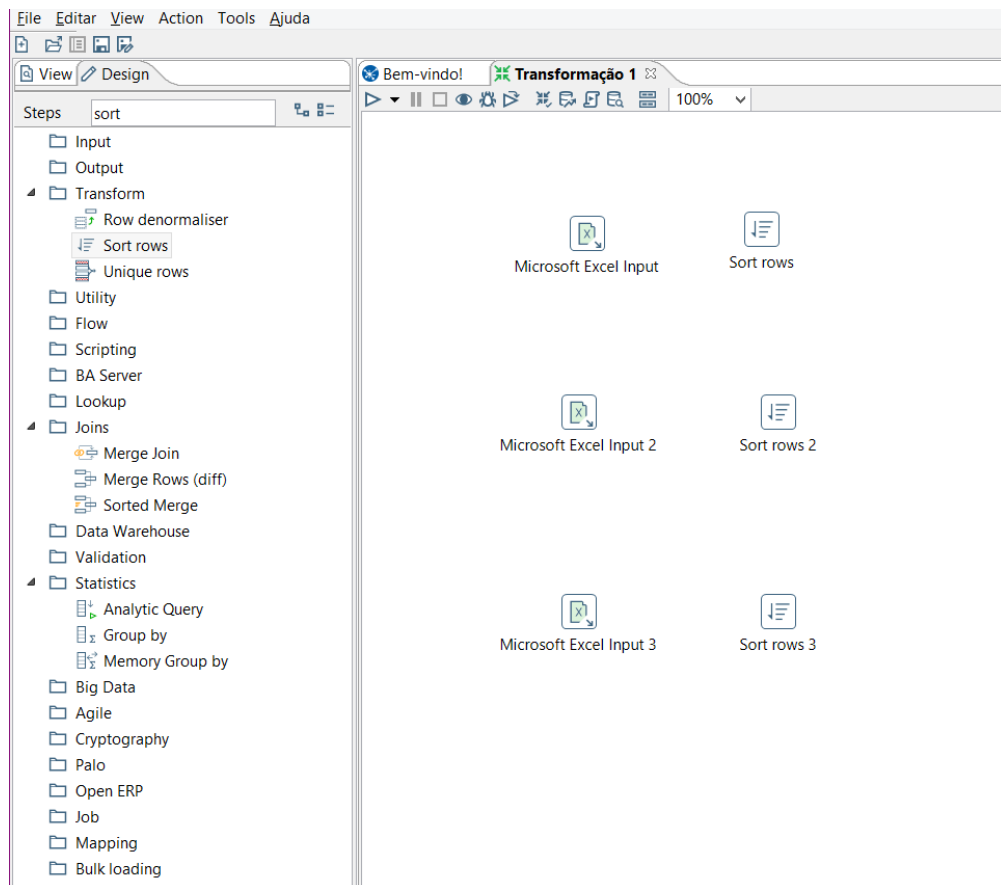
- Aba Fields: Clique no botão Get fields from header now e veja todos os campos disponíveis no arquivo. Dê uma olhada nos dados que serão extraídos do arquivo, clicando no botão Preview rows. Clique Ok e salve a transformação

Files Sheets Content Error Handling Fields Additional output fields										
#	Name	Type	Length	Precision	Trim type	Repeat	Format	Currency	Decimal	Grouping
1	ID_ALIMENTO	Number			none	N				
2	Alimento	String			none	N				

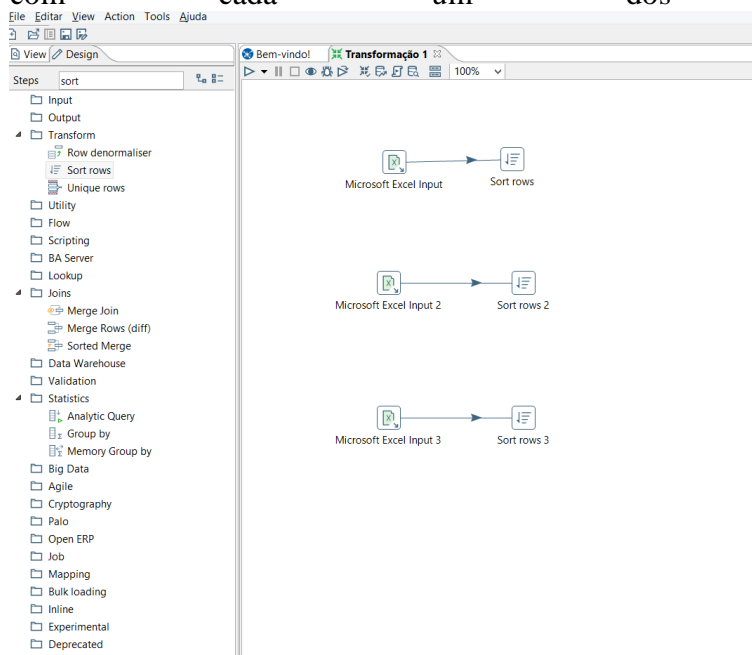
5. Repita os passos: 2, 3 e 4 para adicionar as outras duas planilhas.



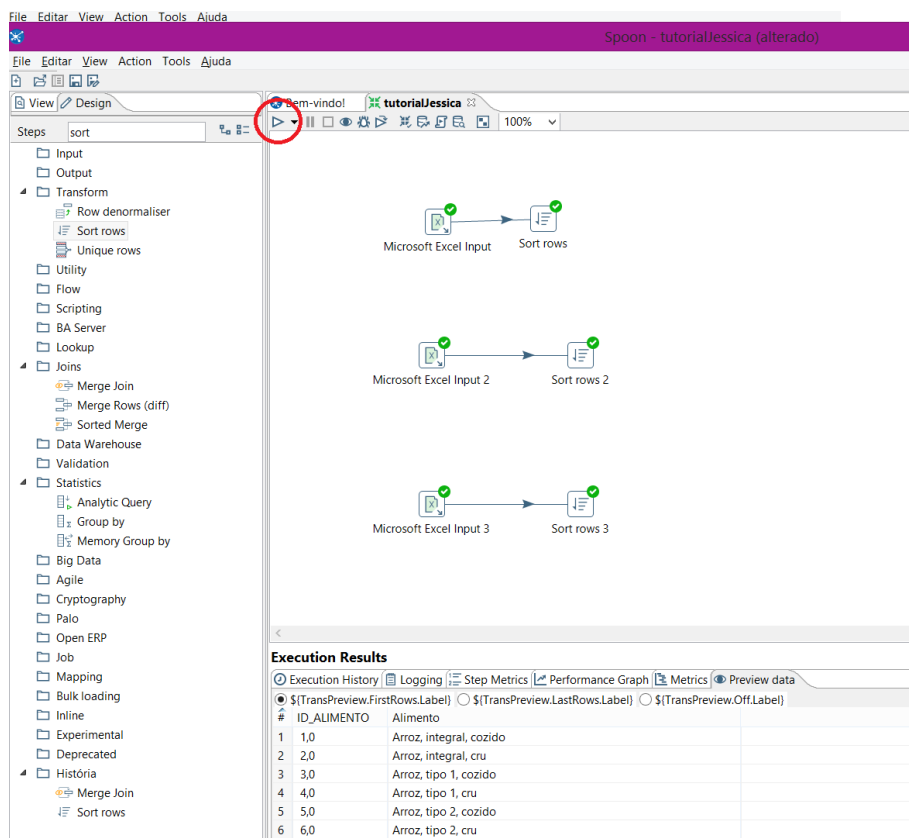
6. Em seguida, adicionar o step “Sort Rows”, step responsável pela ordenação ou classificação de dados que permite ordenar dados através de campos informados em sua lista. O mesmo tem duas formas básicas de ordenação: crescentes ou decrescentes.



7. Fazer hop (clique com botão do meio do mouse e arrastar do step origem a step fim) com cada um dos steps Excel.



8. Na tela de Transformação clicar em Run como indicado na figura abaixo. Se tudo ocorrer bem todos os ícones terão um marcador verde. Além disso, só é possível executar um hop depois de salvar a transformação.



9. Antes de realizar as saídas para xml, txt e banco de dados MySQL precisamos unir as três planilhas. Para isso iremos utilizar o step “Merge Join”. Primeiro com as planilhas 1 e 2 e o resultado desta união iremos unir com a planilha 3. Para isso pesquise o step “Merge Join” e faça um hop com “Sort rows” e “Sort rows 2”.

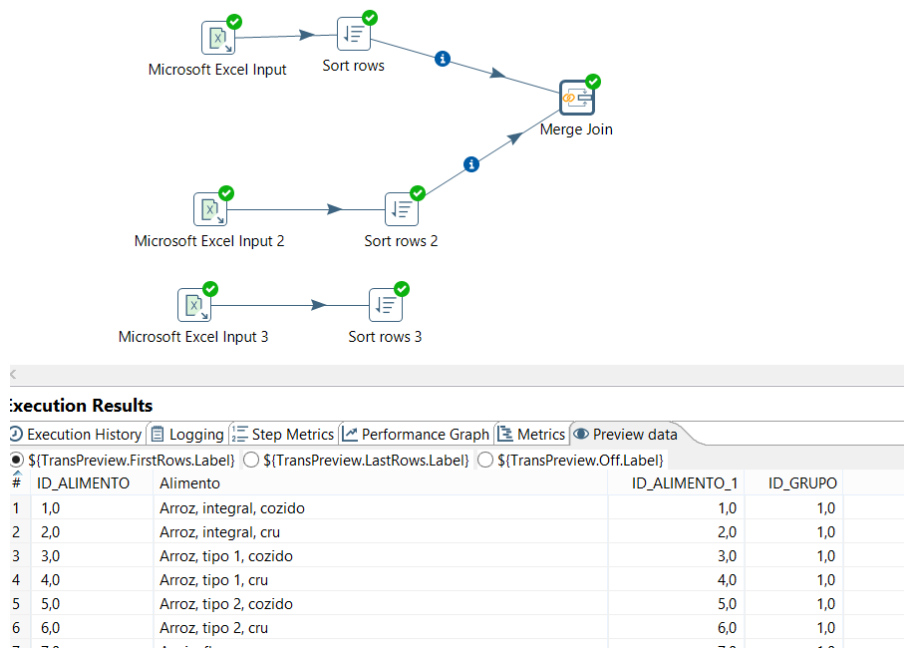
The screenshot shows the Alteryx interface with a workflow named 'merge'. The workflow consists of three parallel paths: 'Microsoft Excel Input' -> 'Sort rows', 'Microsoft Excel Input 2' -> 'Sort rows 2', and 'Microsoft Excel Input 3' -> 'Sort rows 3'. These three paths converge into a 'Merge Join' step. The 'Execution Results' tab is active, displaying a table with 6 rows and 2 columns: ID_ALIMENTO and ID_GRUPO.

#	ID_ALIMENTO	ID_GRUPO
1	1,0	1,0
2	2,0	1,0
3	3,0	1,0
4	4,0	1,0
5	5,0	1,0
6	6,0	1,0

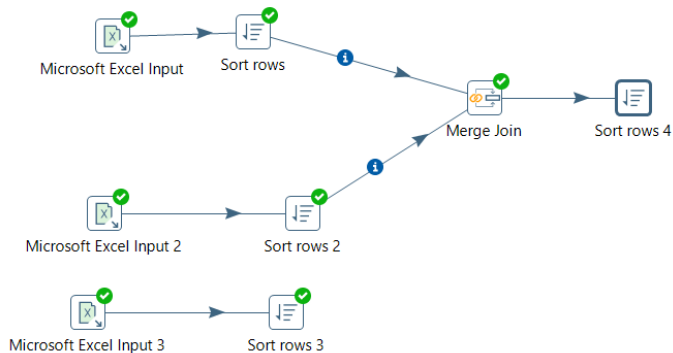
10. Configure o “Merge Join” da seguinte maneira: First Step, escolha o “Sort Rows” e no Second Step escolha o “Sort Rows 2” e em seguida clique em [Get key fields] para cada step e deixe apenas o “ID_ALIMENTO” para cada step e clique em [OK].

The screenshot shows the 'Merge Join' configuration dialog box. The 'Step name' is 'Merge Join'. The 'First Step' is 'Sort rows' and the 'Second Step' is 'Sort rows 2'. The 'Join Type' is 'INNER'. Under 'Keys for 1st step', the key field is 'ID_ALIMENTO'. Under 'Keys for 2nd step', the key field is 'ID_ALIMENTO'. There are 'Get key fields' buttons for both sections and 'Help', 'OK', and 'Cancel' buttons at the bottom.

11. Em seguida clique em Run (ícone mostrado no passo 8). O resultado será a união entre as duas planilhas. Você pode conferir o resultado na aba “Preview Data” como apresentado abaixo:



12. Antes de criarmos outro hop “Merge Join” com “Sort Rows 3”, vamos criar outro step “Sort Rows” e realizar hop com o “Merge Join”.



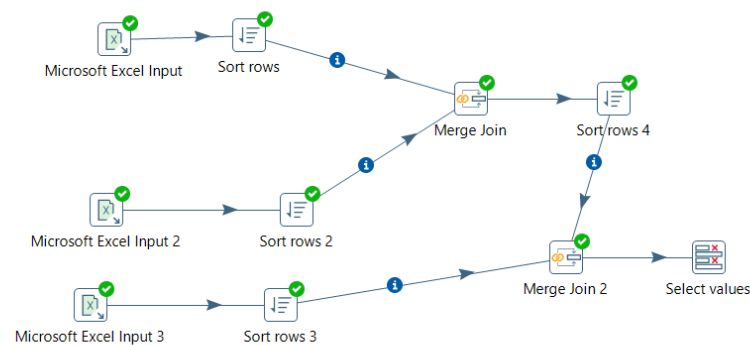
13. Adicione outro “Merge Join” fazendo hop com o “Sort Rows 4” e “Sort Rows 3” de acordo com a configuração abaixo.

Execution Results

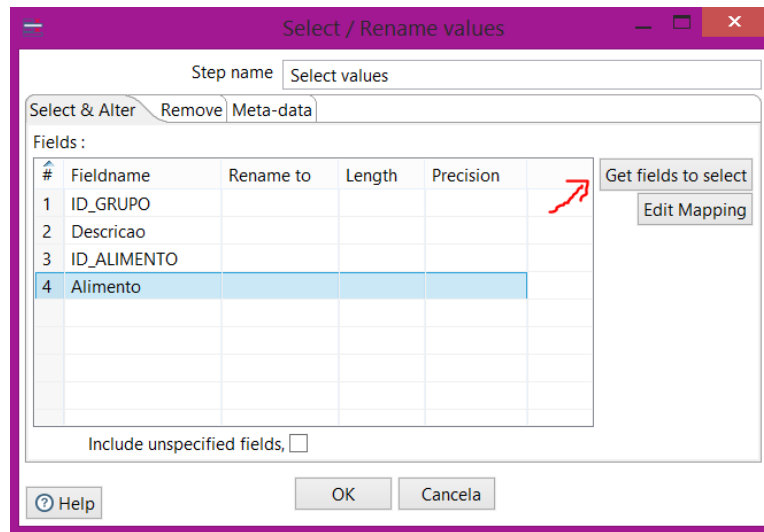
ID_ALIMENTO	Alimento	ID_ALIMENTO_1	ID_GRUPO
1.0	Arroz, integral, cozido	1.0	1.0
2.0	Arroz, integral, cru	2.0	1.0
3.0	Arroz, tipo 1, cozido	3.0	1.0
4.0	Arroz, tipo 1, cru	4.0	1.0
5.0	Arroz, tipo 2, cozido	5.0	1.0
6.0	Arroz, tipo 2, cru	6.0	1.0
7.0	Aveia, flocos, crua	7.0	1.0
8.0	Biscoito, doce, maisena	8.0	1.0
9.0	Biscoito, doce, recheado com chocolate	9.0	1.0
1.10.0	Biscoito, doce, recheado com morango	10.0	1.0
1.11.0	Biscoito, doce, wafer, recheado de chocolate	11.0	1.0

14. Em seguida, clique em Run para executar.

15. Com as três planilhas unidas podemos realizar as saídas. Porém antes iremos adicionar o step “SelectValues” para mapear os valores obtidos no “Merge Join”. Adicione o step e realize hop com o “Merge Join 2”.



16. Clique duas vezes no step “Select Value” e clique em [Get Fields to Select] e escolha quais campos você deseja apresentar na saída e clique em [ok].

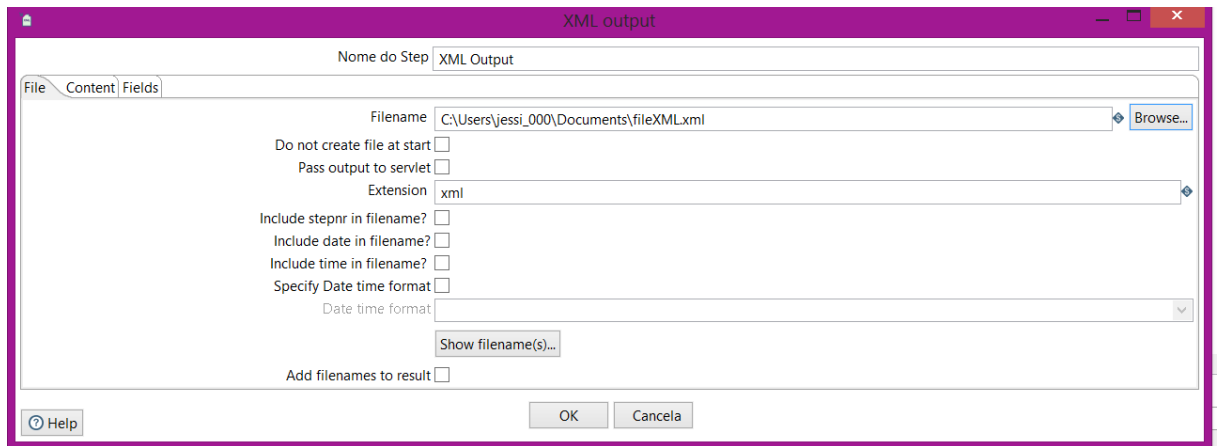


17. Clique em Run. Se tiver tudo ok adicione o step “XML Output” e realize hop.

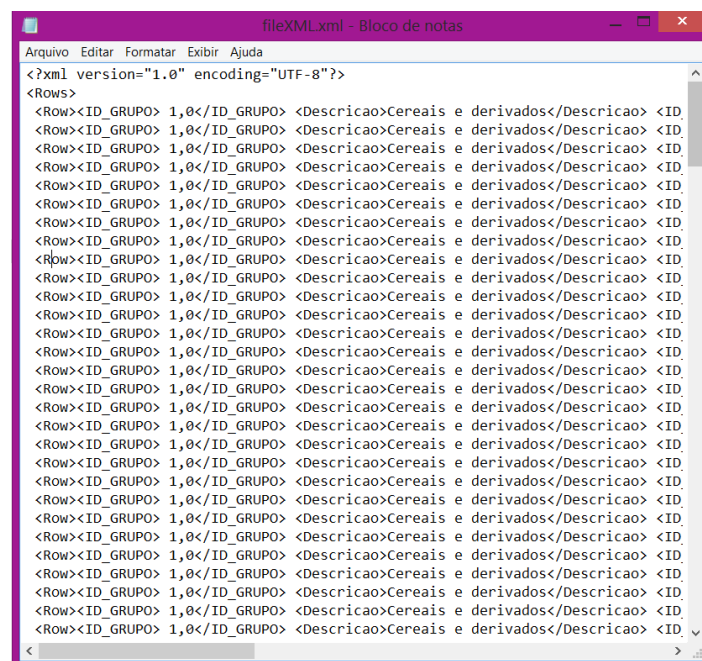
Execution Results

#	ID_GRUPO	Descricao	ID_ALIMENTO	Alimento	ID_ALIMENTO_1	ID_GRUPO_1
1	1,0	Cereais e derivados	1,0	Arroz, integral, cozido	1,0	1,0
2	1,0	Cereais e derivados	2,0	Arroz, integral, cru	2,0	1,0
3	1,0	Cereais e derivados	3,0	Arroz, tipo 1, cozido	3,0	1,0
4	1,0	Cereais e derivados	4,0	Arroz, tipo 1, cru	4,0	1,0
5	1,0	Cereais e derivados	5,0	Arroz, tipo 2, cozido	5,0	1,0
6	1,0	Cereais e derivados	6,0	Arroz, tipo 2, cru	6,0	1,0
7	1,0	Cereais e derivados	7,0	Aveia, flocos, crua	7,0	1,0
8	1,0	Cereais e derivados	8,0	Biscoito, doce, maisena	8,0	1,0
9	1,0	Cereais e derivados	9,0	Biscoito, doce, recheado com chocolate	9,0	1,0
10	1,0	Cereais e derivados	10,0	Biscoito, doce, recheado com morango	10,0	1,0
11	1,0	Cereais e derivados	11,0	Biscoito, doce, wafer, recheado de chocolate	11,0	1,0

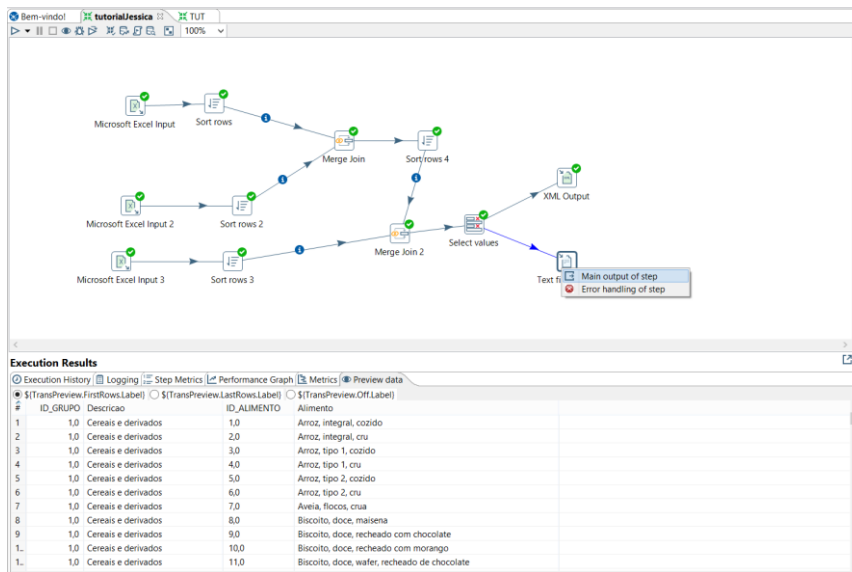
18. Clique duas vezes no step XML e na aba “File” configure para qual diretório você deseja salvar o arquivo clicando em [Browse]. Clique em [ok].



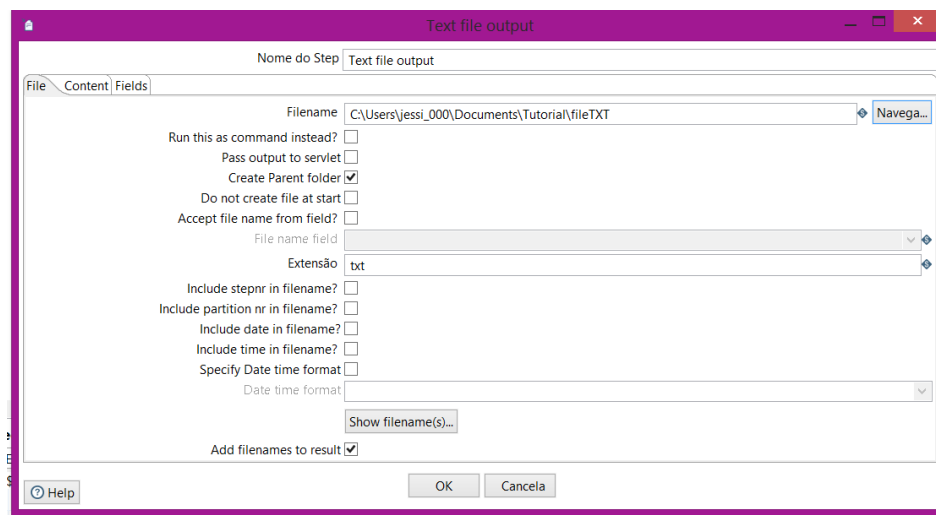
19. Clique em Run. Se tiver tudo ok, vá na pasta onde está o arquivo e verifique o arquivo criado.



20. Para realizar saída em arquivo .txt adicione o step “Text file output” e realize hop com “Select Value”.

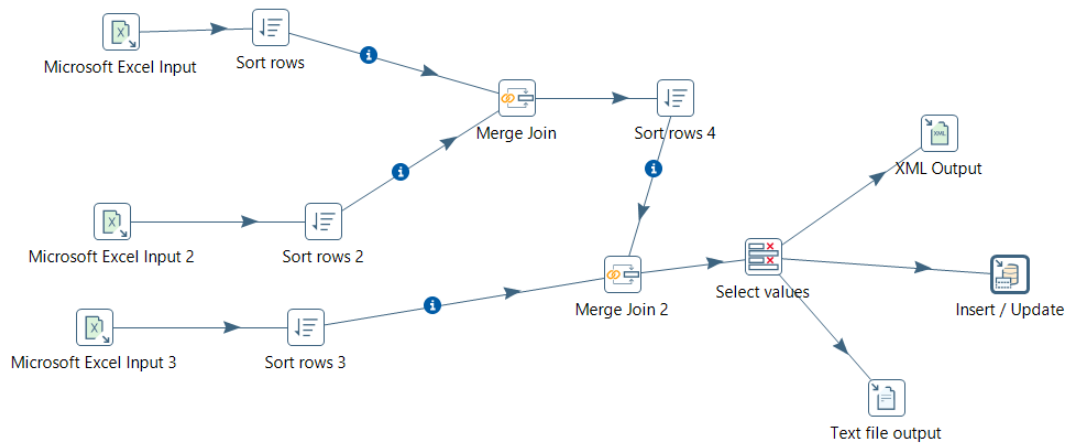


21. Em seguida clique duas vezes no step e na aba “File” configure para qual diretório você deseja salvar o arquivo clicando em [Navegar]. Clique em [ok].



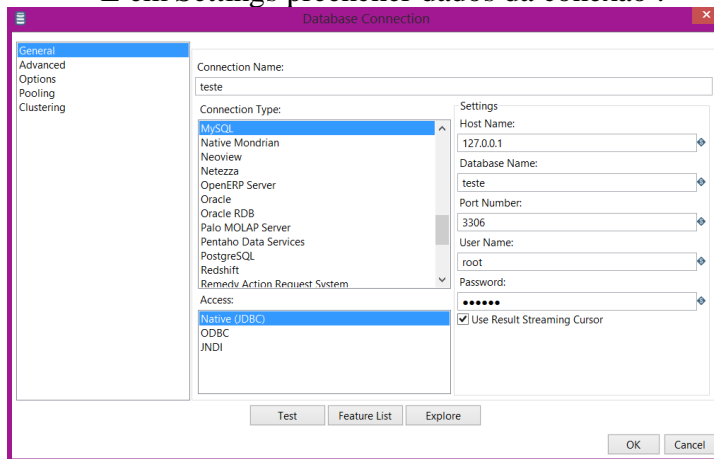
22. Clique em Run. Se tiver tudo ok, vá na pasta onde está o arquivo e verifique o arquivo criado.

23. Por último, adicione o step “Insert Update” e realize hop.

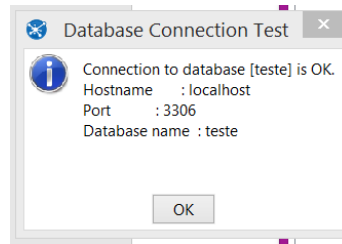


24. Com esse step fazemos a transformação em banco de dados. Para editá-lo clique duas vezes no mesmo com as seguintes configurações:

- Em connection, clicar em New
- Colocar o nome da conexão em “Connnection Name”
- Connection Type escolher MySQL
- Access escolher JDBC (Lembrar de adicionar o construtor na pasta lib do PDI, Pode ser baixado em < <https://dev.mysql.com/downloads/connector/j/>>)
- E em Settings preencher dados da conexão .

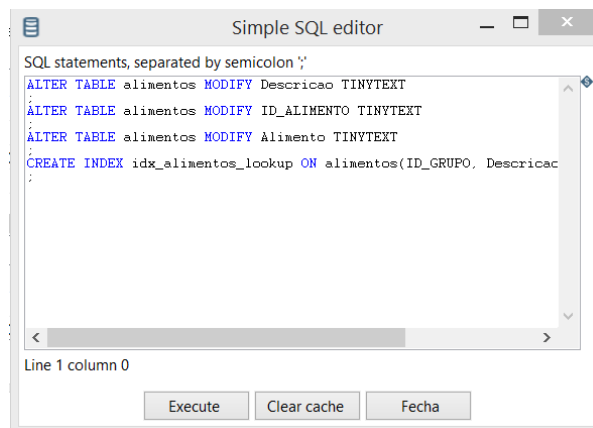


- Clicar em [test], se tiver ok irá aparecer uma janela da conexão e clicar em [ok].



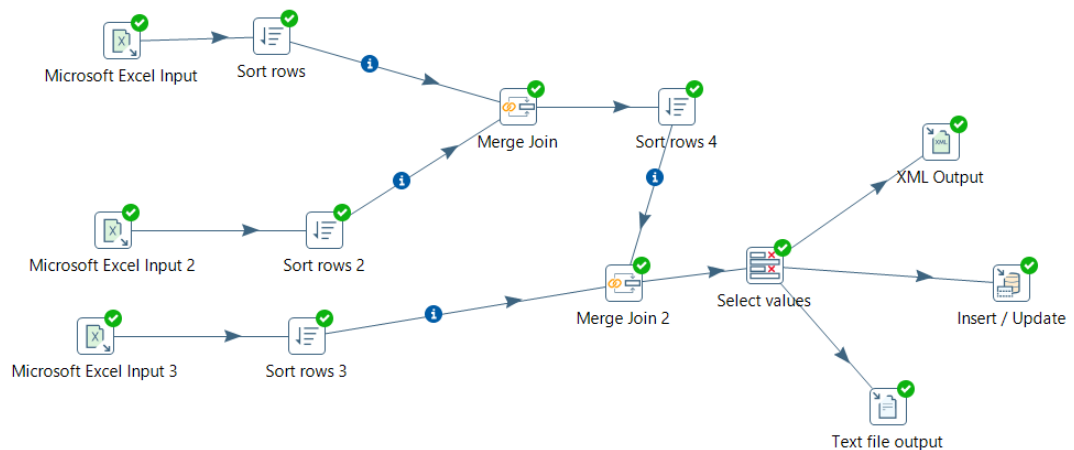
- Clicar em [ok]
- De volta a tela Insert/Update, em Target Table, clicar em [browse] e procurar.
- Em seguida clicar em [Get Fields]

- Em seguida clicar em SQL
- Clicar em [Execute]



- Clicar em [OK]

25. Realizados todos os passos, clicar em Run e se tudo estiver ok verificar no banco.



26. Para o nosso exemplo tivemos:

	ID_GRUPO	Descricao	ID_ALIMENTO	Alimento
	1	Cereais e derivados	3,0	Arroz, tipo 1, cozido
	1	Cereais e derivados	6,0	Arroz, tipo 2, cru
▶	1	Cereais e derivados	9,0	Biscoito, doce, recheado com chocolate
	1	Cereais e derivados	12,0	Biscoito, doce, wafer, recheado de morango
	1	Cereais e derivados	15,0	Bolo, pronto, aipim
	1	Cereais e derivados	18,0	Bolo, pronto, milho
	1	Cereais e derivados	21,0	Cereais, milho, flocos, com sal
	1	Cereais e derivados	24,0	Cereais, mistura para vitamina, trigo, cevada e ...
	1	Cereais e derivados	27,0	Creme de arroz, pó
	1	Cereais e derivados	30,0	Curau, milho verde, mistura para
	1	Cereais e derivados	33,0	Farinha, de milho, amarela
	1	Cereais e derivados	36,0	Farinha, láctea, de cereais
	1	Cereais e derivados	39,0	Macarrão, trigo, cru, com ovos
	1	Cereais e derivados	42,0	Milho, verde, enlatado, drenado