

Aluna: Lays Lopes

Matrícula: 201110005994

Tutorial realizando o ETL através do PDI Pentaho em um banco de dados relacional

Passo 1: Com o PDI já baixado na sua máquina e sua pasta descompactada. Clique no arquivo Spoon.bat. Feito isso irá aparecer uma mensagem se deseja executar e clique em executar (Figura 1)

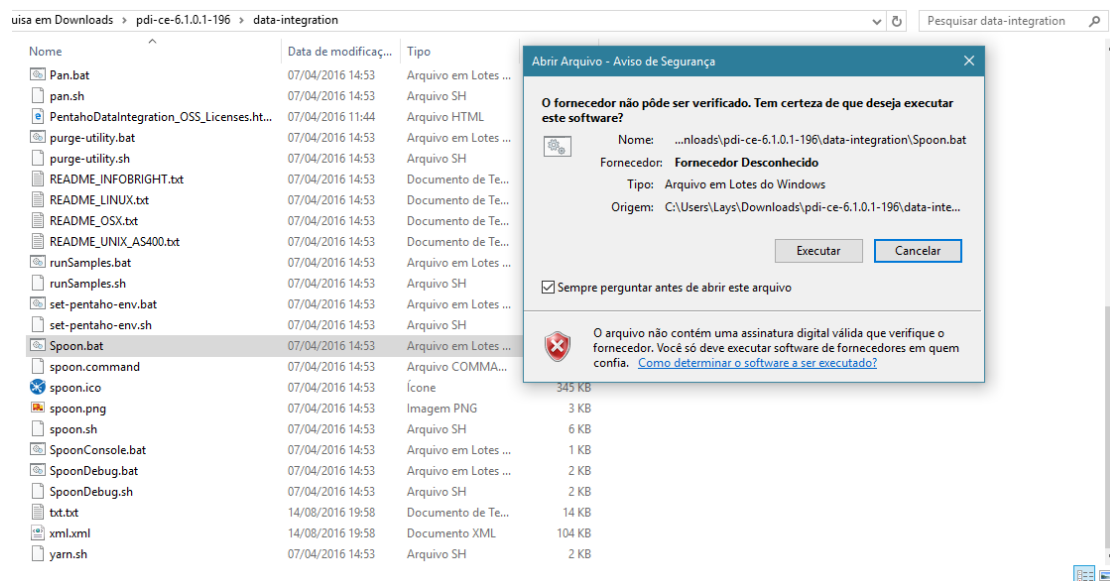


Figura 1

Passo 2. Irá aparecer uma tela carregando o Pentaho Data Integration (Figura 2)



Figura 2

Passo 3. Criando uma transformação, clique em File → Novo → Transformação. (Figura 3)

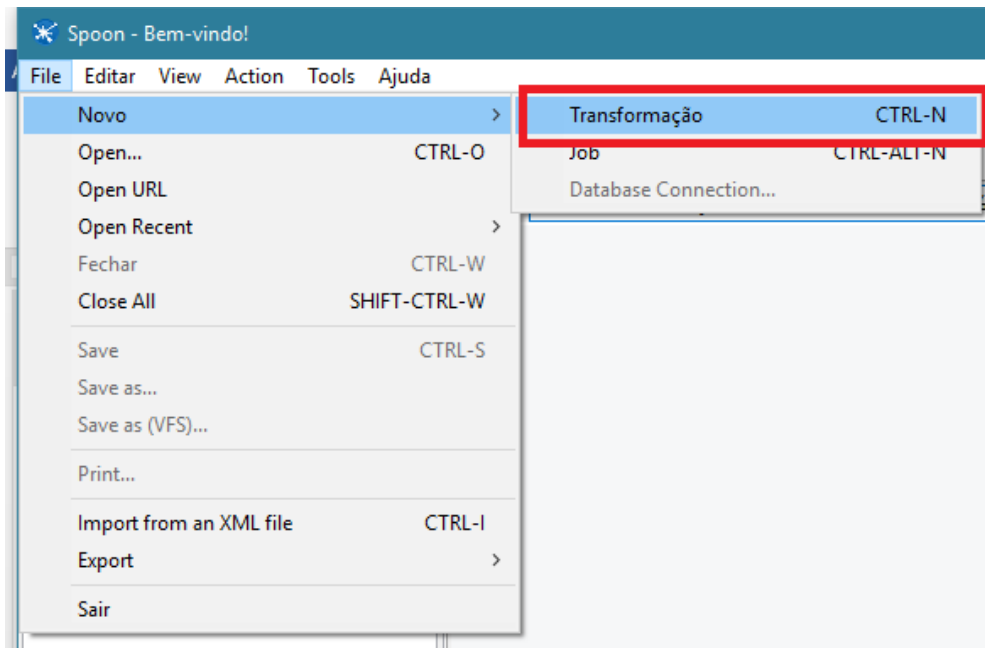


Figura 3

Passo 4. Clique em Design para escolher o tipo de arquivo que irá utilizar. (Figura 4)

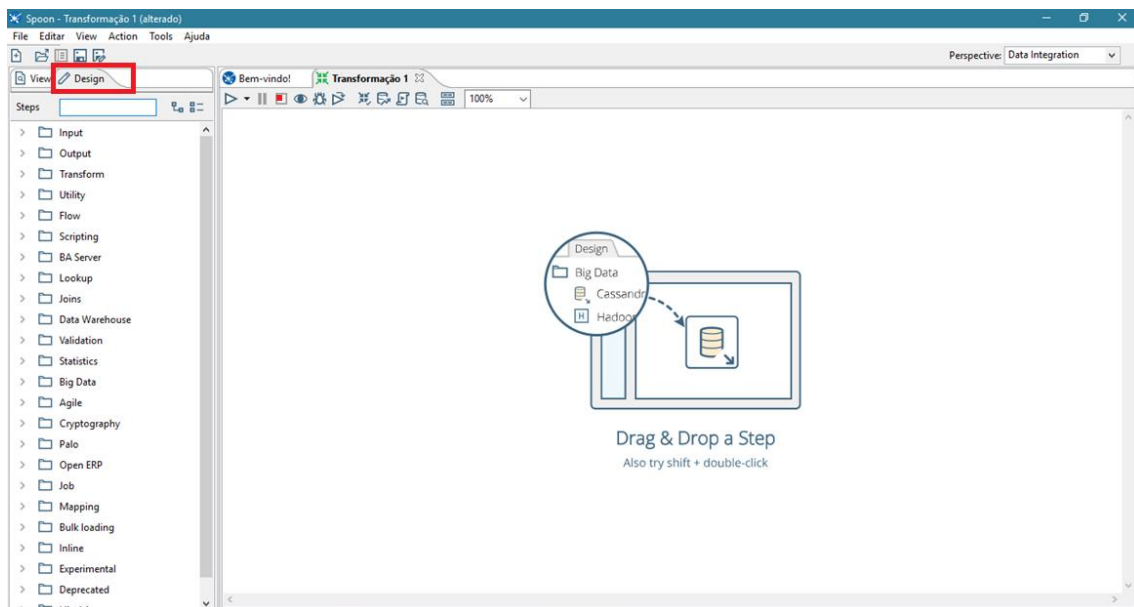


Figura 4

Passo 5. Clique em Input e escolha o tipo arquivo que ser utilizado para realizar o ETL. O arquivo que irei utilizar é .XLSX. Então clique duas vezes Microsoft Excel Input. (Figura 5)

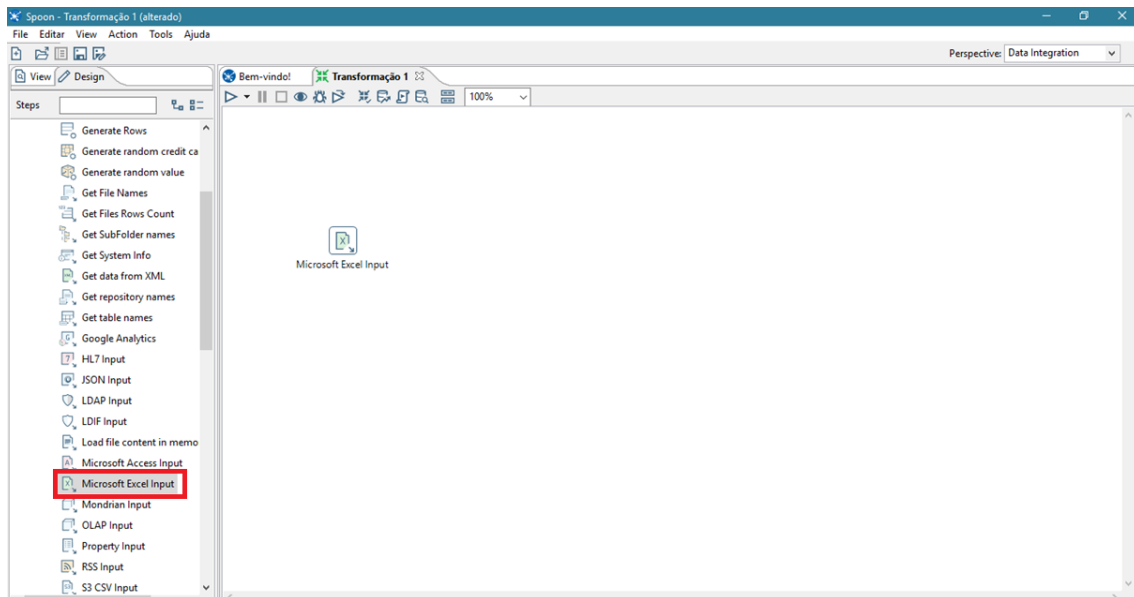


Figura 5

Passo 6. Clique duas vezes Microsoft Excel Input que apareceu na área de transformação para colocar o arquivo e selecionar os dados. Clique em Navegar para escolher o arquivo. Por padrão a aba Files fica selecionada. (Figura 6)

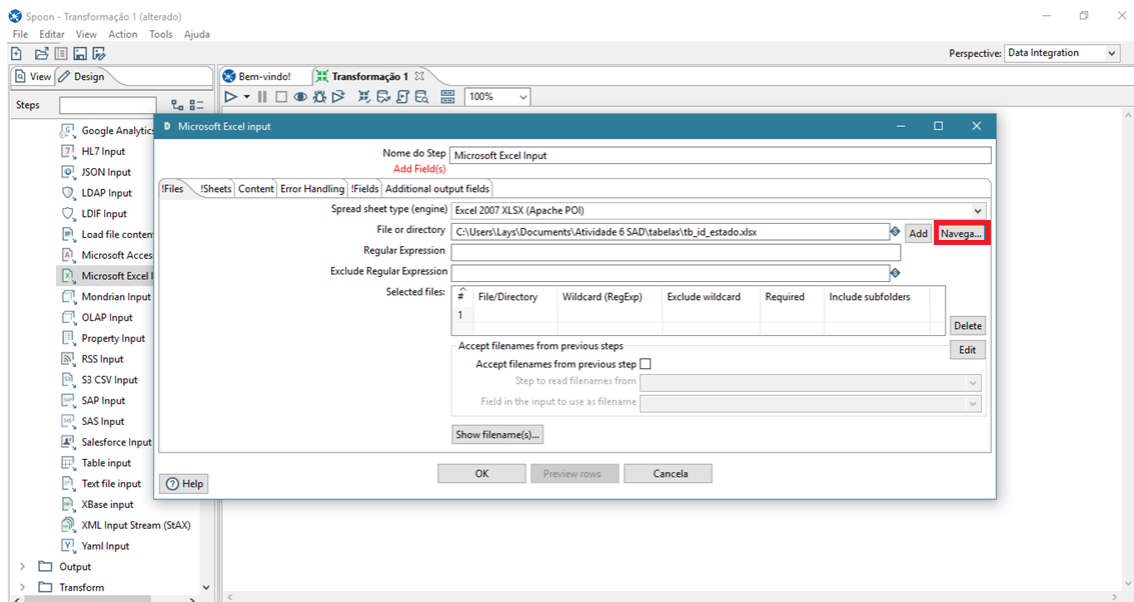


Figura 6

Passo 7. Após escolher o arquivo, clique em Add. Observe que o arquivo irá aparecer em Selected Files. (Figura 7)

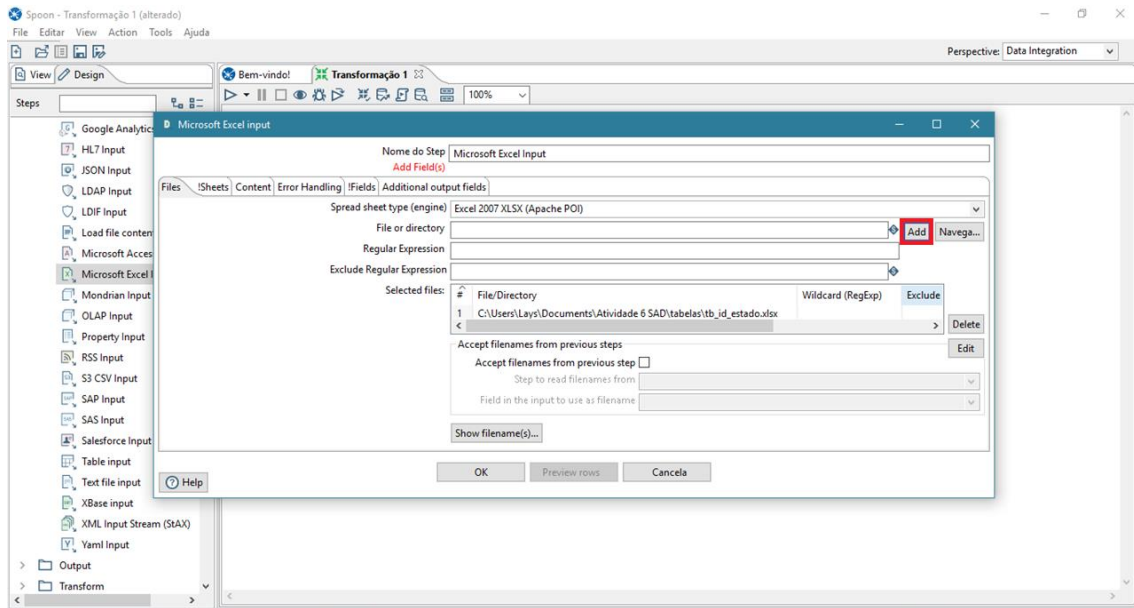


Figura 7

Passo 8. Na aba Sheets, clique em Get Sheetname(s) . (Figura 8)

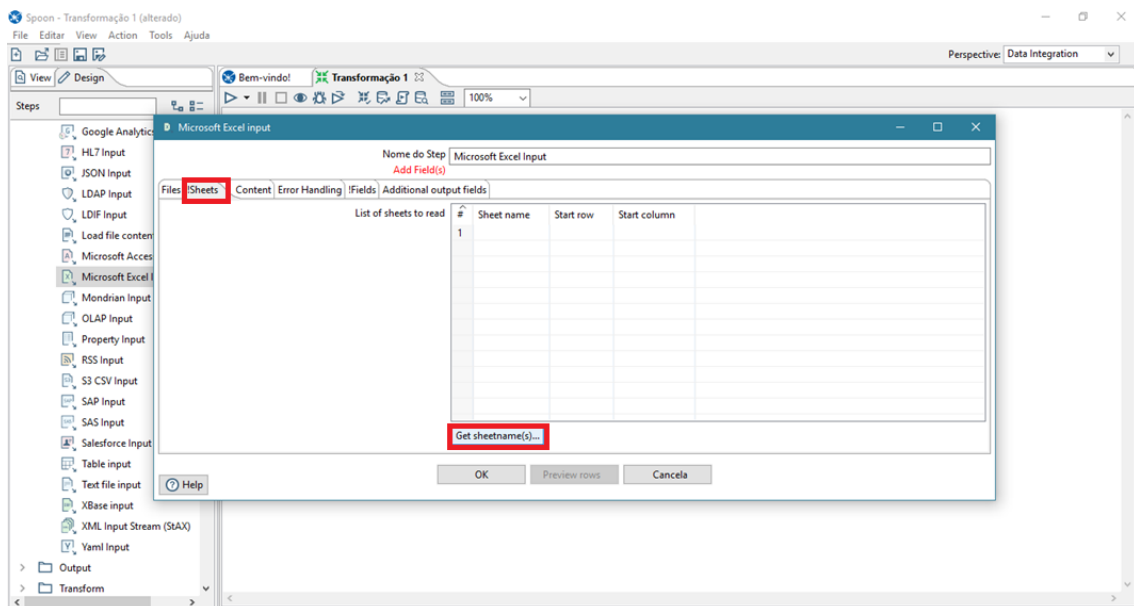


Figura 8

Passo 9. Após clicar em Get Sheetname(s), irá aparecer a planilha com os dados, clique na seta > para selecionar e depois clique em OK . (Figura 9)

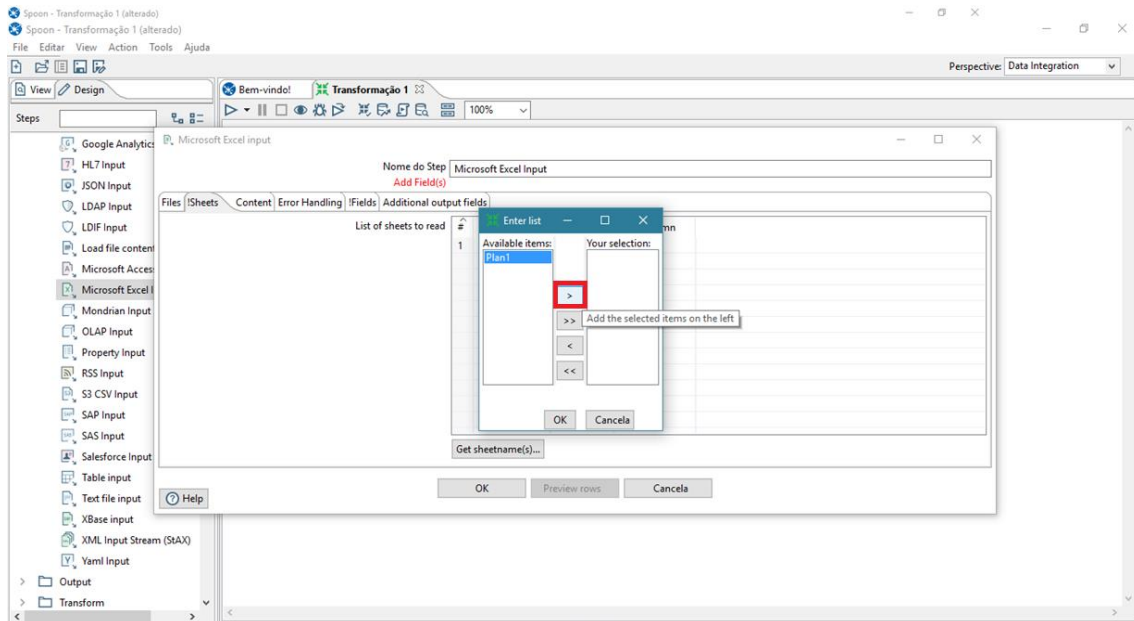


Figura 9

Passo 10. Clique na aba !Fields e posteriormente em Get fields from header row ... para selecionar as colunas da sua planilhas. (Figura 10)

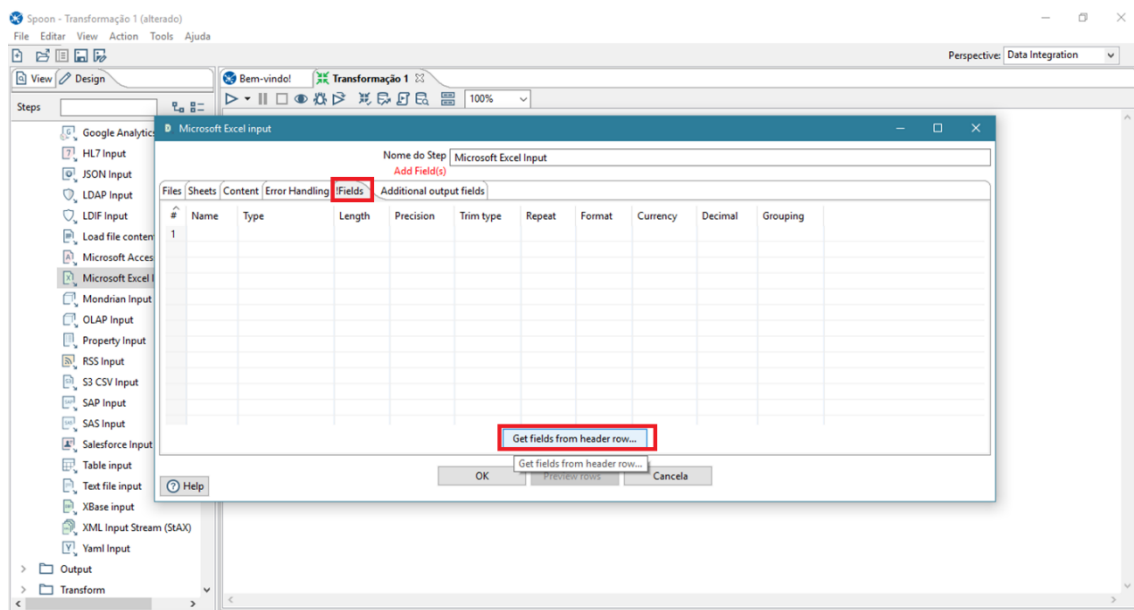


Figura 10

Passo 11. Feito isso irá aparecer as colunas da sua tabela, verifique se o tipo (type) do campo está correto e se no caso da minha tabela no campo id_estado havia aparecido o número com vírgula. Para que isso não ocorresse na coluna format escolha o tipo # pois assim ele irá pegar exatamente como está na coluna do seu arquivo excel. Clique em Preview rows para exibir uma prévia das linhas do seu arquivo. (Figura 11)

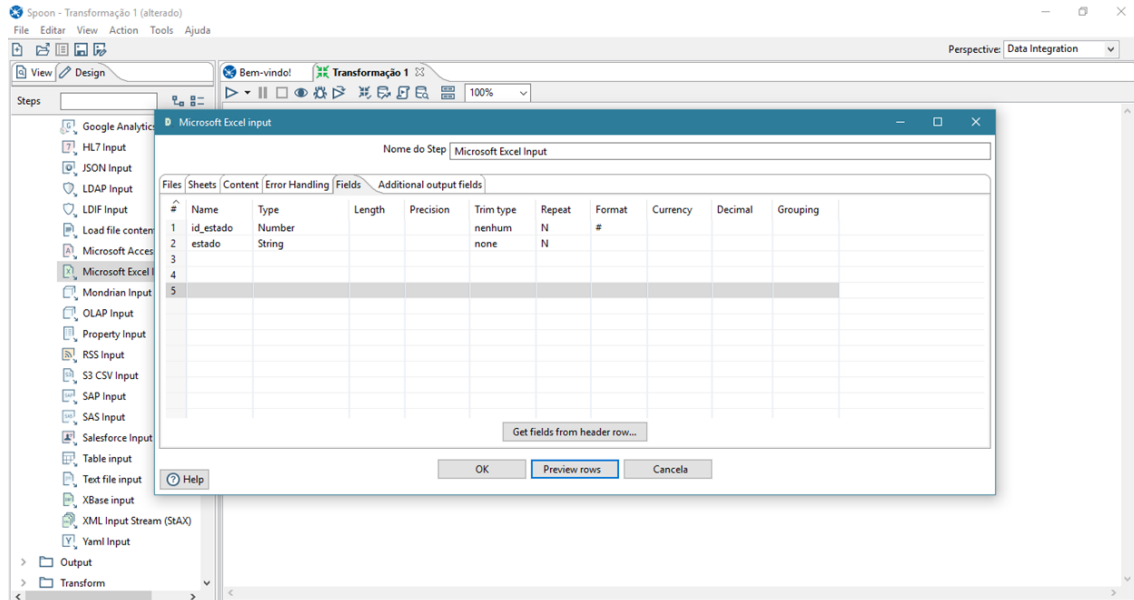


Figura 11

Passo 12. Após selecionar Preview rows irá aparecer uma mensagem perguntando quantas linhas você deseja selecionar, como meu arquivo possui 27 coloquei 30. (Figura 12)

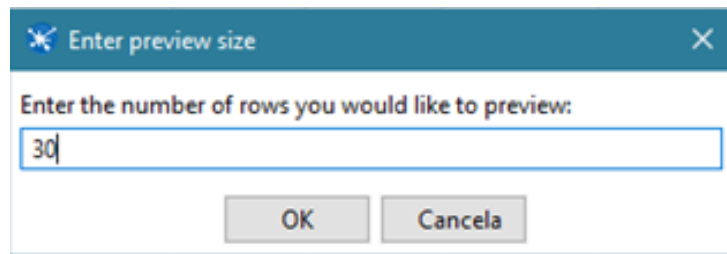
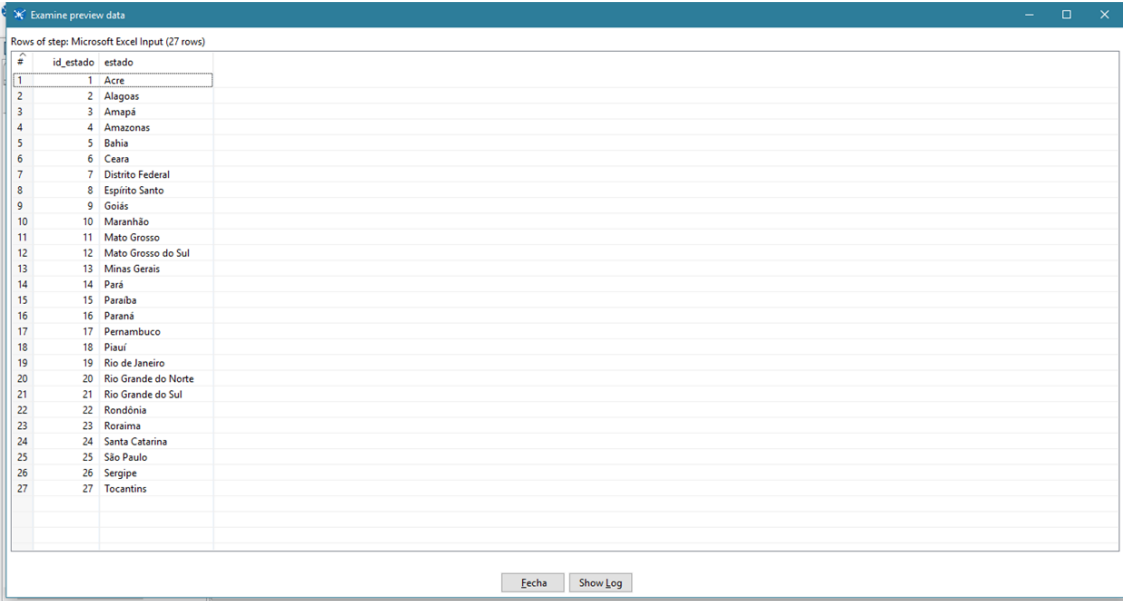



Figura 12

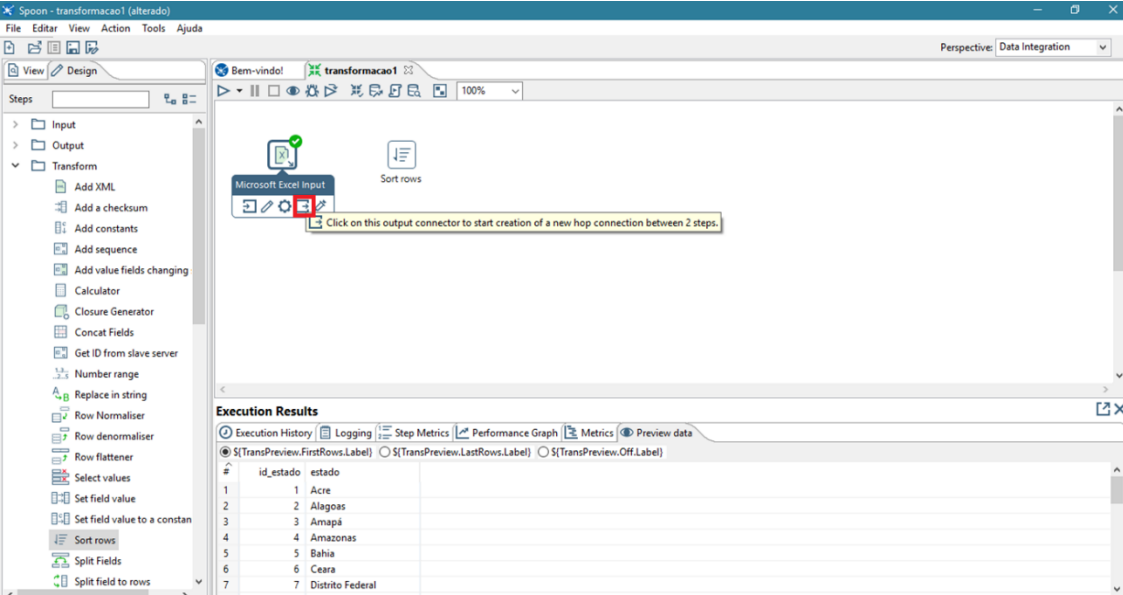
Passo 13. Previa das colunas e linhas do seu arquivo Excel. (Figura 13)



#	id_estado	estado
1	1	Acre
2	2	Alagoas
3	3	Amapá
4	4	Amazonas
5	5	Bahia
6	6	Ceara
7	7	Distrito Federal
8	8	Espírito Santo
9	9	Goiás
10	10	Maranhão
11	11	Mato Grosso
12	12	Mato Grosso do Sul
13	13	Minas Gerais
14	14	Pará
15	15	Paraíba
16	16	Paraná
17	17	Pernambuco
18	18	Piauí
19	19	Rio de Janeiro
20	20	Rio Grande do Norte
21	21	Rio Grande do Sul
22	22	Rorônia
23	23	Roraima
24	24	Santa Catarina
25	25	São Paulo
26	26	Sergipe
27	27	Tocantins

Figura 13

Passo 14. Próximo passo é fazer um SORT ROWS para ordenar as colunas selecionadas. Para isso selecione Transform, procure a transformação SORT ROWS e arraste até a área branca de transformação, dê um clique em Microsoft excel clique no símbolo  ao lado da engrenagem para ligar o arquivo de entrada a transformação de ordenação irá ser feita nesse arquivo. (Figura 14).



#	id_estado	estado
1	1	Acre
2	2	Alagoas
3	3	Amapá
4	4	Amazonas
5	5	Bahia
6	6	Ceara
7	7	Distrito Federal

Figura 14

Passo 15. Clique duas vezes no ícone de SORT ROWS para configurar a ordenação do arquivo. Clique em Obtem Campos, feito isso irá aparecer os campos do arquivo, em seguida em Ascending informe S para ordenar de forma crescente os campos. (Figura 15).

Sort rows

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 keys only) ☐

Fields:

#	Fieldname	Ascending	Case sensitive compare?	Presorted?
1	estado	S	N	N
2	id_estado	S	N	N

Buttons: Help, OK, Cancela, Obtem campos

Figura 15

Passo 16. Repita os passos anteriores para os demais arquivos. (Figura 16).

Execution Results

#	id_estado	estado
1	1	Acre
2	2	Alagoas
3	3	Amapá
4	4	Amazonas
5	5	Bahia
6	6	Ceará
7	7	Distrito Federal

Figura 16

Passo 17. Agora vamos juntar dois arquivos em um. No lado esquerdo selecione Joins, procure o MERGE JOIN, arraste para área branca e clique em SORT ROWS e SORT ROWS 2 ligue ao MERGE JOIN . (Figura 17)

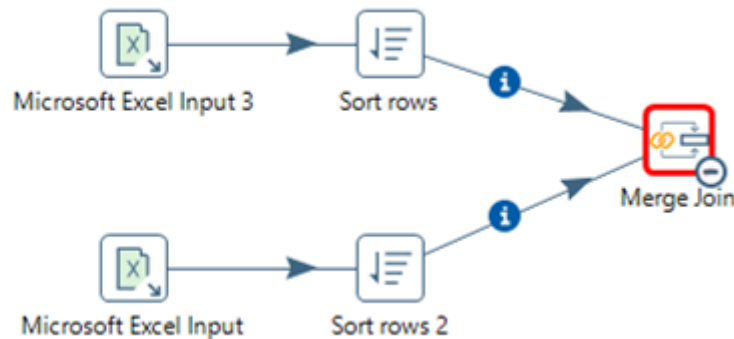


Figura 17

Feito isso agora vamos configurar o MERGE JOIN. Selecione o primeiro arquivo que deseja que apareça e depois selecione o segundo e o tipo de seleção se é JOIN, LEFT JOIN(...). Como queremos todos os dados iremos fazer um JOIN. Clique em Get key fields nas duas tabelas para comparar as chaves das tabelas , pois caso sejam iguais irá realizar o INNER JOIN, depois clique em OK. Irá aparecer uma mensagem dizendo que para fazer o MERG JOIN é necessário que os campos das chaves estejam ordenados de forma crescente, como já fizemos isso no SORT ROWS não precisamos nos preocupar.

A janela 'Merge Join' apresenta as seguintes configurações:

- Step name: Merge Join
- First Step: Sort rows
- Second Step: Sort rows 2
- Join Type: INNER

Abas para configuração de chaves:

#	Key field
1	id_estado


Get key fields

#	Key field
1	id_estado

Get key fields

Botões: ? Help, OK, Cancela

Figura 18

Passo 18. Próximo passo é fazer um SORT ROWS para ordenar as colunas geradas pelo MERG JOIN . Para isso selecione Transfor, procure a transformação SORT ROWS e arraste até a área branca de transformação, dê um clique em Microsoft excel clique no símbolo  ao lado da engrenagem para ligar o arquivo de entrada a transformação de ordenação irá ser feita nesse arquivo. (Figura 19).

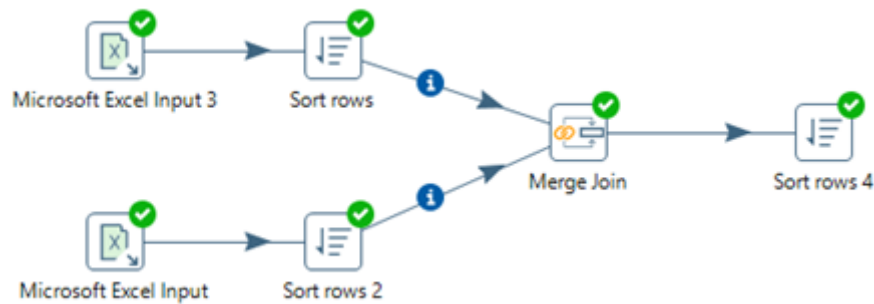


Figura 19

Passo 19. Agora vamos juntar o arquivo gerado do MERG JOIN e ordenado com o SORT ROWS 4 com o terceiro arquivo. No lado esquerdo selecione Joins, procure o MERGE JOIN, arraste para área branca e clique em SORT ROWS 3 e SORT ROWS 4 ligue ao MERGE JOIN. (Figura 20)

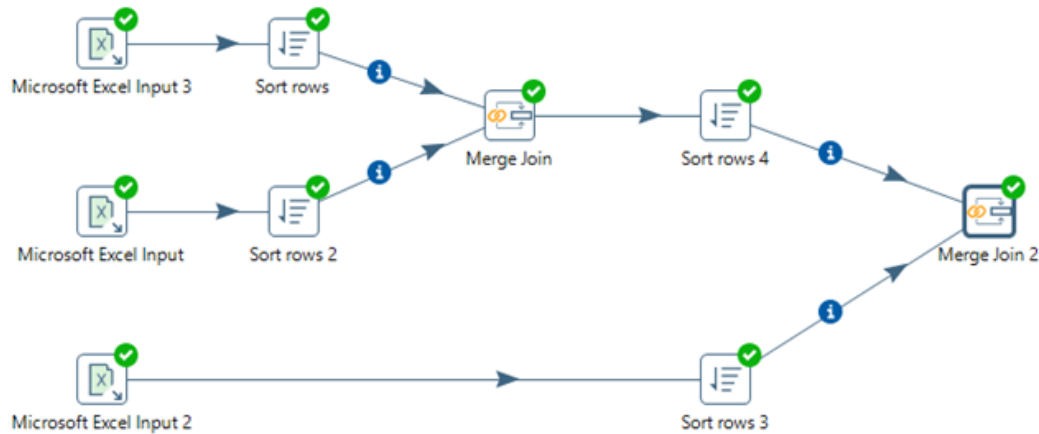


Figura 20

Todas as colunas e linhas dos três arquivos unidos em um único arquivo.

Execution Results

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

☒ S{TransPreview.FirstRows.Label} ☐ S{TransPreview.LastRows.Label} ☐ S{TransPreview.Off.Label}

#	id_estado	sigla	id_estado_1	estado	id_estado_2	regiao	id_regiao
1	1	AC	1	Acre	1	NORTE	1
2	2	AL	2	Alagoas	2	NORDESTE	2
3	3	AP	3	Amapá	3	NORTE	1
4	4	AM	4	Amazonas	4	NORTE	1
5	5	BA	5	Bahia	5	NORDESTE	2
6	6	CE	6	Ceara	6	NORDESTE	2

Figura 21

Passo 20. Feito o MERGE JOIN vamos selecionar quais colunas (valores) iremos quer dessa junção. No lado esquerdo clique em Transfor e arraste o ícone Select Values até a área branca e faça a ligação com o MERG JOIN 2. (Figura 21)

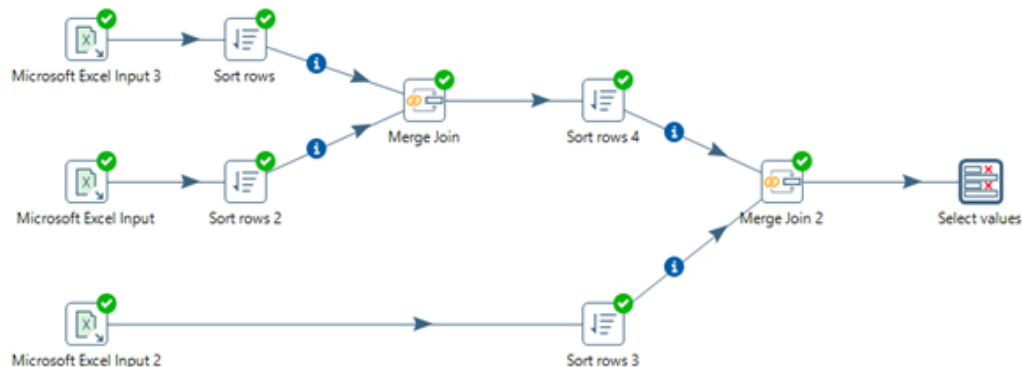


Figura 22

Clique em Select values e vamos selecionar quais colunas iremos querer. Clique em GET FIELDS TO SELECT para obter todos os campos do MERG JOIN. Caso deseje em Fieldname você poderá alterar o nome da coluna. (Figura 23)

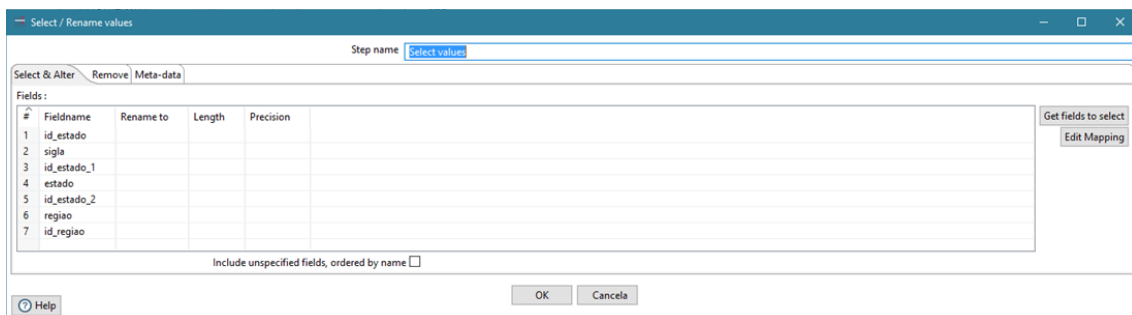


Figura 23

Caso deseje remover algum campo repetido, clique na aba “Remove” e depois em GET FIELDS TO REMOVE e deixe apenas os campos que você deseja remover. Clique em OK. (Figura 24)

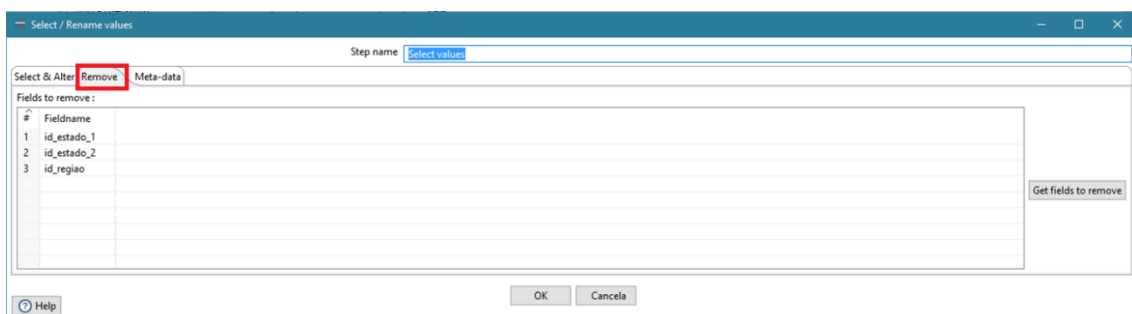



Figura 24

Passo 21. Clique no botão  para executar a transformação e verificar se os dados que você selecionou saíram corretamente. (Figura 25)

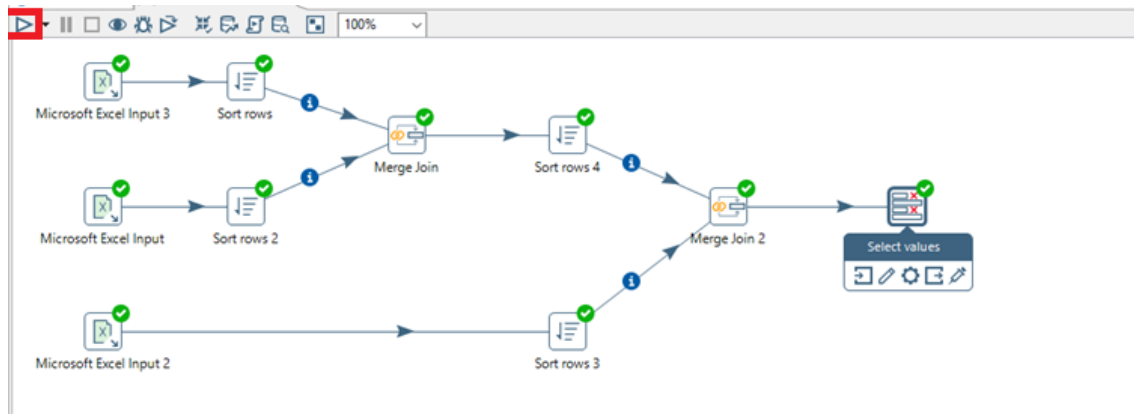


Figura 25

Como resultado verificamos que veio corretamente o que foi selecionado. (Figura 26)

Execution Results

Execution History Logging Step Metrics Performance Graph Metrics Preview data					
<input checked="" type="radio"/> \${TransPreview.FirstRows.Label} <input type="radio"/> \${TransPreview.LastRows.Label} <input type="radio"/> \${TransPreview.Off.Label}					
#	id_estado	sigla	estado	regiao	
1	1	AC	Acre	NORTE	
2	2	AL	Alagoas	NORDESTE	
3	3	AP	Amapá	NORTE	
4	4	AM	Amazonas	NORTE	
5	5	BA	Bahia	NORDESTE	
6	6	CE	Ceara	NORDESTE	
7	7	DF	Distrito Federal	CENTRO OESTE	

Figura 26

Passo 22. Agora vamos exportar o arquivo em XML e TXT. No lado esquerdo clique em Output e selecione Text File output e XML output, faça as ligações do Select value com o Text File output e XML output. Irá aparecer um alerta clique em Main output of step. (Figura 27)

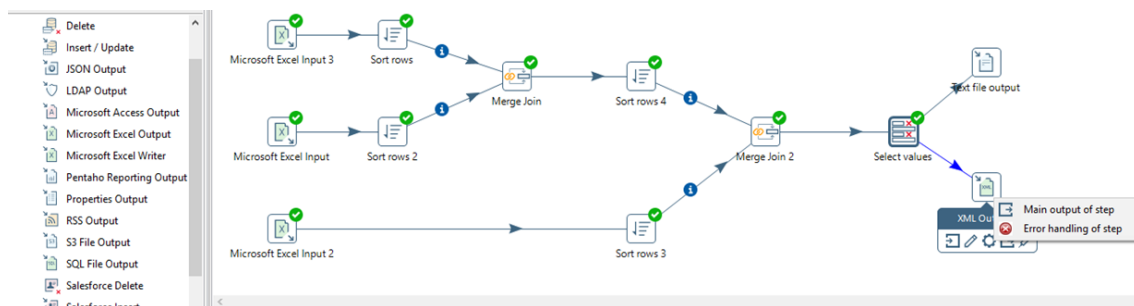


Figura 27

Como iremos transformar em mais de uma saída ao tentar ligar o Select value ao segundo step de saída irá aparecer uma mensagem perguntando se deseja distribuir as linhas ou copiar clique em “Copiar”, pois assim todas as linhas serão enviadas para todas as saídas. (Figura 28)

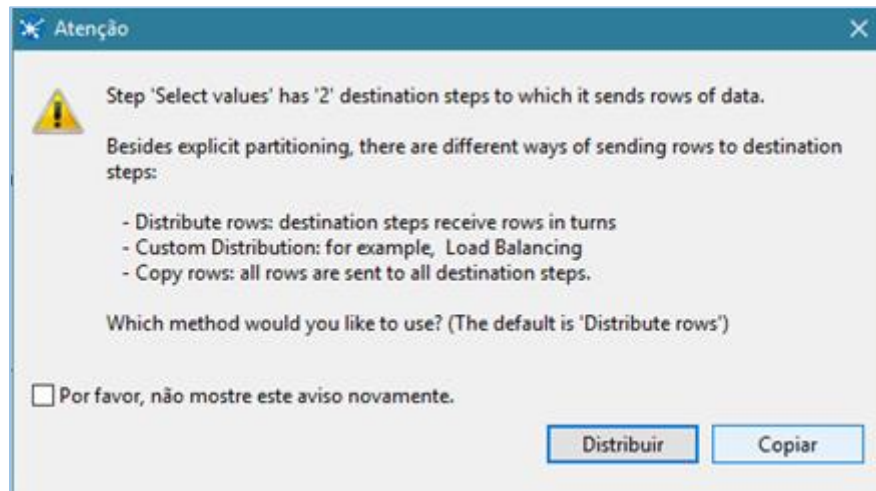


Figura 28

Clique com o botão direito do mouse no step de saída para configurar o nome do arquivo que será gerado. Em File no campo “Filename” digite o nome do arquivo e clique em Navega... para escolher a pasta onde será salvo o mesmo e depois clique em “OK”. Faça o mesmo para as demais saídas. (Figura 29)

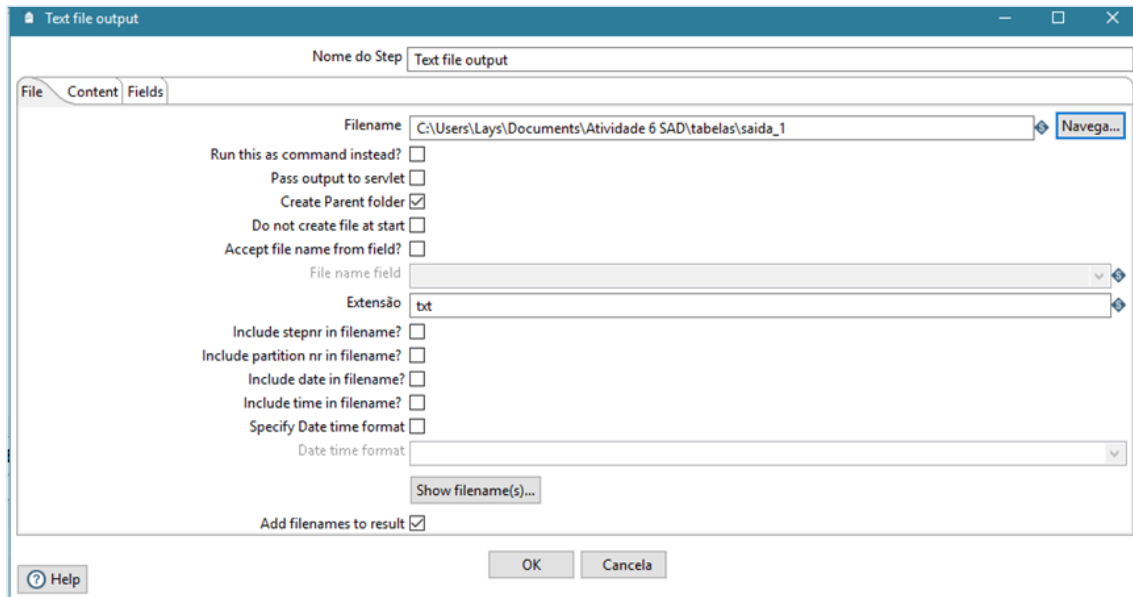



Figura 29

Passo 23. Clique no botão  para executar a transformação e verificar se os dados foram exportados corretamente.

Passo 24. Exportando a tabela e criando ela em um banco de dados. No lado direito clique em Output e selecione Insert/Update e arraste para a sua transformação e ligue com o Select Value. (Figura 30)

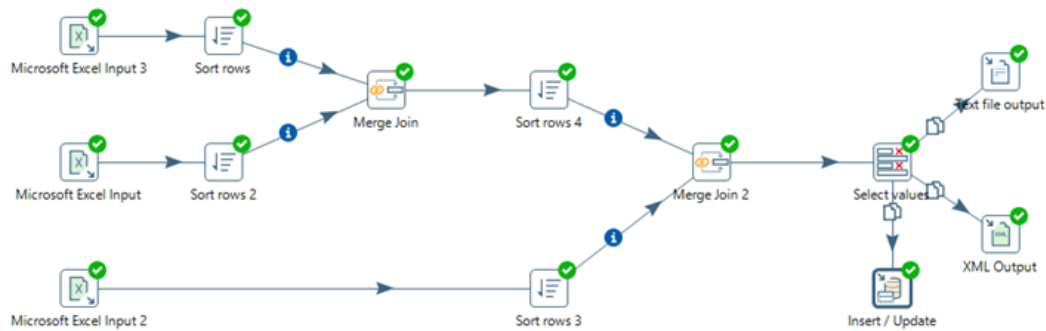


Figura 30

Clique duas vezes sobre o ícone do INSERT/UPDATE e vamos configurar. Primeiro passo é se conectar ao banco o qual deseja inserir a tabela. Clique em Wizard para selecionar o banco já existente. (Figura 31)

Step name: Insert / Update

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

Target schema: [dropdown] Navega...

Target table: lookup table Browse...

Commit size: 100

Don't perform any updates: ☐

The key(s) to look up the value(s):

#	Table field	Comparator	Stream field1	Stream field2
1				

Get fields

Update fields:

#	Table field	Stream field	Update
1			

Get update fields

Edit mapping

Help OK Cancela SQL

Figura 31

Em Name of database(...) Coloque o nome da conexão, selecione o tipo de banco que deseja conectar no caso utilizo o PostgreSQL e o tipo de acesso Native. Clique em Next. (Figura 32)

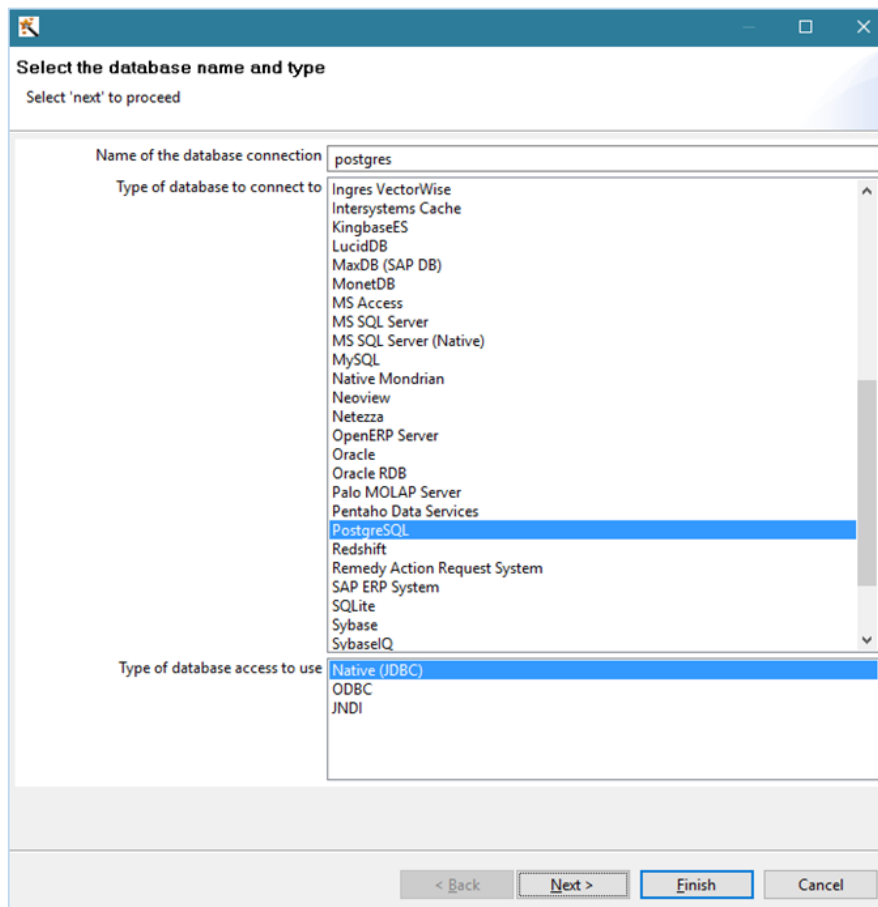
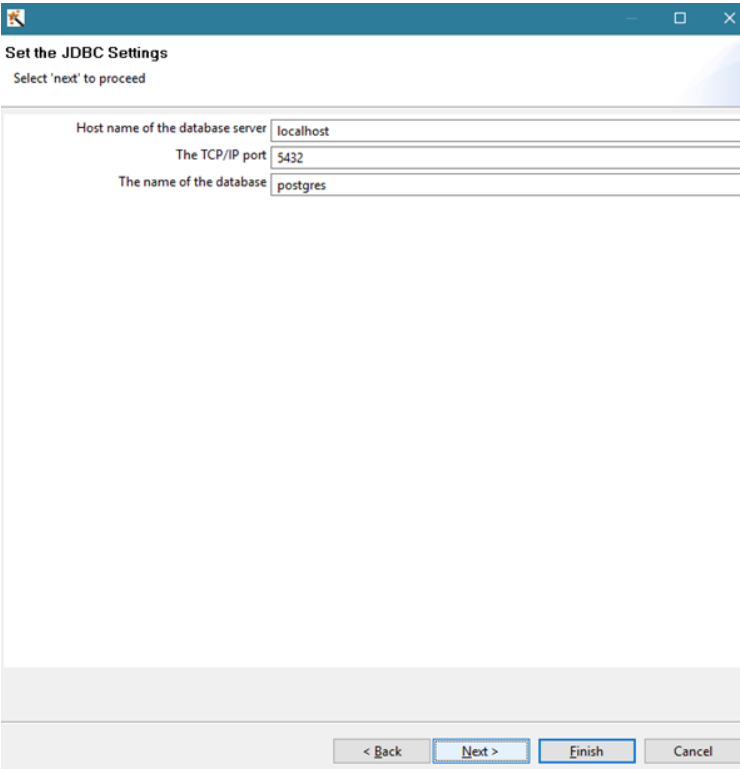


Figura 32

Coloque o nome do host do banco e o nome da base de dados que deseja inserir esses dados.



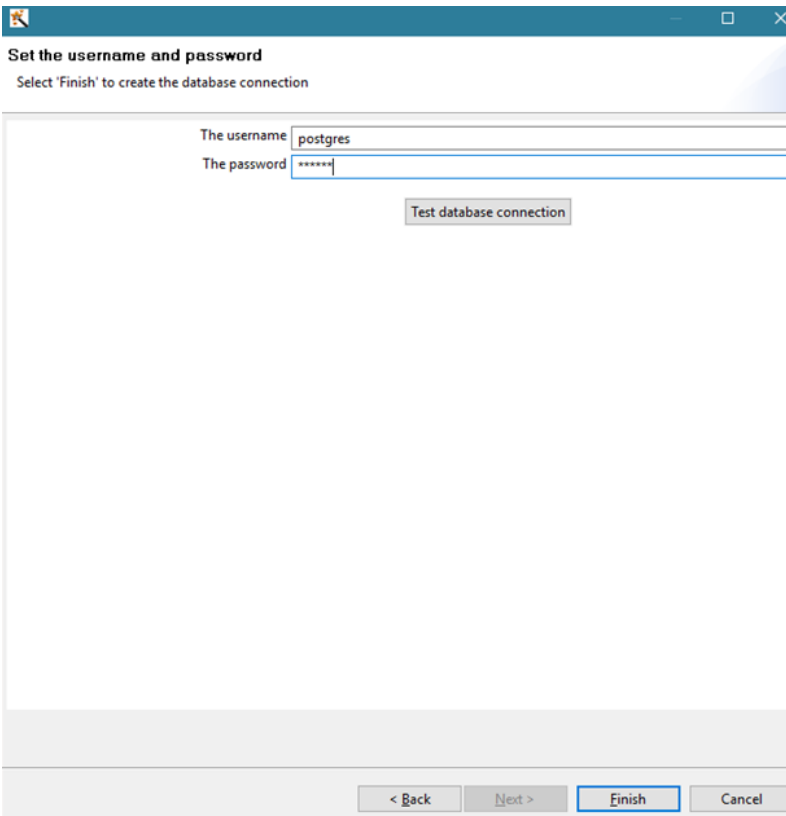
Set the JDBC Settings
Select 'next' to proceed

Host name of the database server localhost
The TCP/IP port 5432
The name of the database postgres

< Back Next > Finish Cancel

Figura 33

Coloque o nome do superusuário do banco e a senha. E clique em Test database connection.



Set the username and password
Select 'Finish' to create the database connection

The username postgres
The password *****

Test database connection

< Back Next > Finish Cancel

Figura 34

Após clicar caso a conexão dê OK. Pronto seu banco consegue se comunicar com o ETL. (Figura 35)

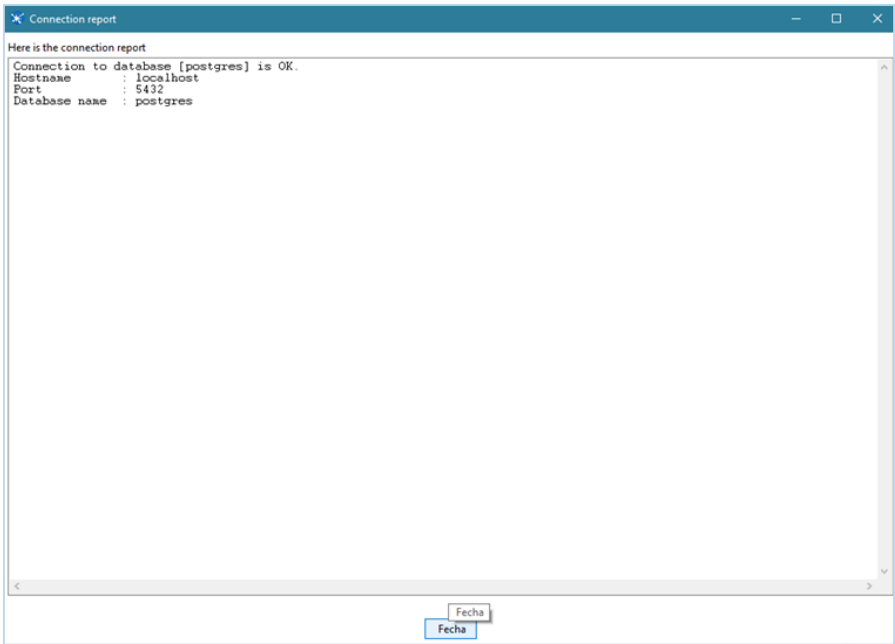


Figura 35

Feito isso vamos continuar a configurar o INSERT/UPDATE. A connection já vai aparecer que foi a que configuramos anteriormente. No Target schema irá aparecer os esquemas da sua base de dados , selecione um para que seja colocado nele a nova tabela. Em target table coloque o nome da sua tabela ou caso vá fazer um update em uma selecione ela. Clique em GET FIELDS para selecionar os campos da tabela e clique em GET UPDATE para atualizar os campos das tabelas. (Figura 36)

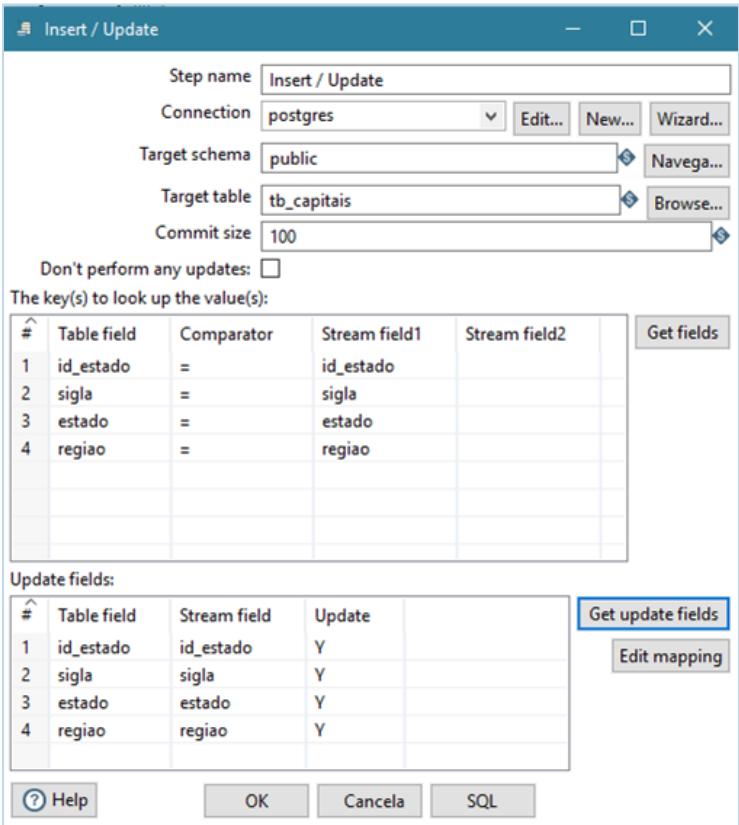


Figura 36

Clique SQL para criar a tabela no banco caso esta não existe e depois clique em executar. (Figura 37)

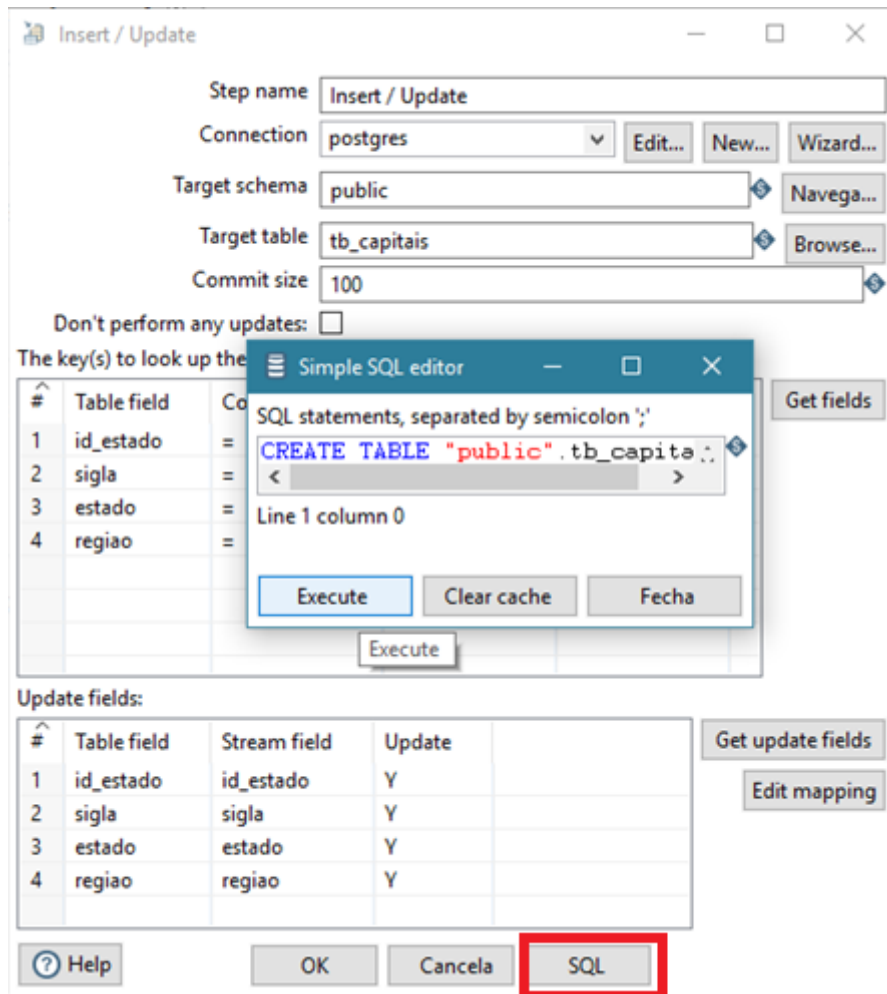


Figura 37

Ir   exibir o Script do Create e depois clique em OK. (Figura 38)

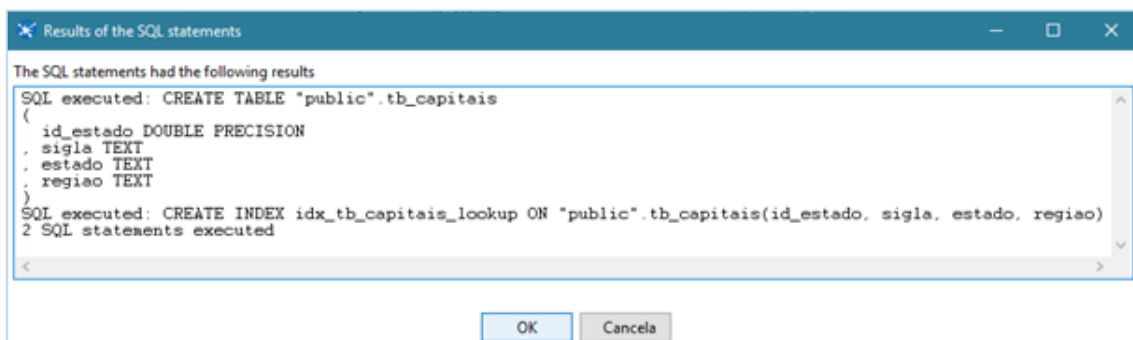



Figura 38

Passo 25. Clique no botão  para executar a transformação e verificar se os dados foram inseridos na sua base de dados. (Figura 39)

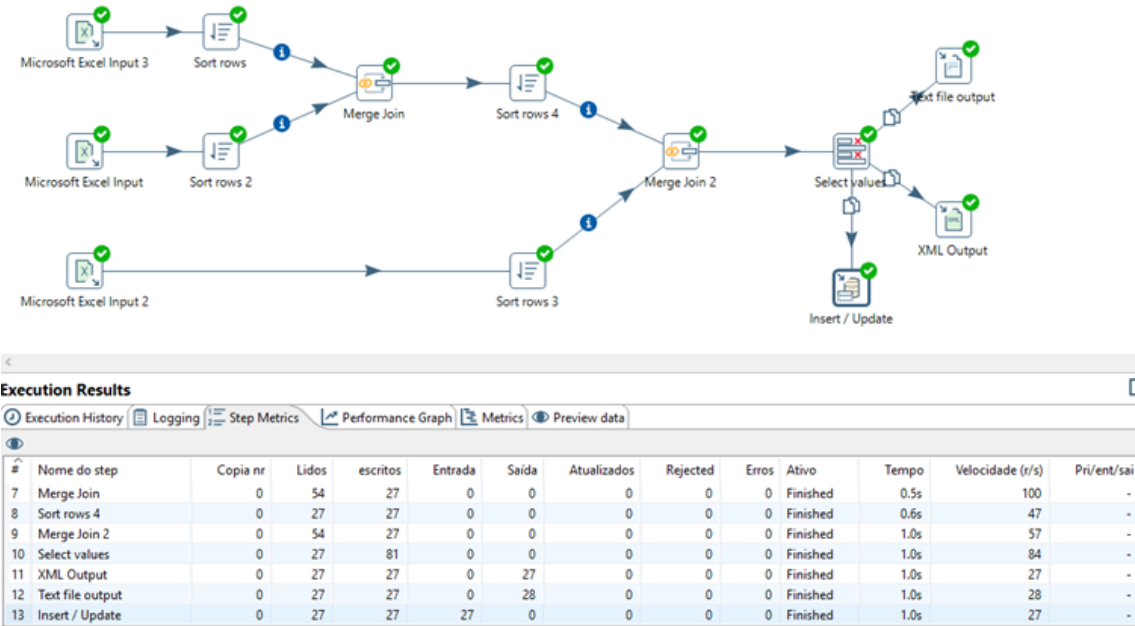


Figura 39

Passo 26. Vá no banco de dados e verifique se a tabela foi criada e dê um select para ver se os dados foram inseridos. (Figura 40)

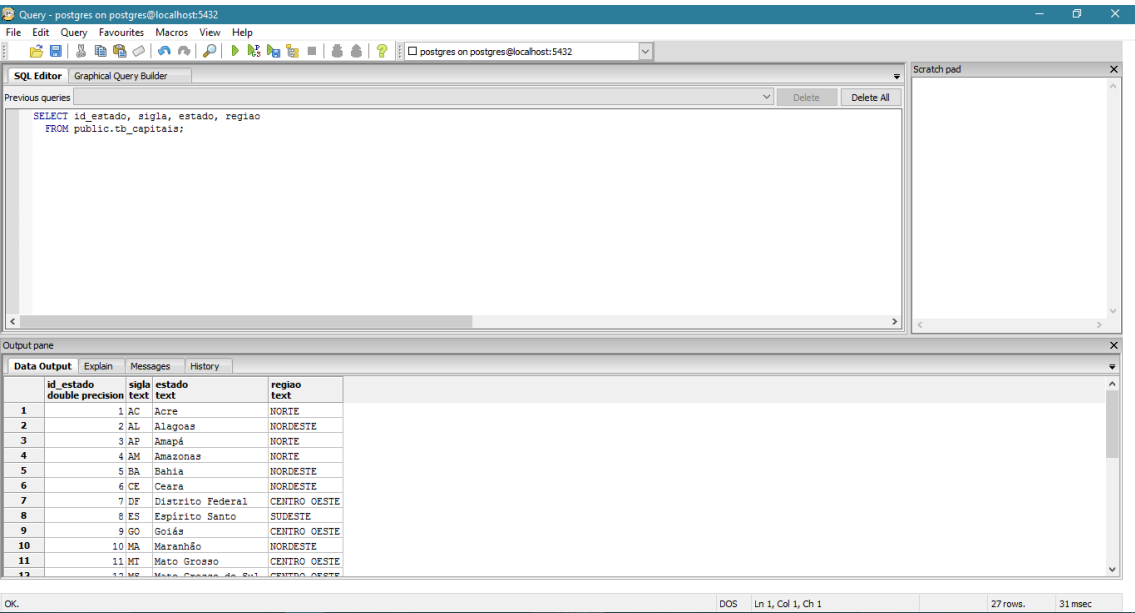


Figura 40