

Aluna: Lays Lopes

Matrícula: 201110005994

Tutorial de criação de um ETL no Pentaho no Windows.

Passo 1: Acesse o site

<https://sourceforge.net/projects/pentaho/files/Data%20Integration/6.1/> e faça o download do **pdi-ce-6.1.0.1-196.zip**. (Figura 1)

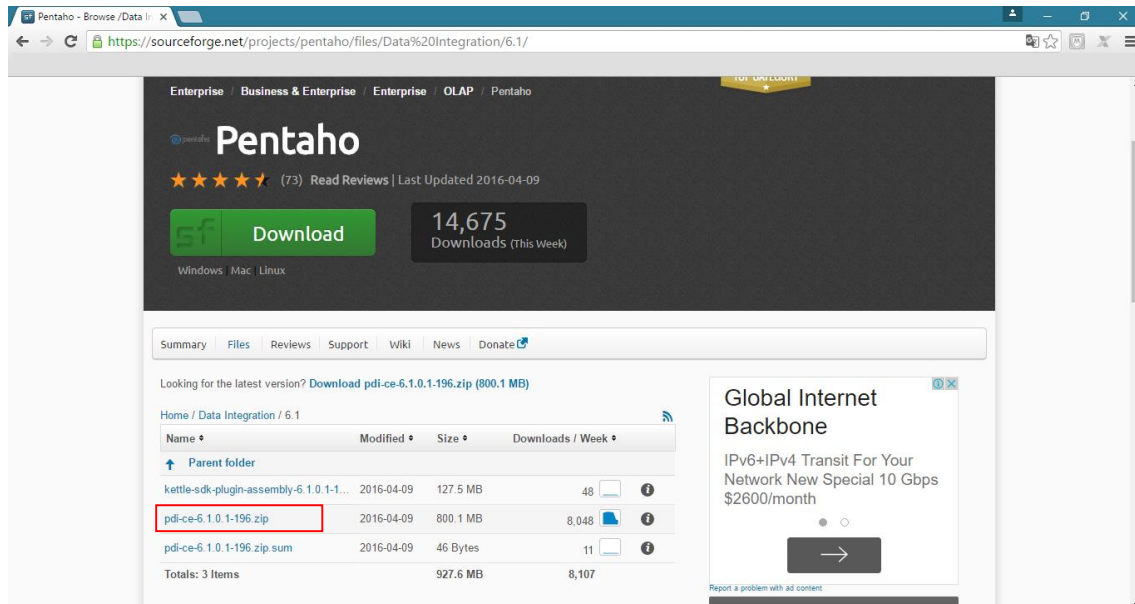


Figura 1 - Site para fazer o download do pentaho data integration

Passo 2: Após o download extraia o arquivo. (Figura 2)

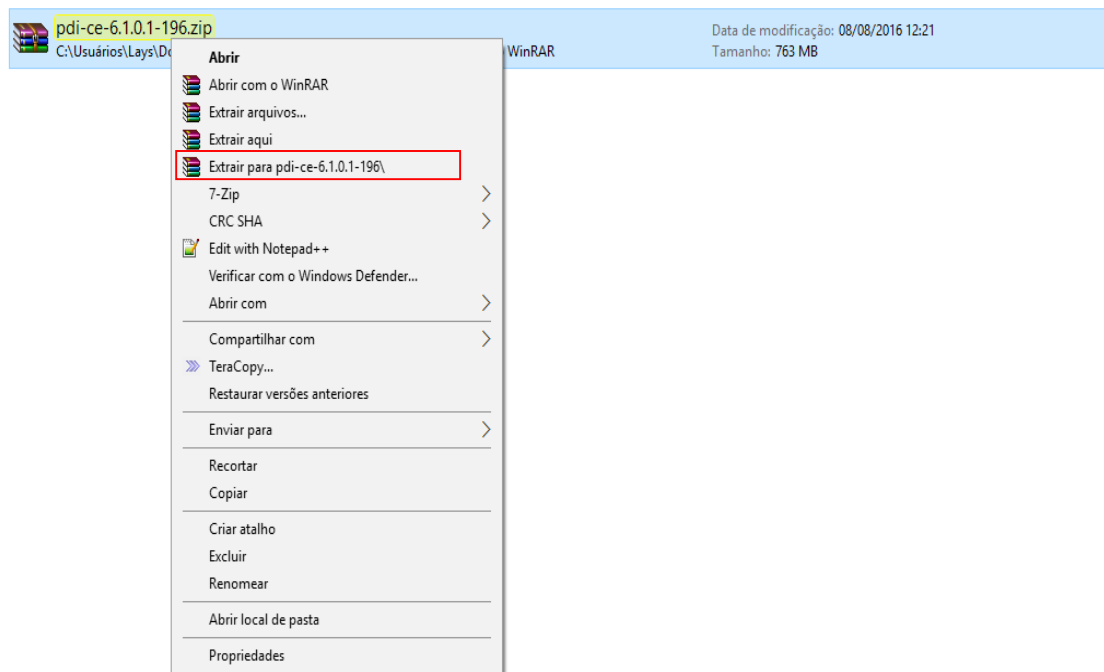


Figura 2 – Extrair zip do pdi

Passo 2: Antes de executar o arquivo start-pentaho.bat. Verifique se possui Sun Java Developer Kit (JDK) e JRE no seu computador. Através do comando no cmd `java -version` e verifique se a versão é 1.7 ou inferior. Caso esteja usando uma versão mais atual, desinstale e coloque a versão 1.7. (Figura 3)

```
C:\> Prompt de Comando
Microsoft Windows [versão 10.0.10586]
(c) 2015 Microsoft Corporation. Todos os direitos reservados.

C:\Users\Lays> java -version
java version "1.7.0_55"
Java(TM) SE Runtime Environment (build 1.7.0_55-b13)
Java HotSpot(TM) 64-Bit Server VM (build 24.55-b03, mixed mode)

C:\Users\Lays>
```

Figura 3 – Versão do java

Passo 3: Configurando as variáveis de ambiente Java para que se possa executar o Pentaho.

As variáveis que serão configuradas são:

- JAVA_HOME
- JRE_HOME
- PATH
- CLASSPATH

Etapa 1 -Vá no painel de controle do computador e clique em “Sistema e Segurança”.
(Figura 4)



Figura 4 – Painel de controle

Etapa 2 - Clique em sistema. (Figura 5)

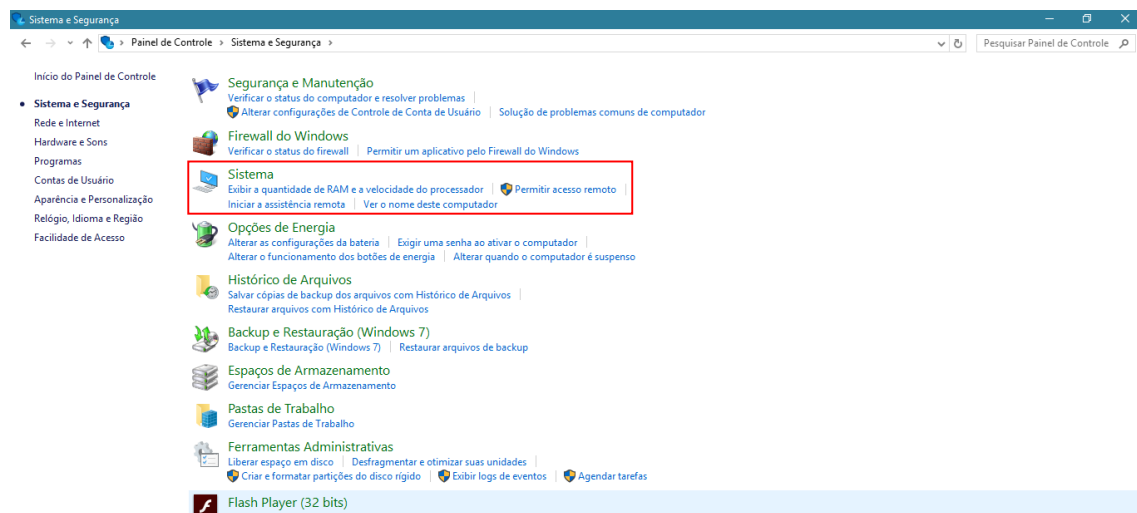


Figura 5 – Sistema e Segurança

Etapa 3 - Clique em “Configurações avançadas do sistema” no canto superior esquerdo. (Figura 6)

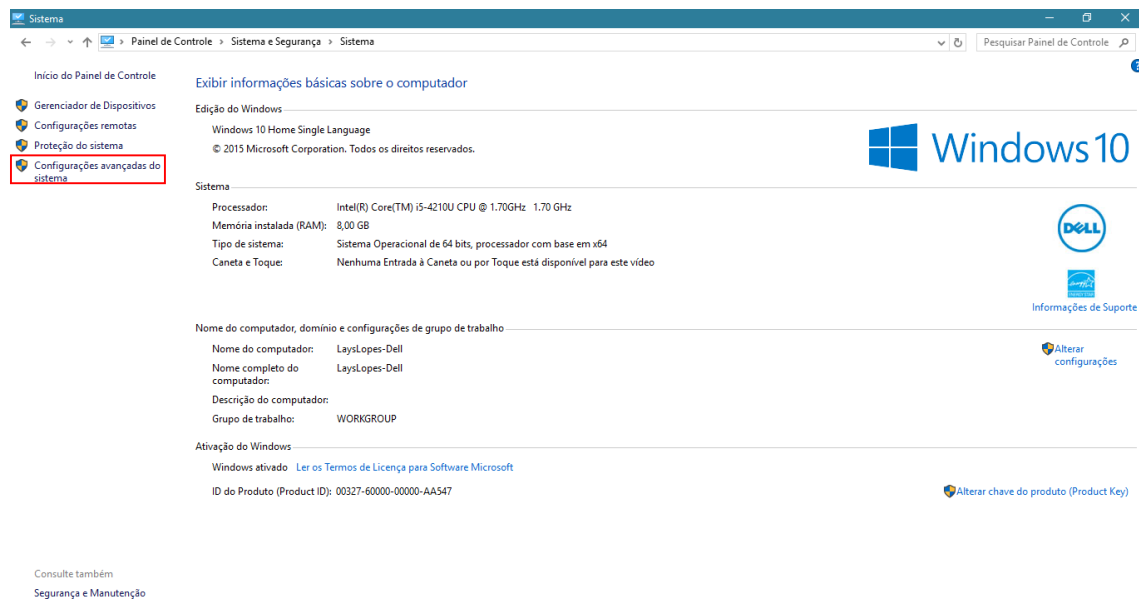


Figura 6 – Sistema

Etapa 4 – Clique em variáveis de ambiente. (Figura 7)

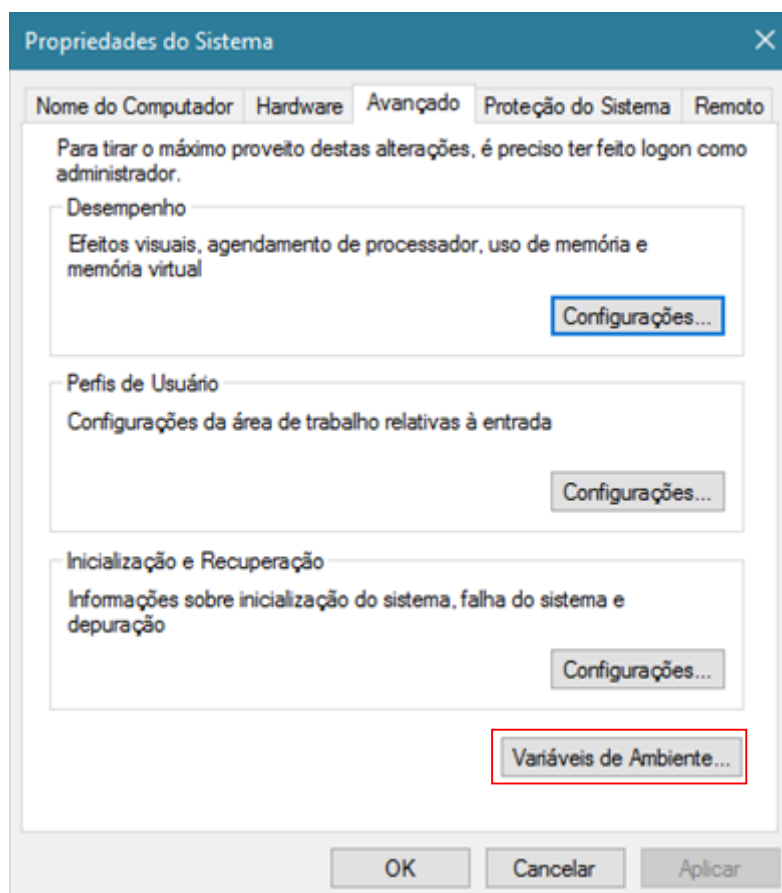


Figura 7 – Configurações avançadas do sistema

Etapa 4 – Configurando as variáveis de usuário (Figura 8)

- JAVA_HOME
- JRE_HOME
- PATH
- CLASSPATH

1. Adicionando a variável JAVA_HOME:

Clique Novo:

- Nome da variável: JAVA_HOME
- Valor da variável: C:\Program Files\Java\jdk1.7.0_55 (Esse local é onde está instalado o seu Java)

2. Adicionando a variável JRE_HOME:

- Nome da variável: JRE_HOME
- Valor da variável: JAVA_HOME

3. Adicionando a variável PATH:

- Nome da variável: PATH
- Valor da variável: C:\Program Files\Java\jdk1.7.0_55\bin

4. Adicionando a variável CLASSPATH:

- Nome da variável: CLASSPATH
- Valor da variável: JAVA_HOME

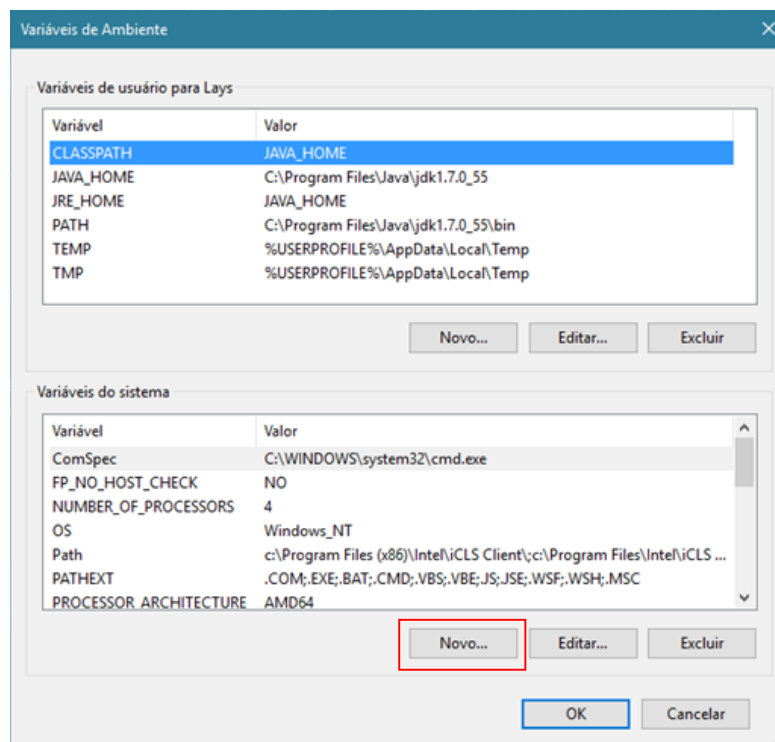


Figura 8 – Configurando as variáveis de ambiente

Passo 4: Para executar o Pentaho Data Integration, abra a pasta descompactada e clique no arquivo Spoon.bat. Feito isso irá aparecer uma mensagem se deseja executar (Figura 9)

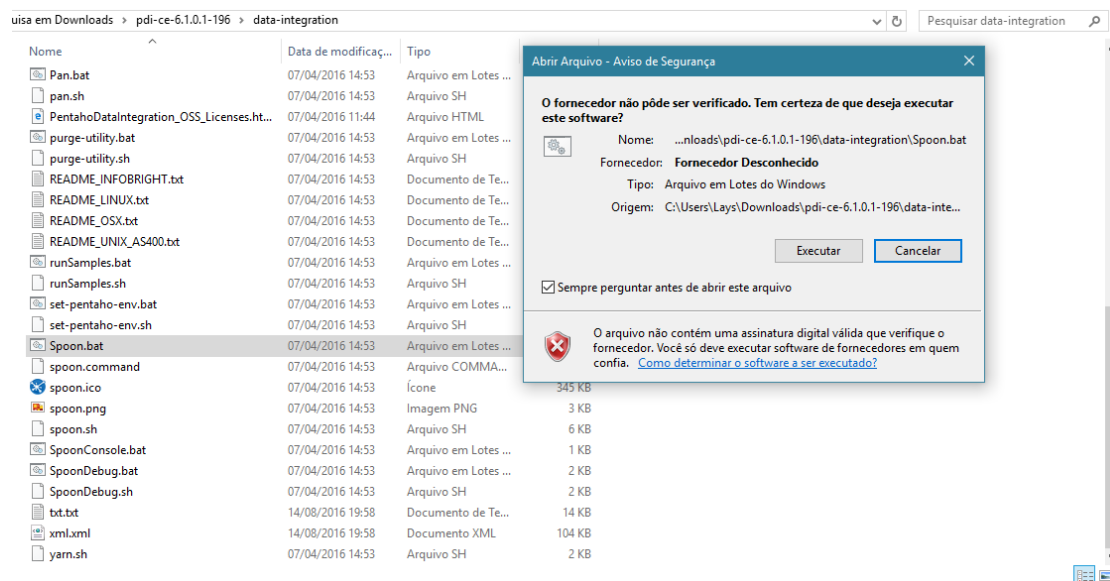


Figura 9 –Executando o spoon.bat

Passo 5. Irá aparecer uma tela carregando o Pentaho Data Intagration (Figura 10)



Figura 10 – PDI

Passo 6. Criando uma transformação. Na tela inicial do Pentaho Data Integration (PDI), clique transformações. (Figura 11)

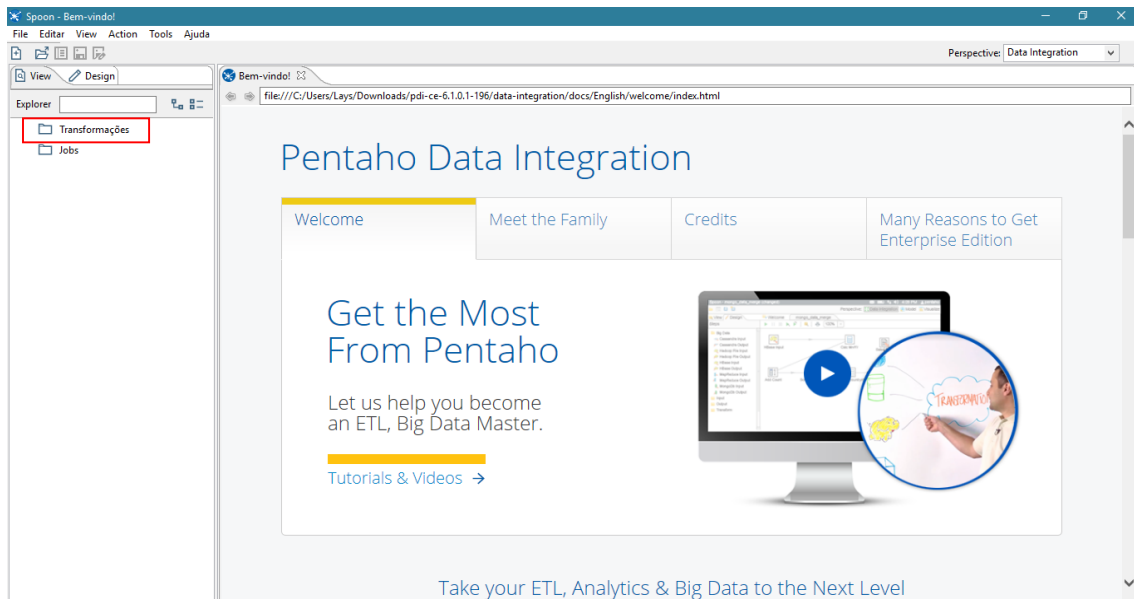


Figura 11 – Tela inicial do PDI

Passo 7. Em transformação, escolha uma entrada, ou seja, qual tipo de dados que deseja transformar. (Figura 12)

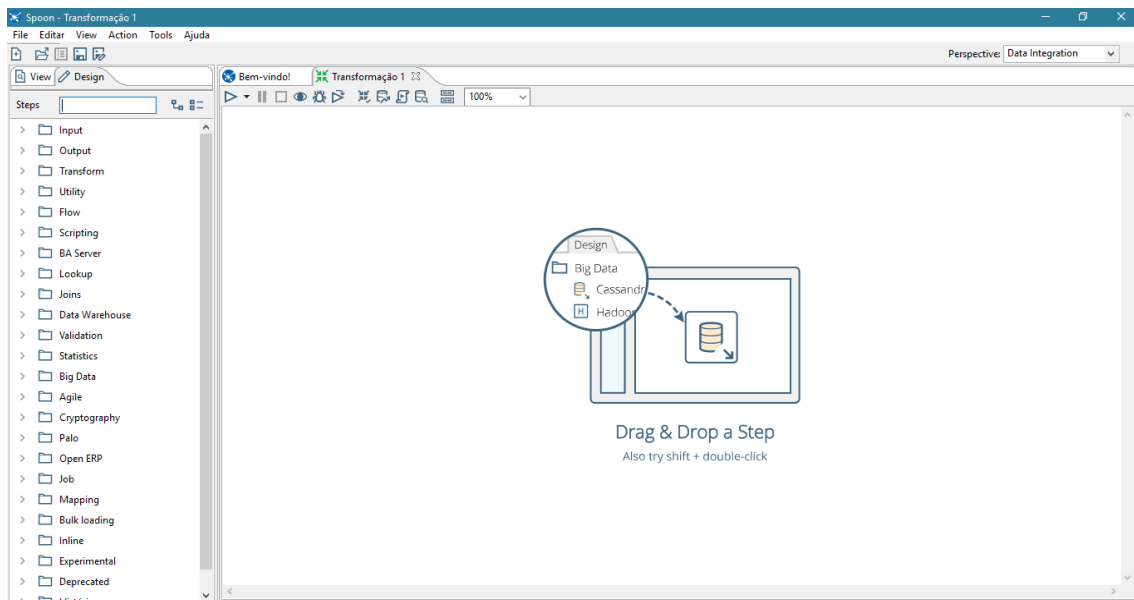


Figura 12 – Criando uma transformação

Passo 8. Como base de dados utilizei o sitio **<http://dados.gov.br/dataset/cidades-digitais>**. E faça o download do arquivo .CSV (Figura 13)



Figura 13 – Base de Dados para ser transformada

Passo 9. Clique em Input e escolha o formato da sua base a ser transformada. O arquivo que irei utilizar é .CSV. Então clique duas vezes no CSV file Input. (Figura 14)

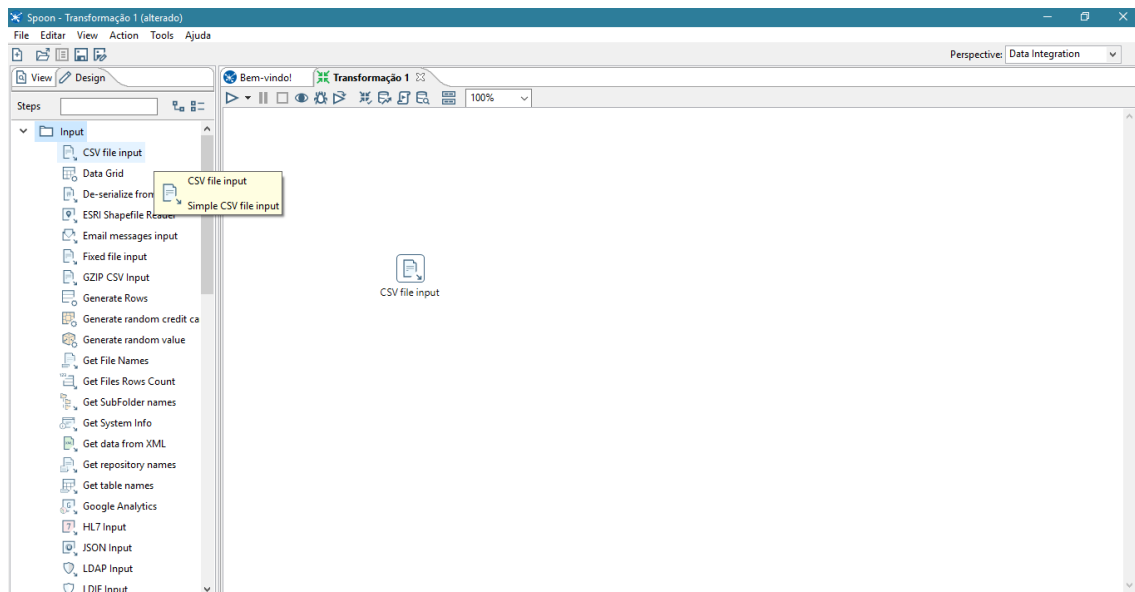


Figura 14 – Escolhendo o formato do arquivo de entrada

Passo 10. Clique em Output e escolha o formato que deseja transformar sua base. Irei transformar em saídas: JSON, XLS, TXT e XML. Arraste a saída JSON Output para área branca e faça isso para as demais saídas que desejar. (Figura 15)

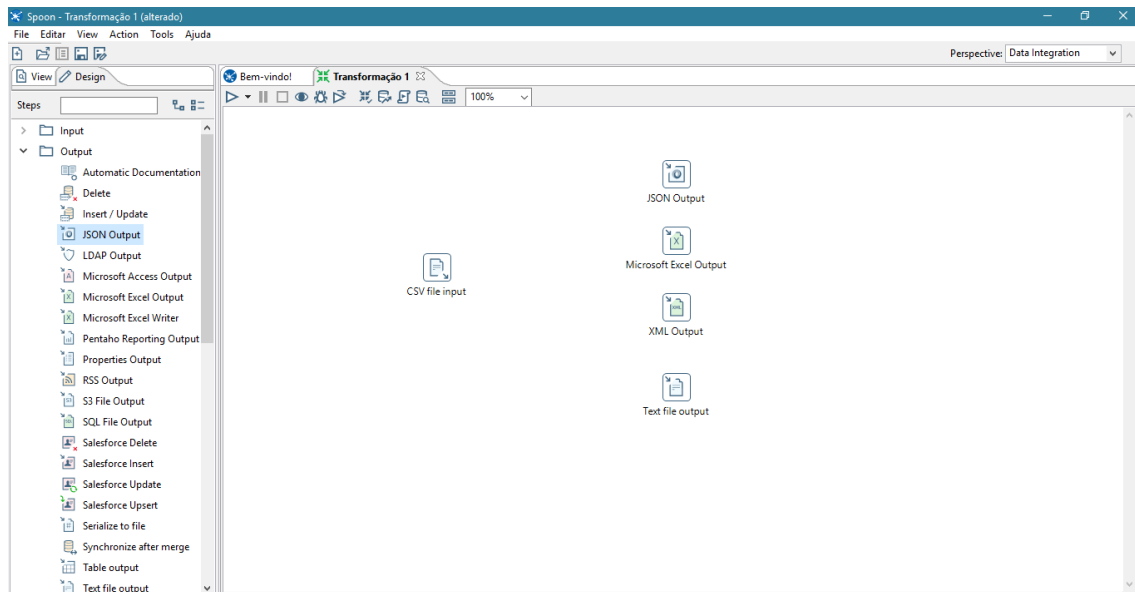



Figura 15 – Escolhendo o formato do arquivo de saída

Passo 11. Coloque o curso do mouse sobre o desenho do CSV file input e clique no símbolo  ao lado da engrenagem para ligar o arquivo de entrada ao de saída. (Figura 16)

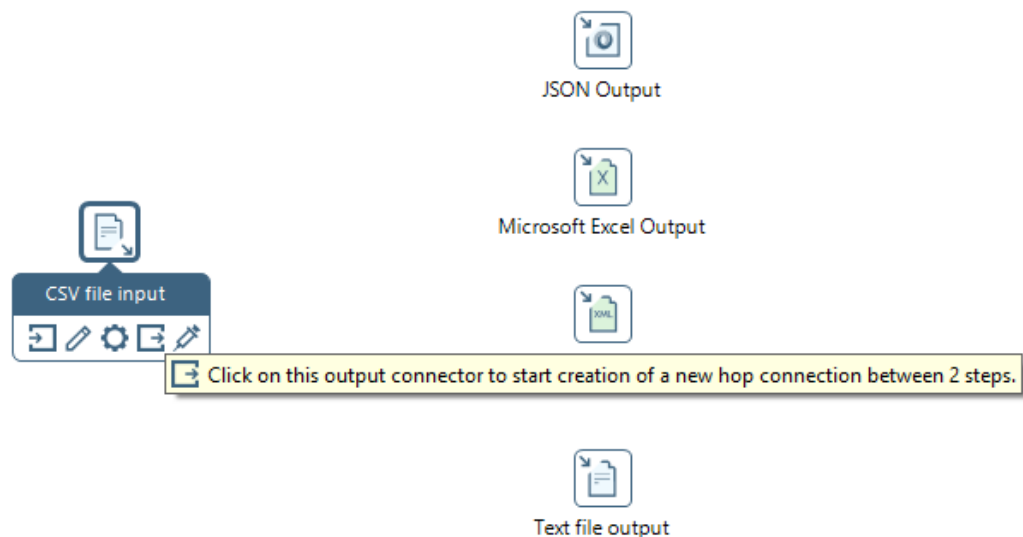


Figura 16 – Criando conexão entre dois steps

Passo 12. Ao aparecer a seta de ligação do CSV file input e leve-o até a saída desejada. Irá aparecer a mensagem perguntando se deseja que aquele step seja o formato de saída, clique em “Main output of step”, para confirmar. (Figura 17)

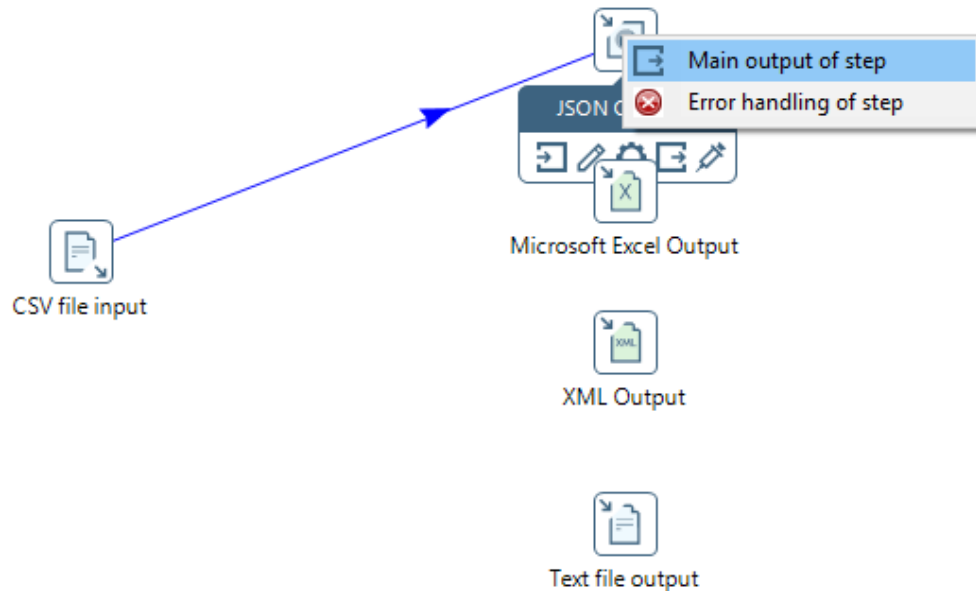


Figura 17 – Ligando a entrada a saída

Passo 13. Como iremos transformar em mais de uma saída ao tentar ligar o CSV file input 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 18)

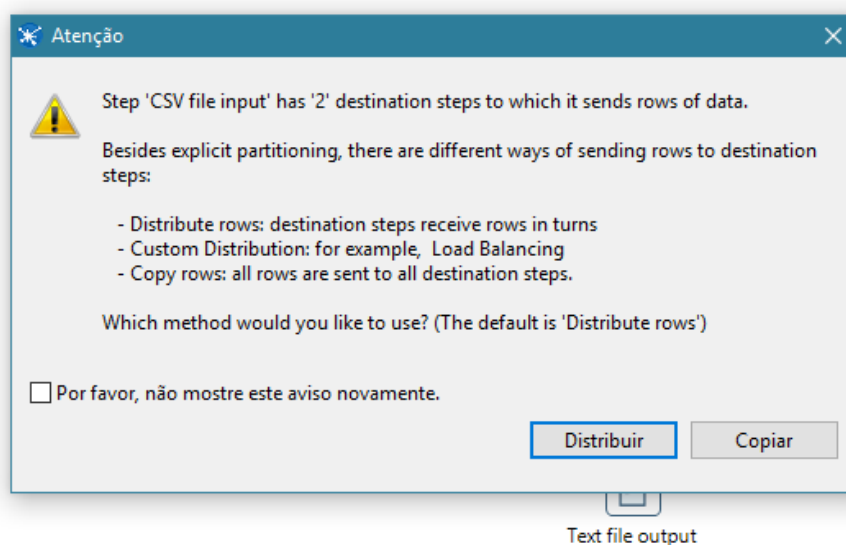


Figura 18 – Mensagem ao ligar entrada a uma segunda saída

Passo 14. Com o cursor mouse do mouse posicionado sobre o CSV file output clique no símbolo da engrenagem e depois em editar step. (Figura 19)

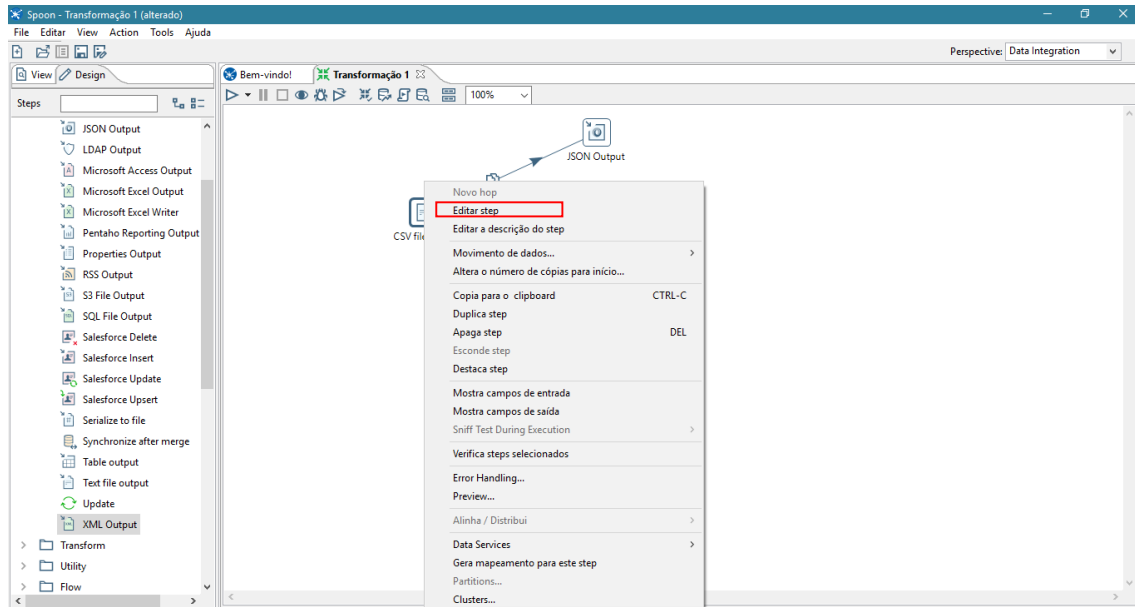


Figura 18 – Editando o step de entrada

Passo 15. Escolha o arquivo .CSV de que utilizará para transformar, observe se os dados dos arquivos são separados por vírgula ou ponto e vírgula e coloque no campo “Delimiter” o símbolo. Feito isso clique em “Obtem campos”. (Figura 19)

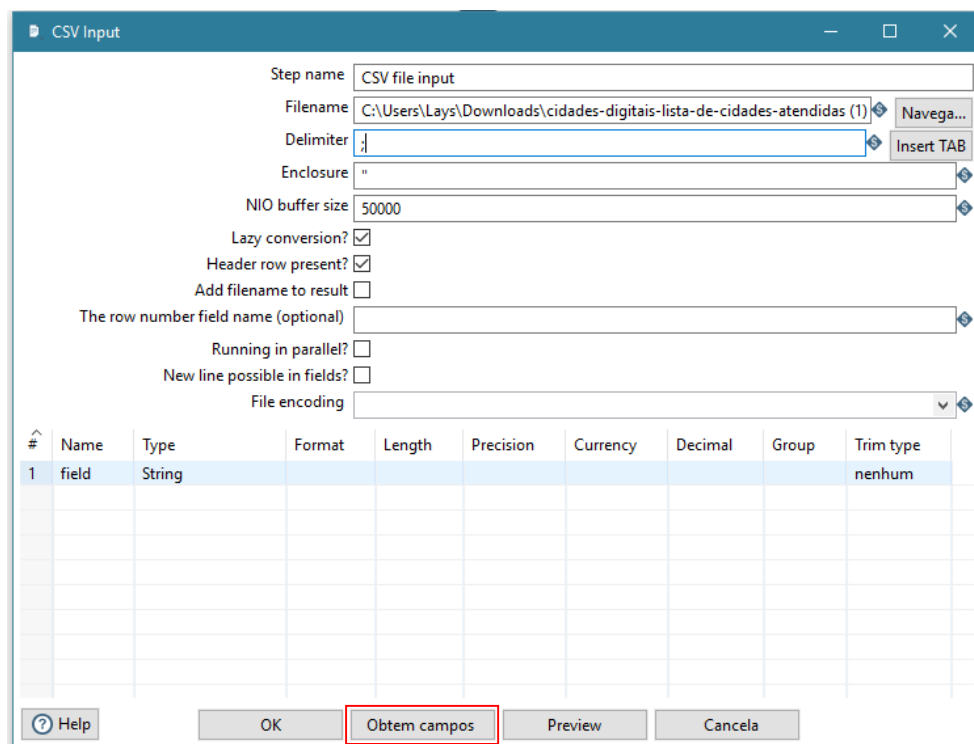


Figura 19 – Importando a base que irá transformar

Passo 16. Ao clicar em obter campos irá aparecer uma mensagem para que você informe quantas linhas dos dados deseja obter. Como meu arquivo possui 339 eu coloquei 400. Clique em “OK”. (Figura 20)

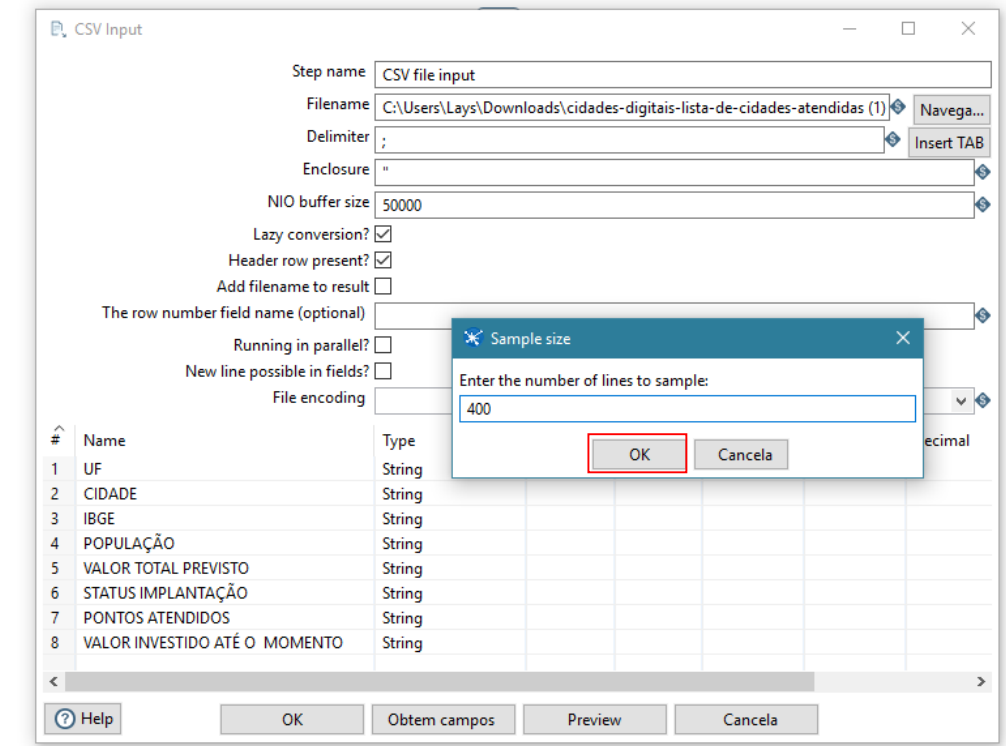


Figura 20 – Informando o número de linhas

Passo 17. Irá aparecer o resultado dos campos obtidos e o número de linhas obtidas. Verifique se importou os campos e a quantidade de linhas. Clique em Fechar. (Figura 21)

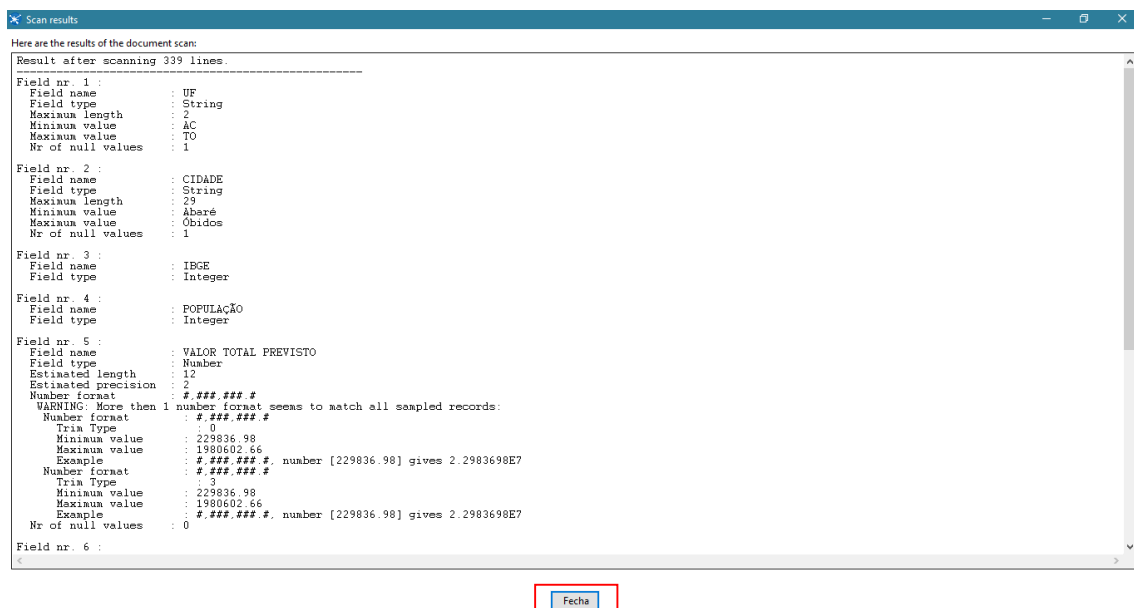


Figura 21 – Resultado obtido

Passo 18. Clique em Ok. Para finalizar a configuração do CSV file input. (Figura 22)

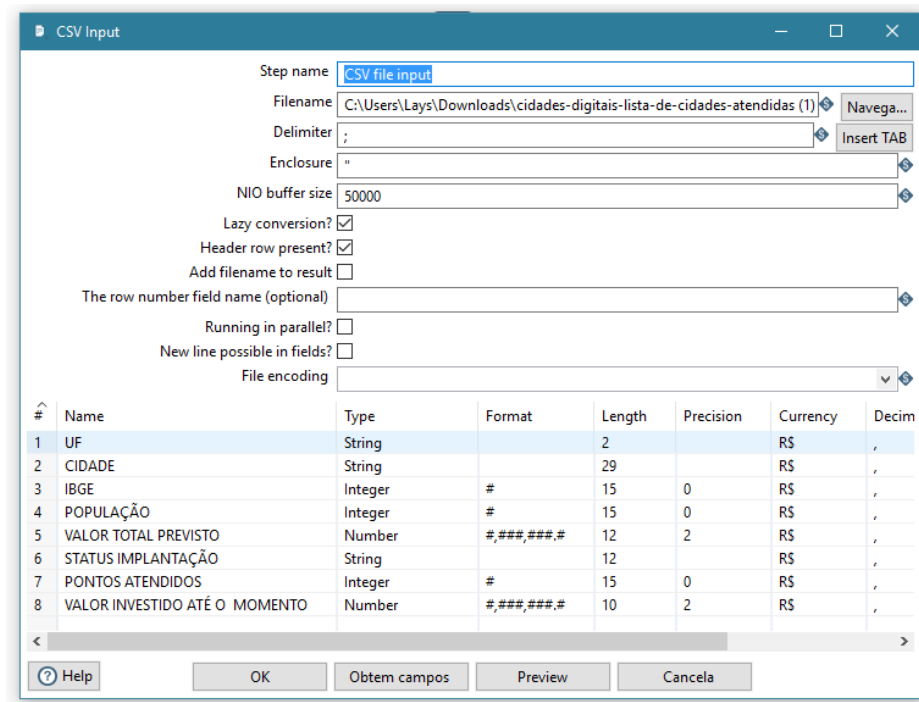


Figura 22 – Finalização da configuração do CSV file input

Passo 19. Clique com o botão direito do mouse no step de saída para configurar o nome do arquivo que será gerado. Em Output 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 23)

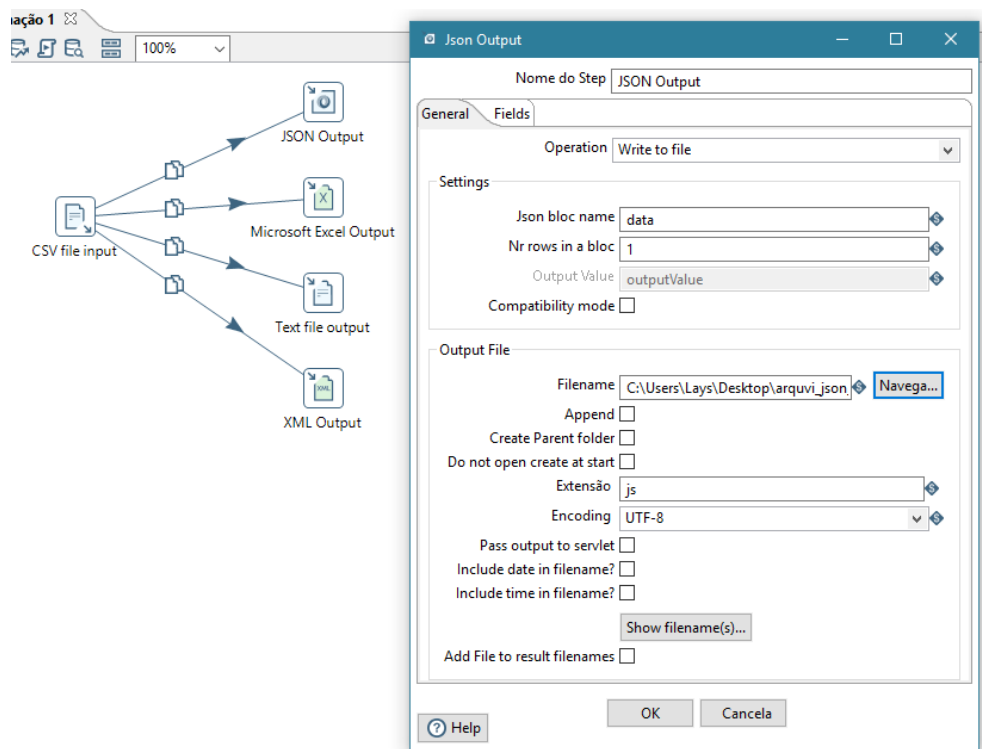


Figura 23 – Configuração do nome do arquivo de saída

Passo 20. Clique com o botão  para executar a transformação. (Figura 24)

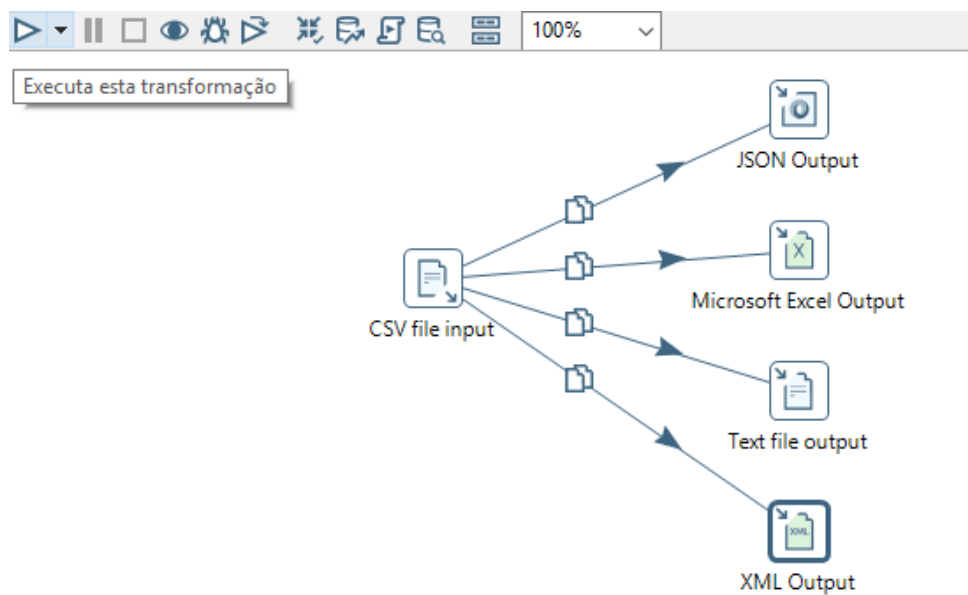
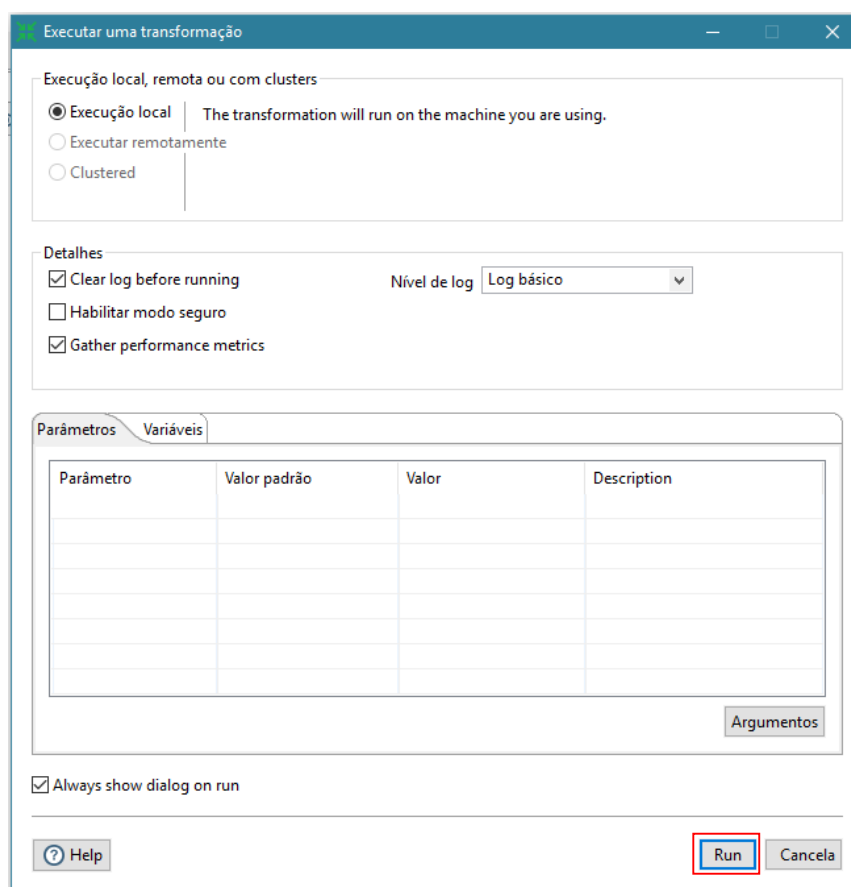


Figura 24 – Executando a transformação

Passo 21. Escolha o local de execução e clique em “Run”. (Figura 25)



The image shows a dialog box titled "Executar uma transformação". It has two tabs: "Parâmetros" and "Variáveis". The "Parâmetros" tab is active. It contains the following sections:

- Execução local, remota ou com clusters:** Three radio buttons are present: "Execução local" (selected), "Executar remotamente", and "Clustered". A note next to "Execução local" says "The transformation will run on the machine you are using."
- Detalhes:** A section with three checkboxes: "Clear log before running" (checked), "Habilitar modo seguro" (unchecked), and "Gather performance metrics" (checked). To the right is a "Nível de log" dropdown menu set to "Log básico".
- Parâmetros:** A table with four columns: "Parâmetro", "Valor padrão", "Valor", and "Description". The table is currently empty.
- Argumentos:** A button labeled "Argumentos" is located at the bottom right of the table area.
- Always show dialog on run:** A checkbox that is checked.
- Buttons:** At the bottom, there are three buttons: "Help" (with a question mark icon), "Run" (highlighted with a red box), and "Cancela".

Figura 25 – Local de execução

Passo 22. Resultado da execução. Como todos ficaram com o símbolo verde então a transformação ocorreu com sucesso e no resultado da execução podemos verificar a saída se o número de linhas corresponde. (Figura 26)

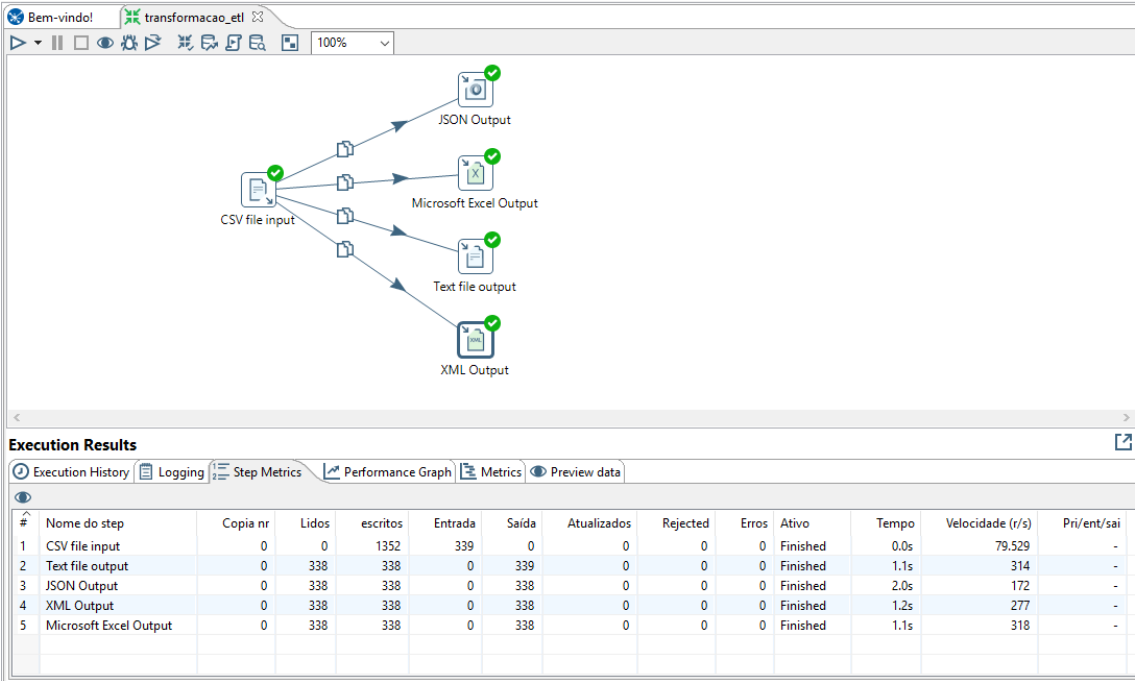


Figura 26 – Resultado da execução

Passo 22. Arquivos gerados como saída. (Figura 27)

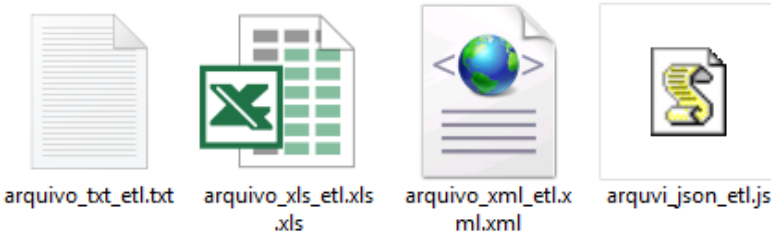


Figura 26 – Arquivos gerados como saída