Information Visualization

Daniel Gonçalves, Sandra Gama

03

Pentaho Data Integration

Tutorial

01

PENTAHO DATA INTEGRATION:

INTRODUCTION

PDI - Pentaho Data Integration (Kettle)

“Delivers powerful

Extraction,

Transformation

Loading

capabilities, using a groundbreaking, matadata-driven approach.”

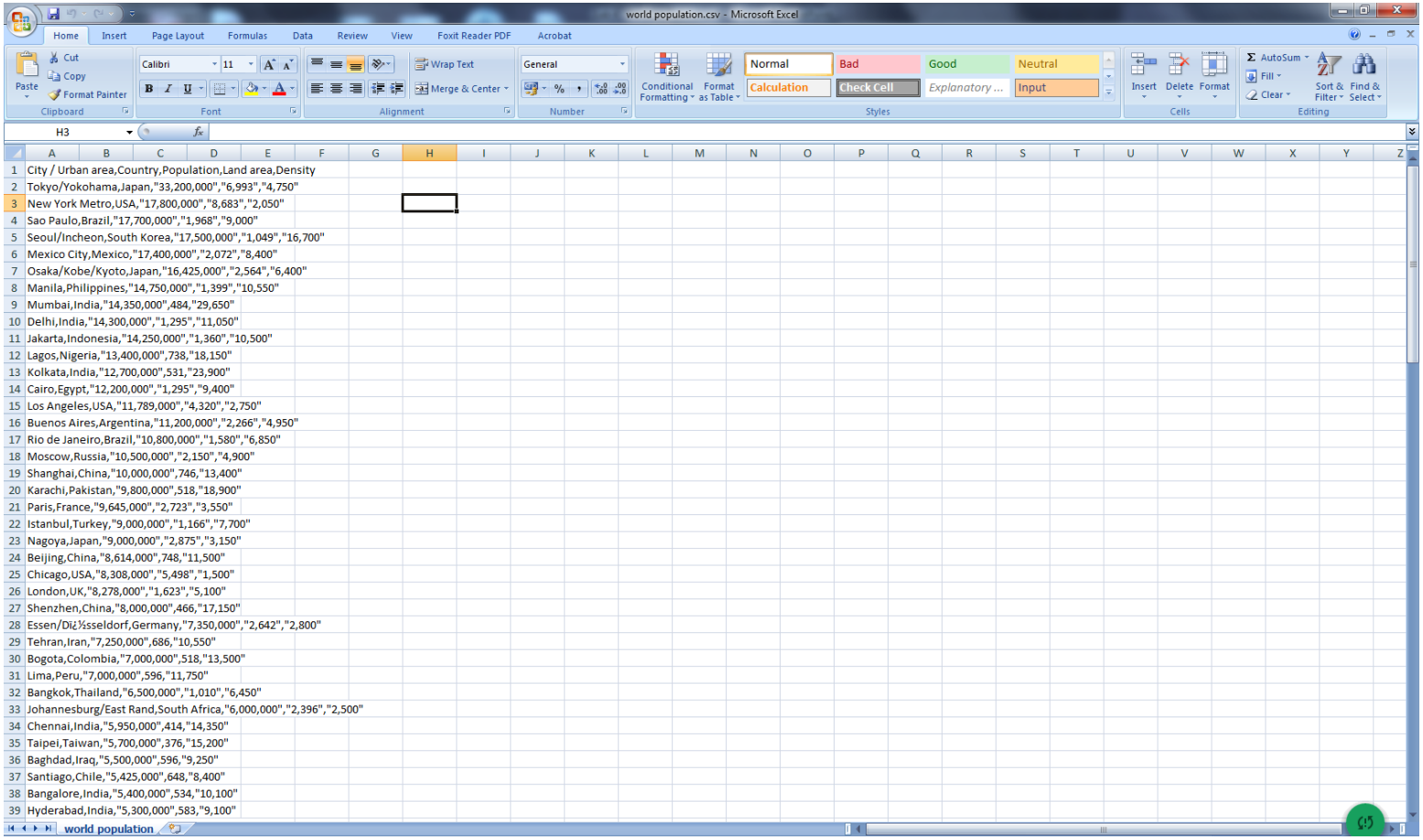
Starting point: raw data file

Download the data file

(course page > labs' section)

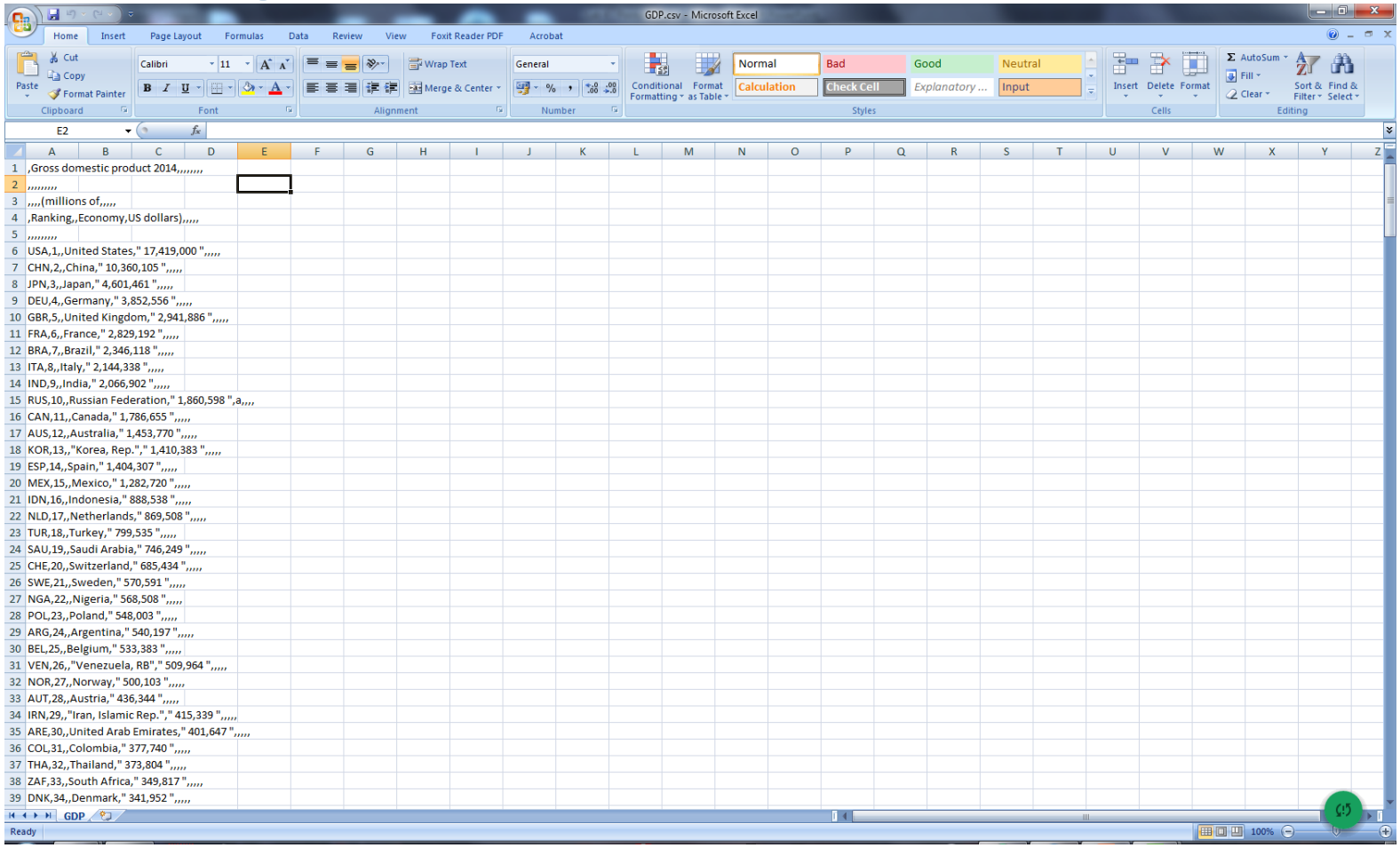
Check your data (two files)!

Open with Excel: world population.csv



Check your data (two files)!

Open with Excel: GDP.csv

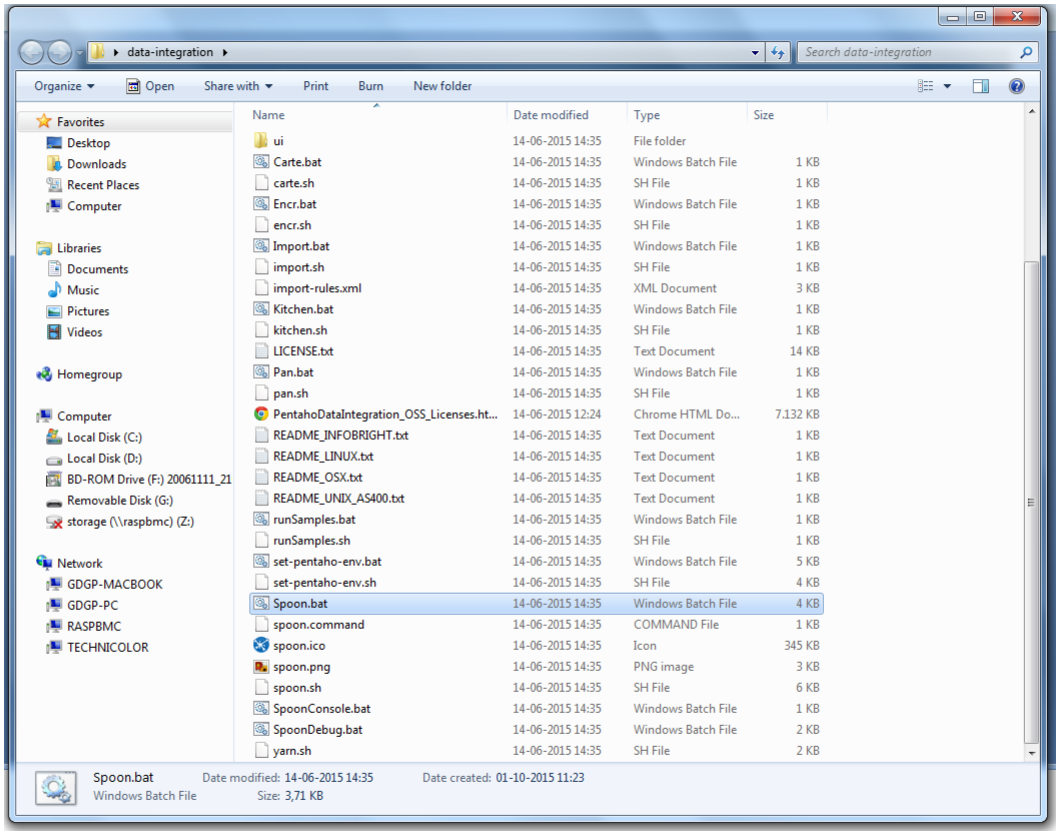


Let's transform it with Pentaho Data Integration

Download

http://community.pentaho.com/projects/data-integration/

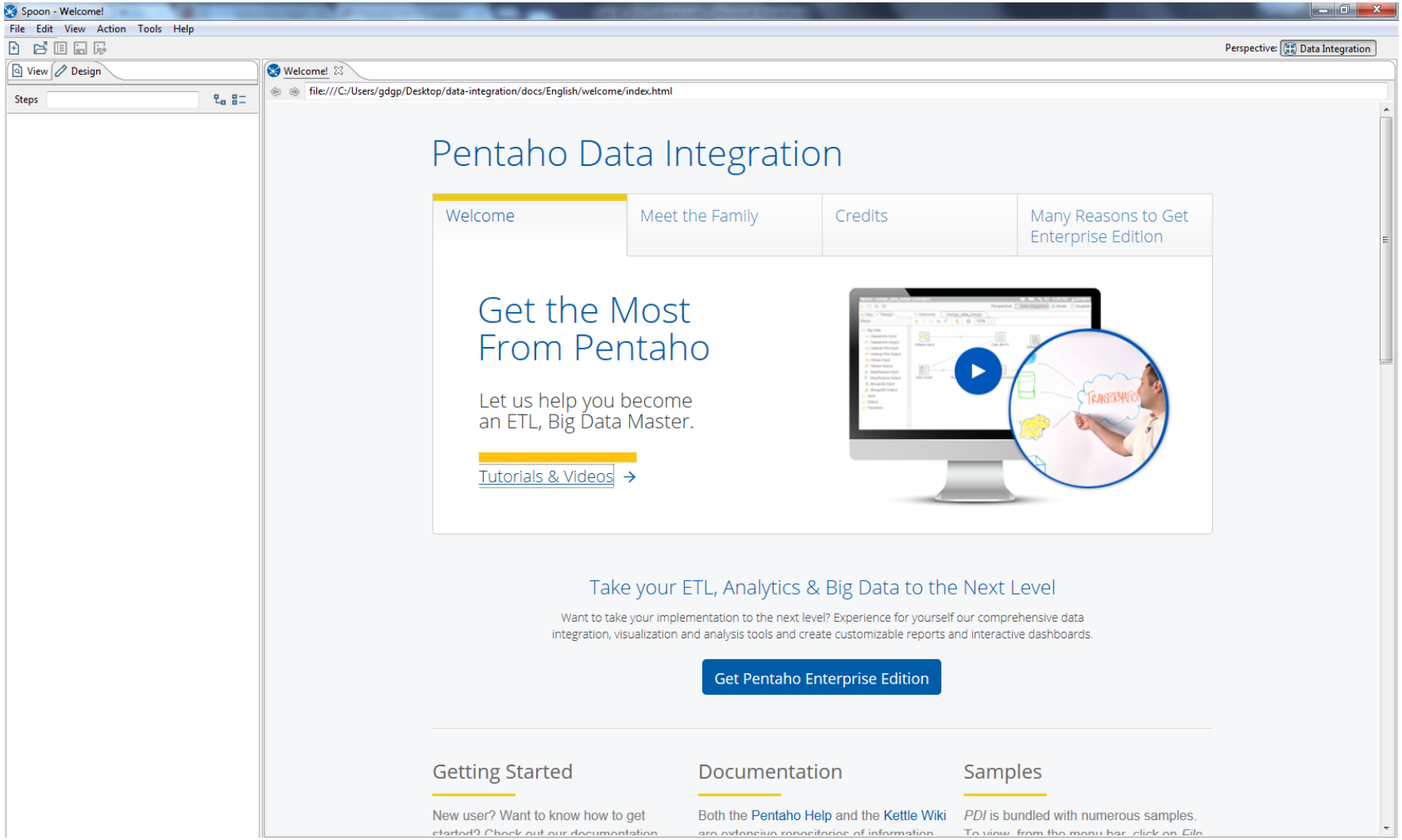
Unzip (anywhere) & run!



Run

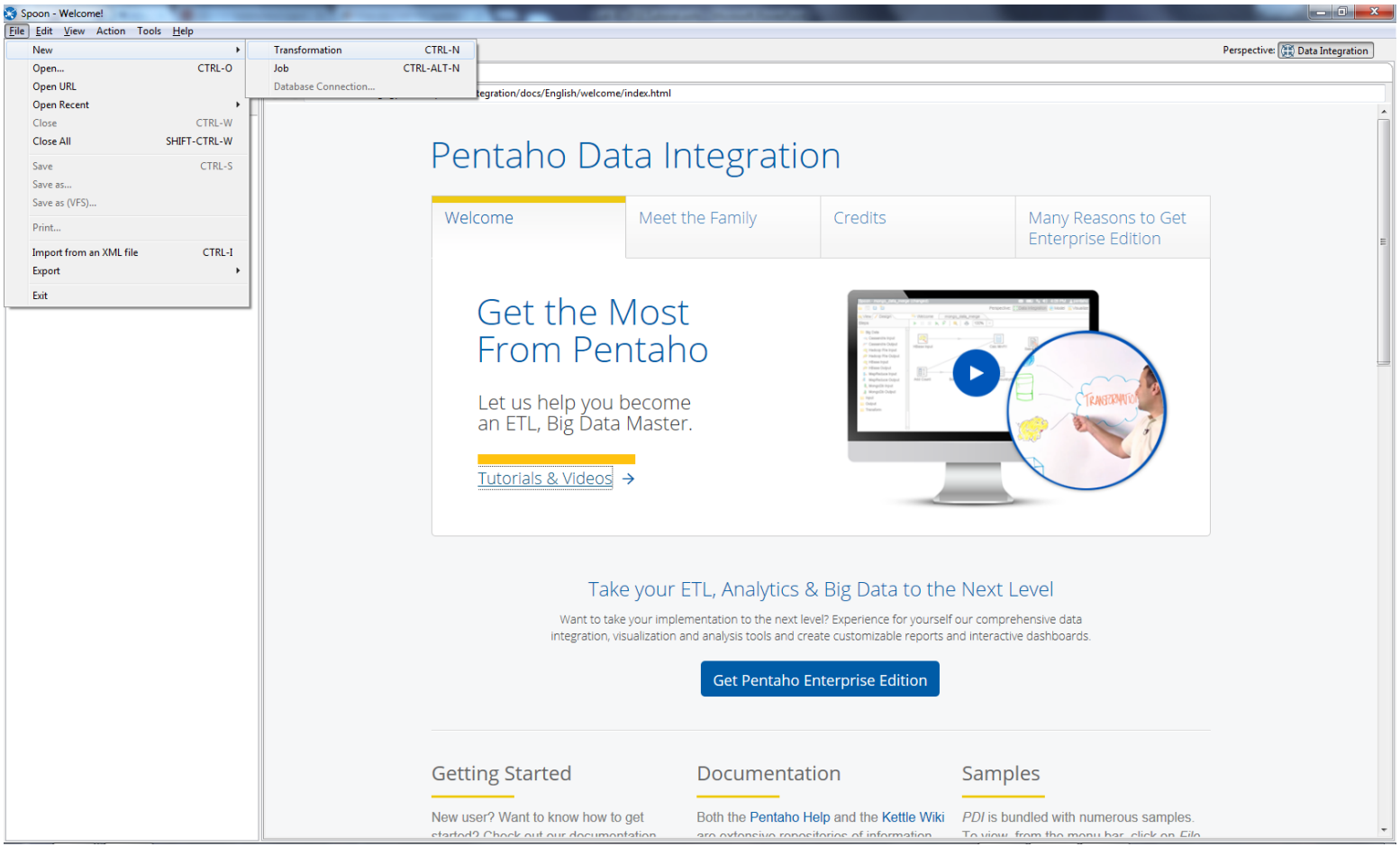
Spoon.bat

PDI – Spoon Interface

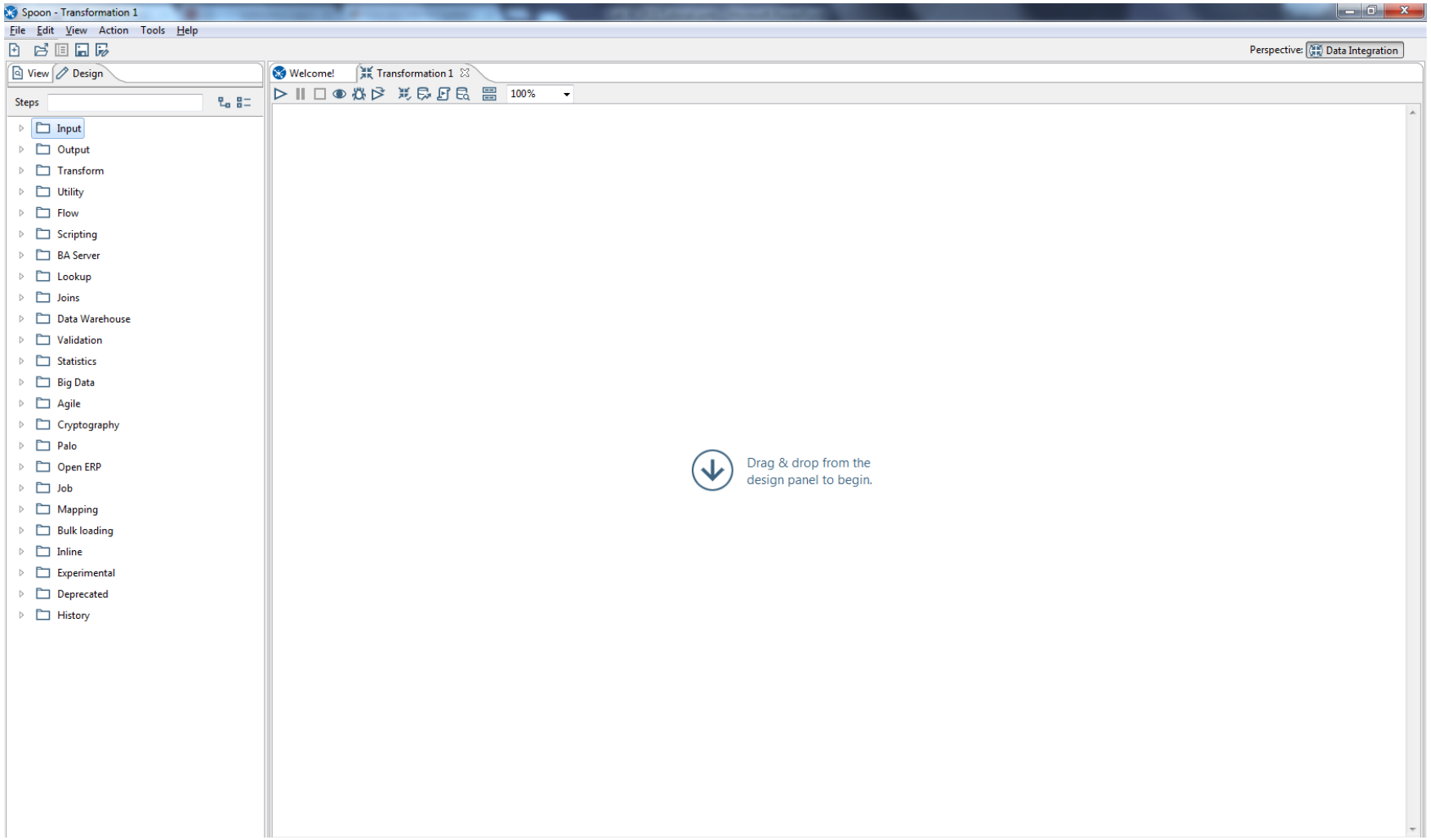


Create new Data Transformation

File > New > Transformation

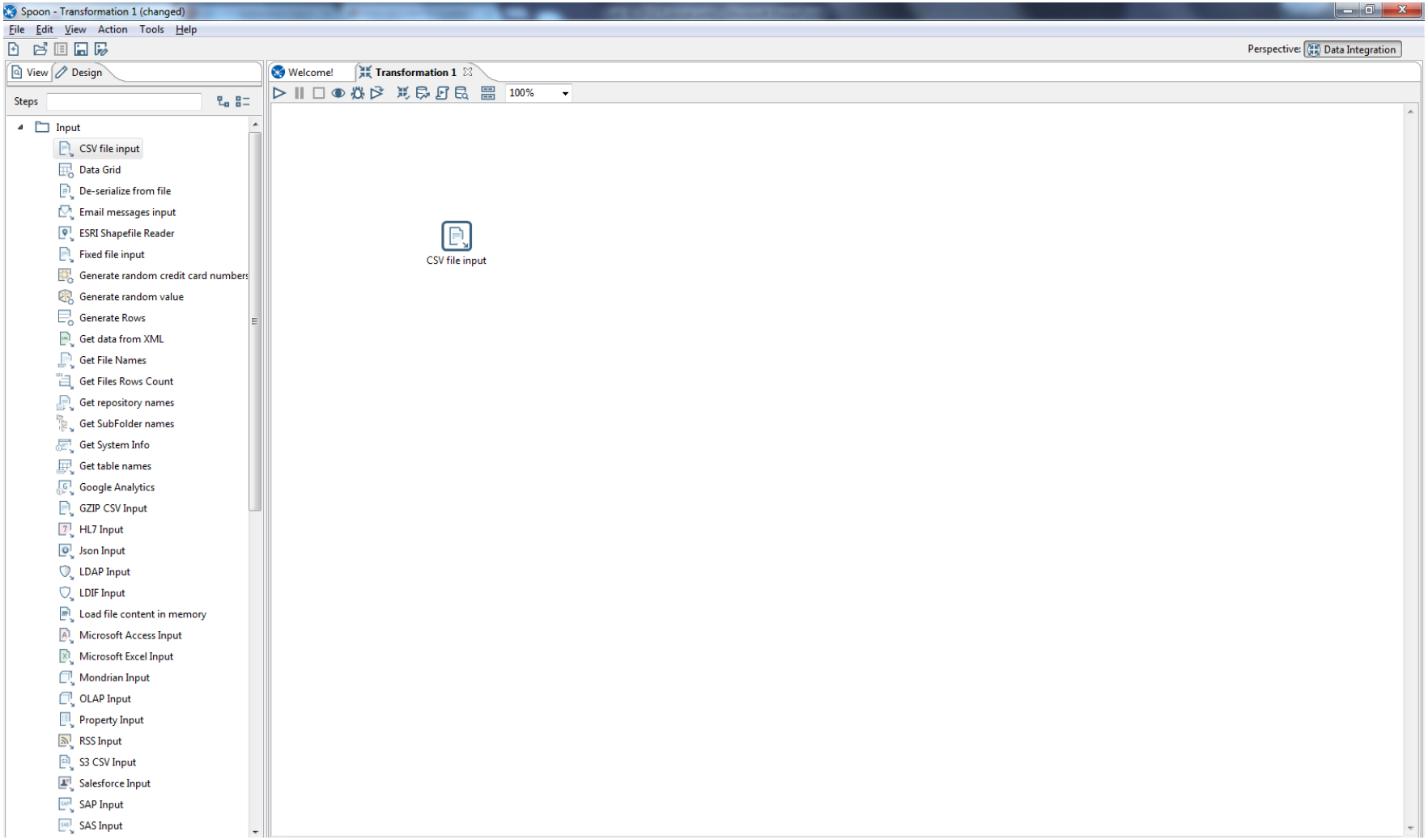


New transformation created



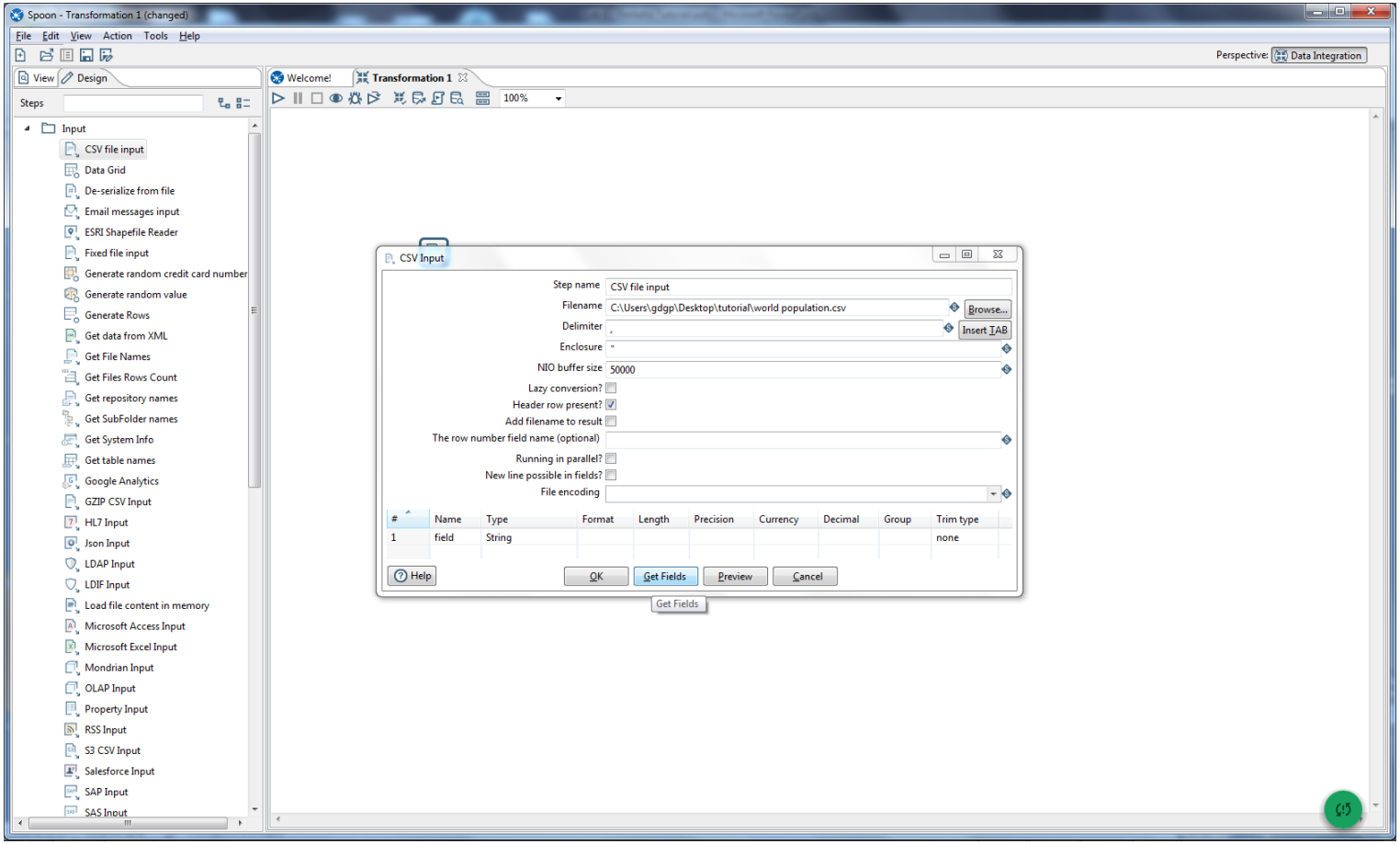
Now let's use our data file as input.

Input > CSV File input > drag & drop

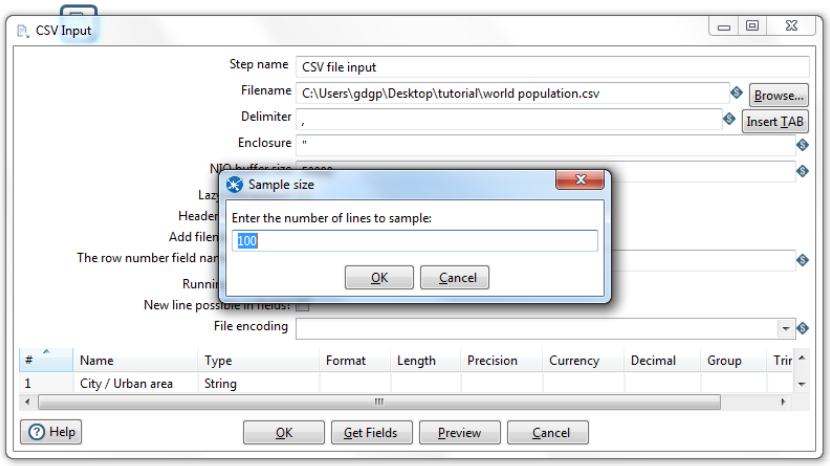


Right click to adjust properties

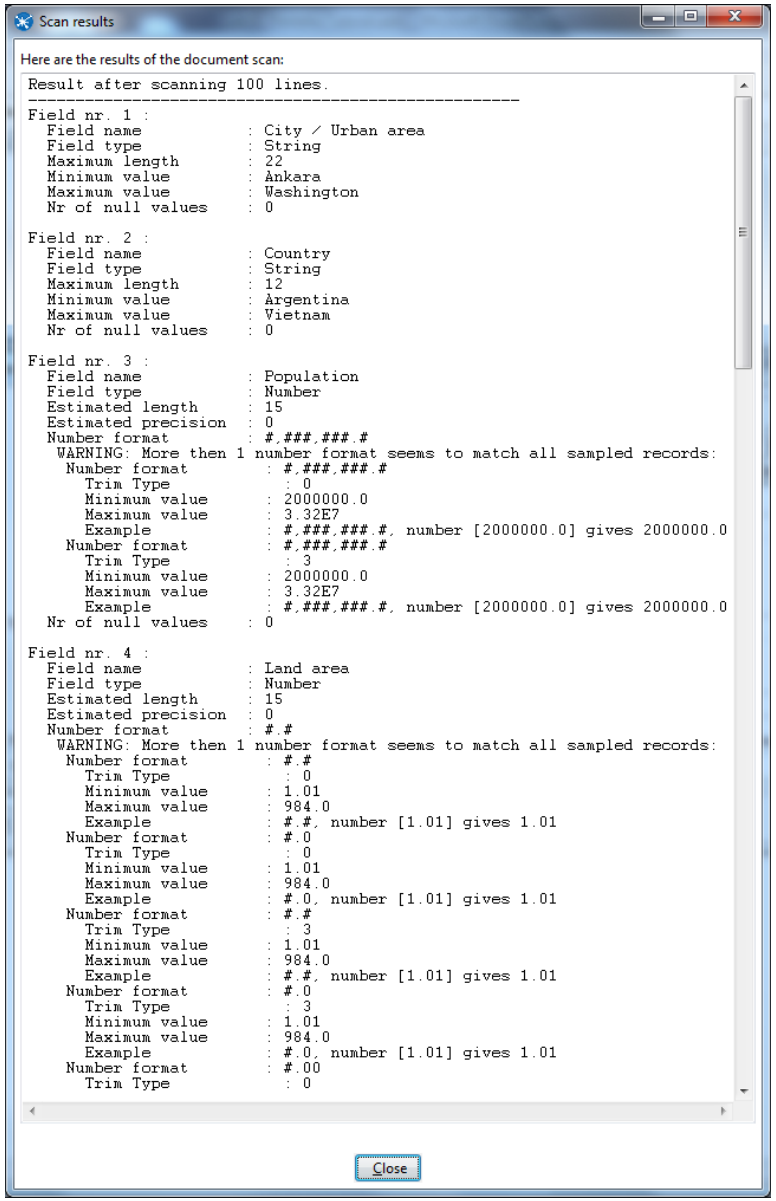
Define step name and filename (browse your data file)



Get fields…

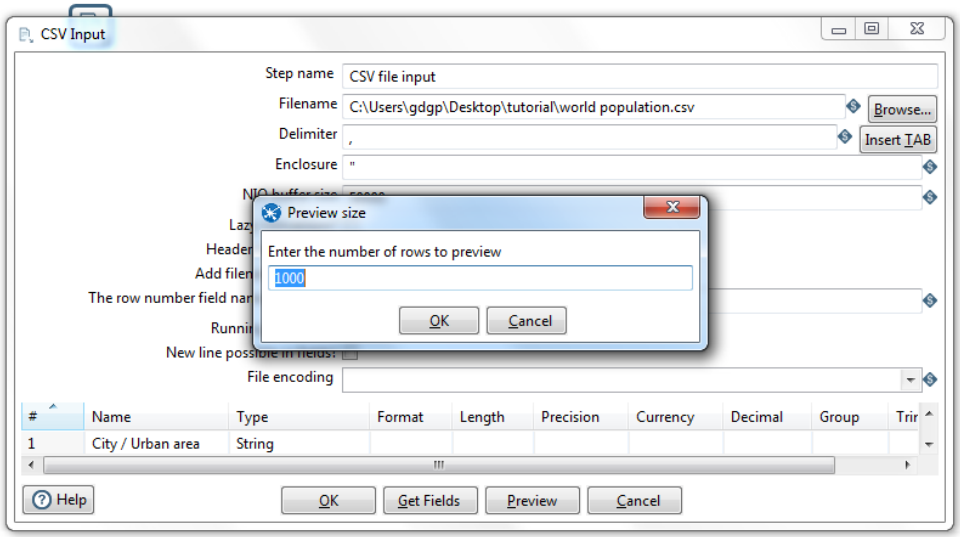


>



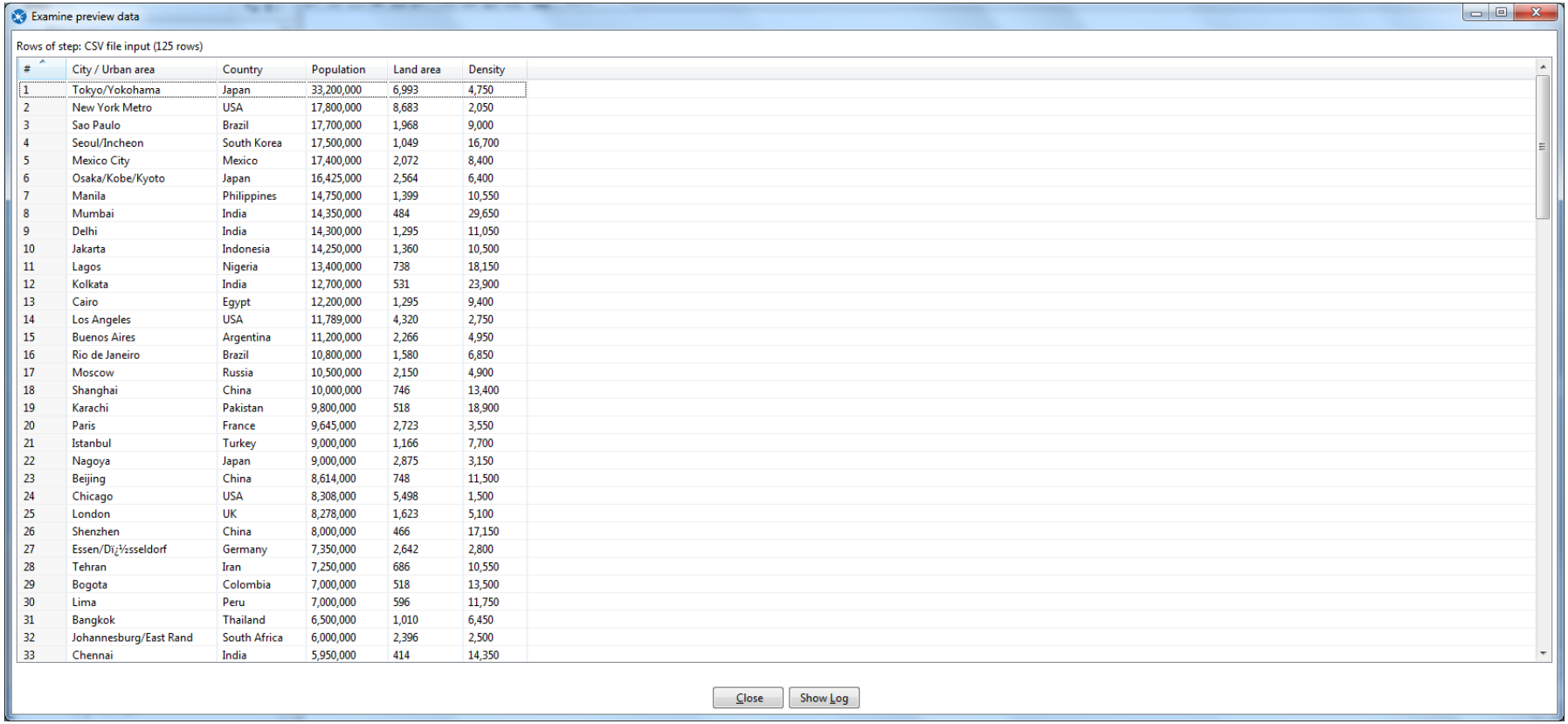
Now we have all the fields

Preview…


CSV Input
Step name
CSV file input
Filename
C:\Users\gdgp\Desktop\tutorial\world population.csv
Browse...
Delimiter
,
Insert TAB
Enclosure
"
Preview size
Enter the number of rows to preview
1000
OK
Cancel
File encoding
Name Type Format Length Precision Currency Decimal Group
1 City / Urban area String
Help
OK
Get Fields
Preview
Cancel

Preview…

There you go!

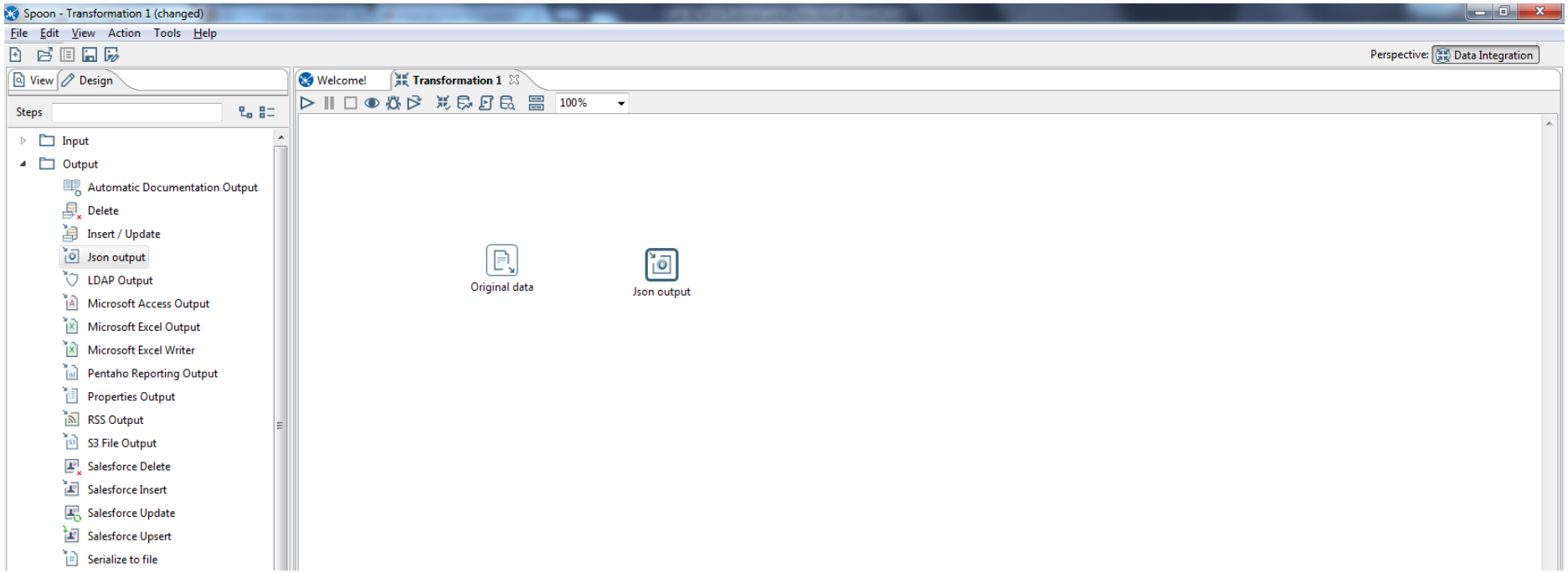


#	City / Urban area	Country	Population	Land area	Density
1	Tokyo/Yokohama	Japan	33,200,000	6,993	4,750
2	New York Metro	USA	17,800,000	8,683	2,050
3	Sao Paulo	Brazil	17,700,000	1,968	9,000
4	Seoul/Incheon	South Korea	17,500,000	1,049	16,700
5	Mexico City	Mexico	17,400,000	2,072	8,400
6	Osaka/Kobe/Kyoto	Japan	16,425,000	2,564	6,400
7	Manila	Philippines	14,750,000	1,399	10,550
8	Mumbai	India	14,350,000	484	29,650
9	Delhi	India	14,300,000	1,295	11,050
10	Jakarta	Indonesia	14,250,000	1,360	10,500
11	Lagos	Nigeria	13,400,000	738	18,150
12	Kolkata	India	12,700,000	531	23,900
13	Cairo	Egypt	12,200,000	1,295	9,400
14	Los Angeles	USA	11,789,000	4,320	2,750
15	Buenos Aires	Argentina	11,200,000	2,266	4,950
16	Rio de Janeiro	Brazil	10,800,000	1,580	6,850
17	Moscow	Russia	10,500,000	2,150	4,900
18	Shanghai	China	10,000,000	746	13,400
19	Karachi	Pakistan	9,800,000	518	18,900
20	Paris	France	9,645,000	2,723	3,550
21	Istanbul	Turkey	9,000,000	1,166	7,700
22	Nagoya	Japan	9,000,000	2,875	3,150
23	Beijing	China	8,614,000	748	11,500
24	Chicago	USA	8,308,000	5,498	1,500
25	London	UK	8,278,000	1,623	5,100
26	Shenzhen	China	8,000,000	466	17,150
27	Essen/Dï¿½sseldorf	Germany	7,350,000	2,642	2,800
28	Tehran	Iran	7,250,000	686	10,550
29	Bogota	Colombia	7,000,000	518	13,500
30	Lima	Peru	7,000,000	596	11,750
31	Bangkok	Thailand	6,500,000	1,010	6,450
32	Johannesburg/East Rand	South Africa	6,000,000	2,396	2,500
33	Chennai	India	5,950,000	414	14,350

What to do with this data?

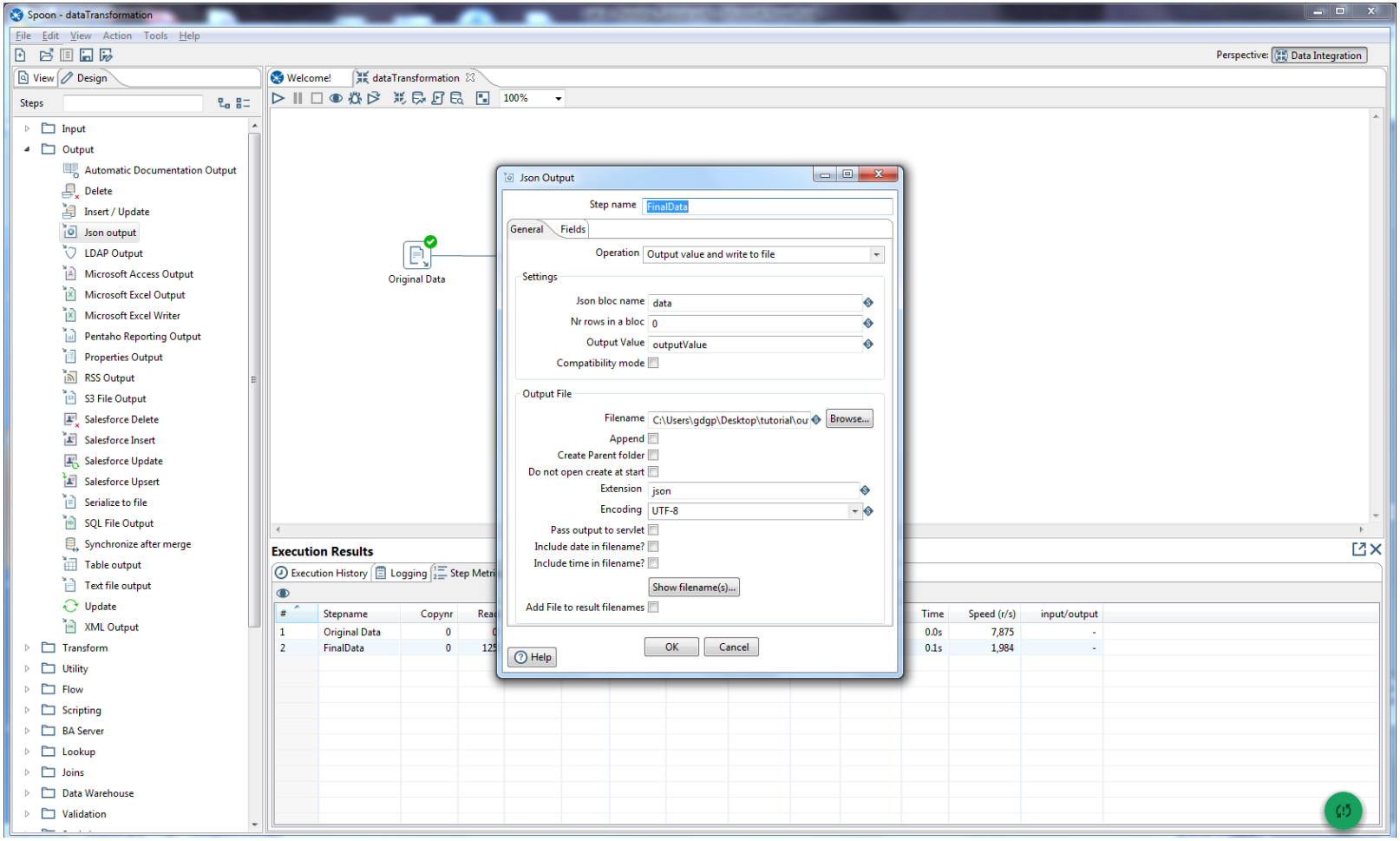
Let’s write it to a .json file

Output > Json output > drag&drop



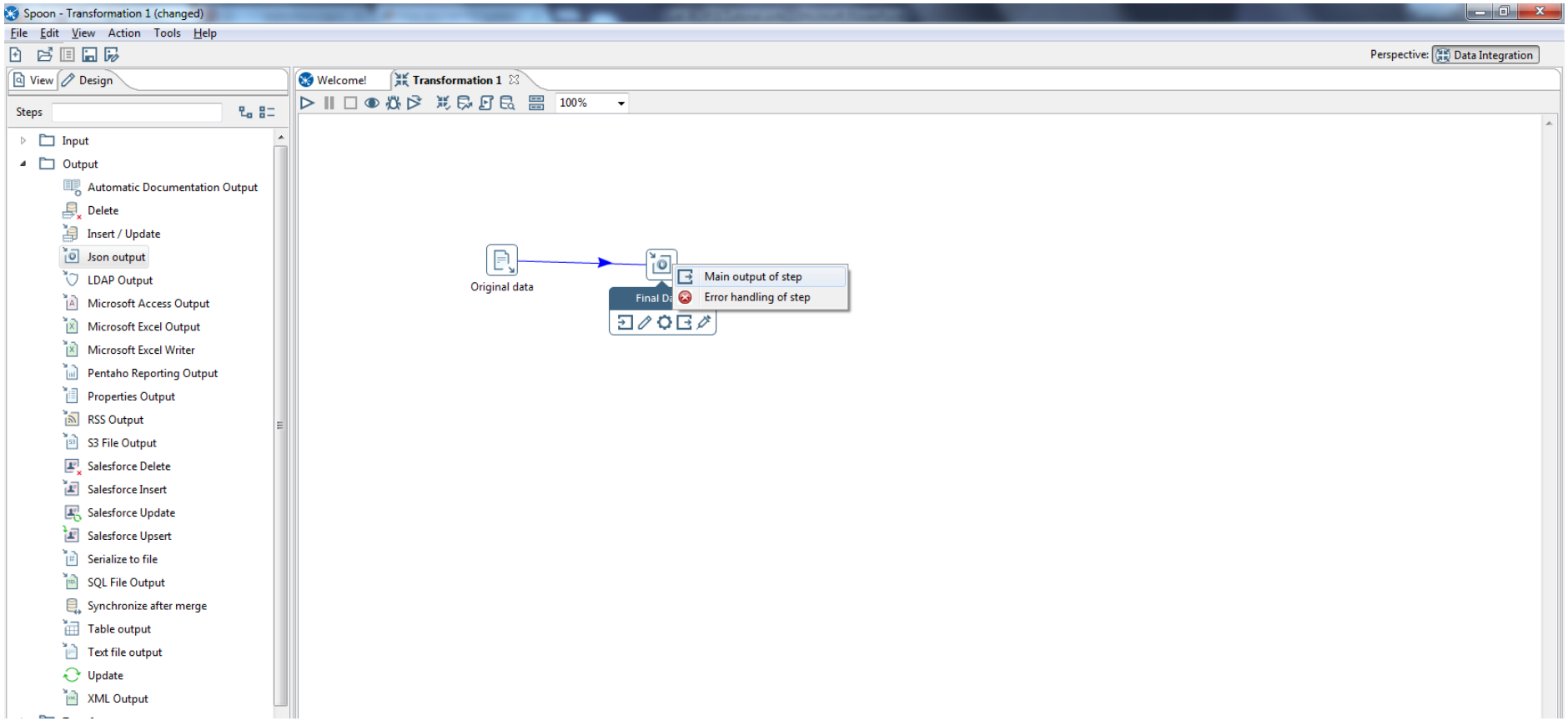
Let’s configure the output

Configure step name, file name and extension



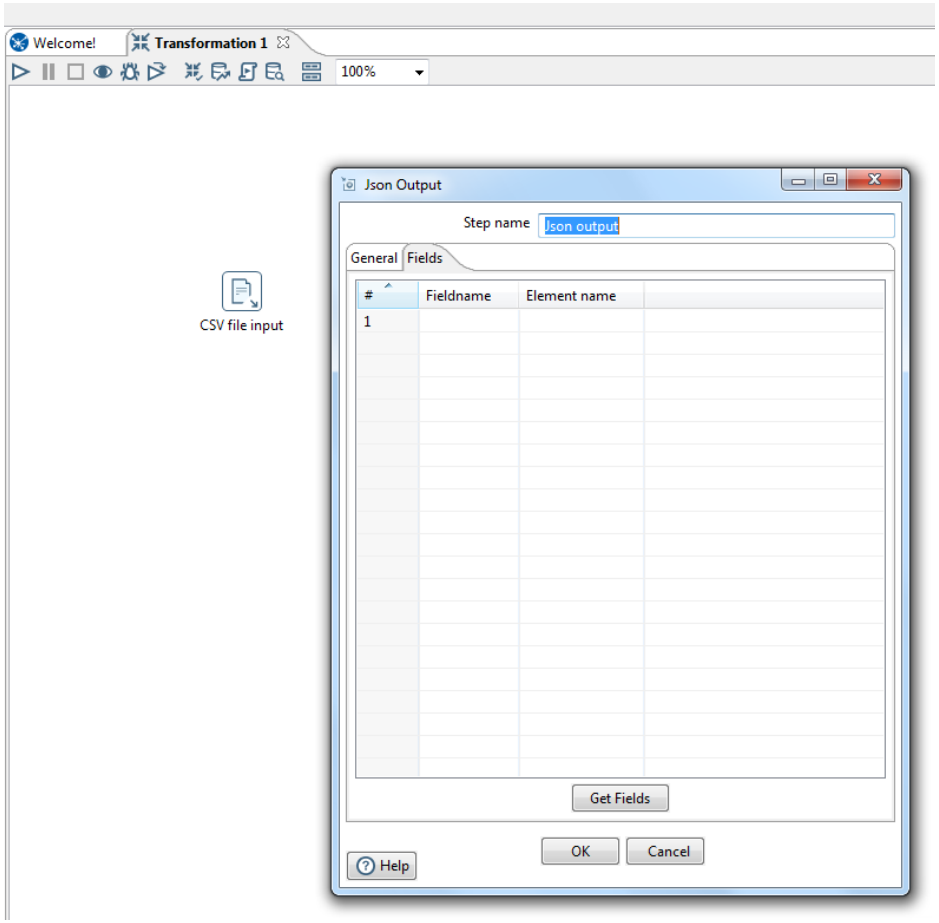
Connect input and output

Shift + click, drag, click



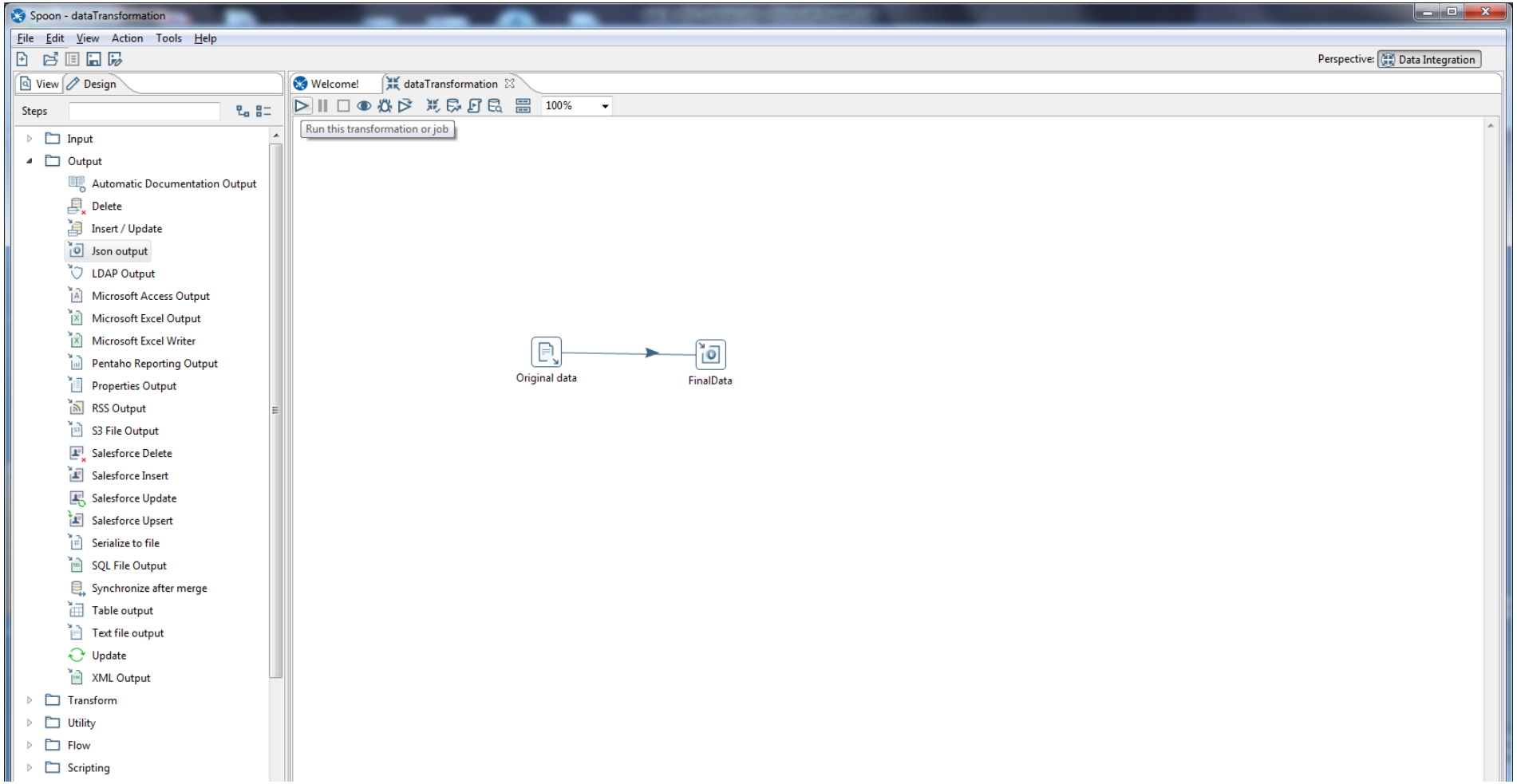
Let’s configure the output

Fields > get fields



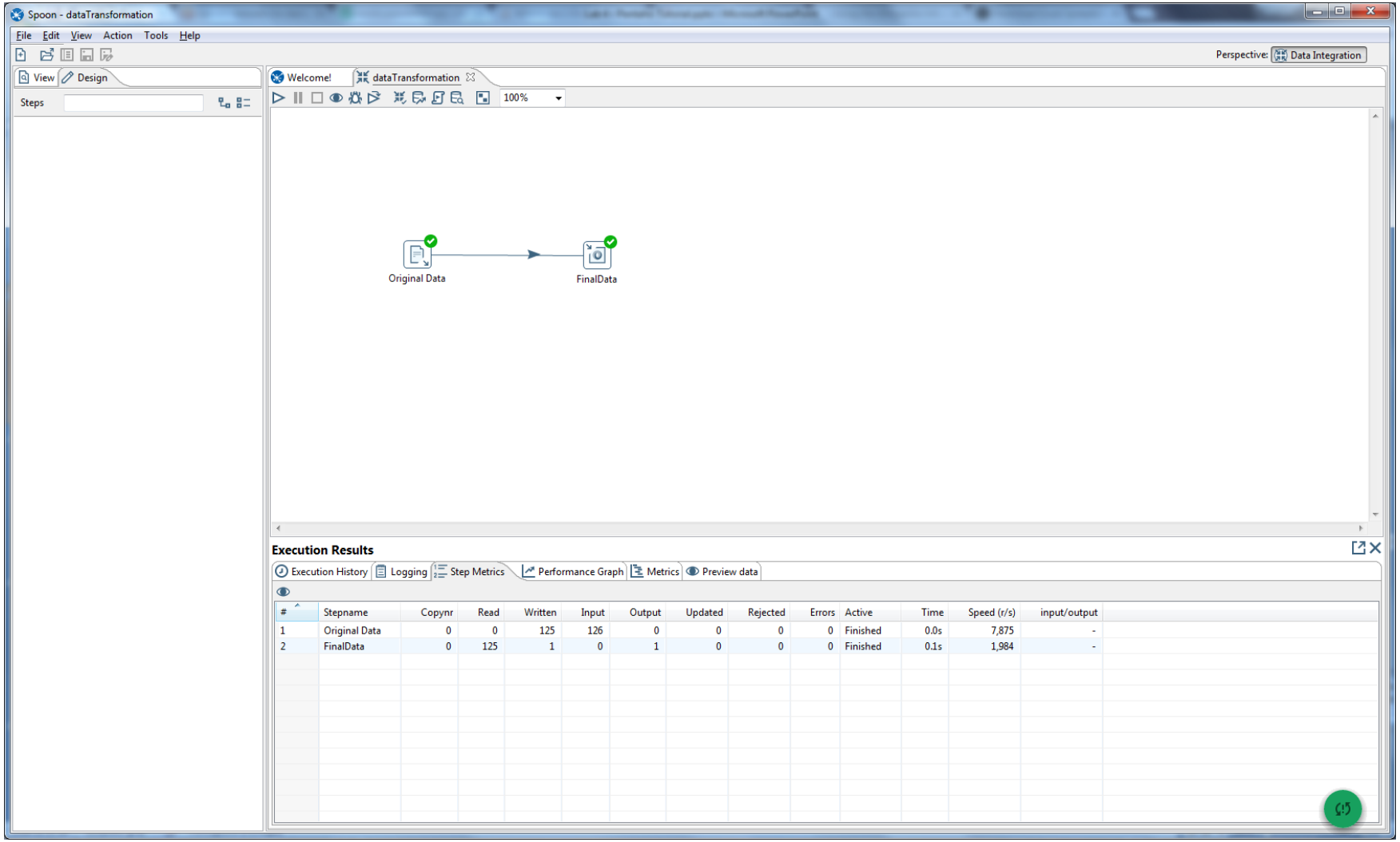
Run!

Save transformation, run and launch

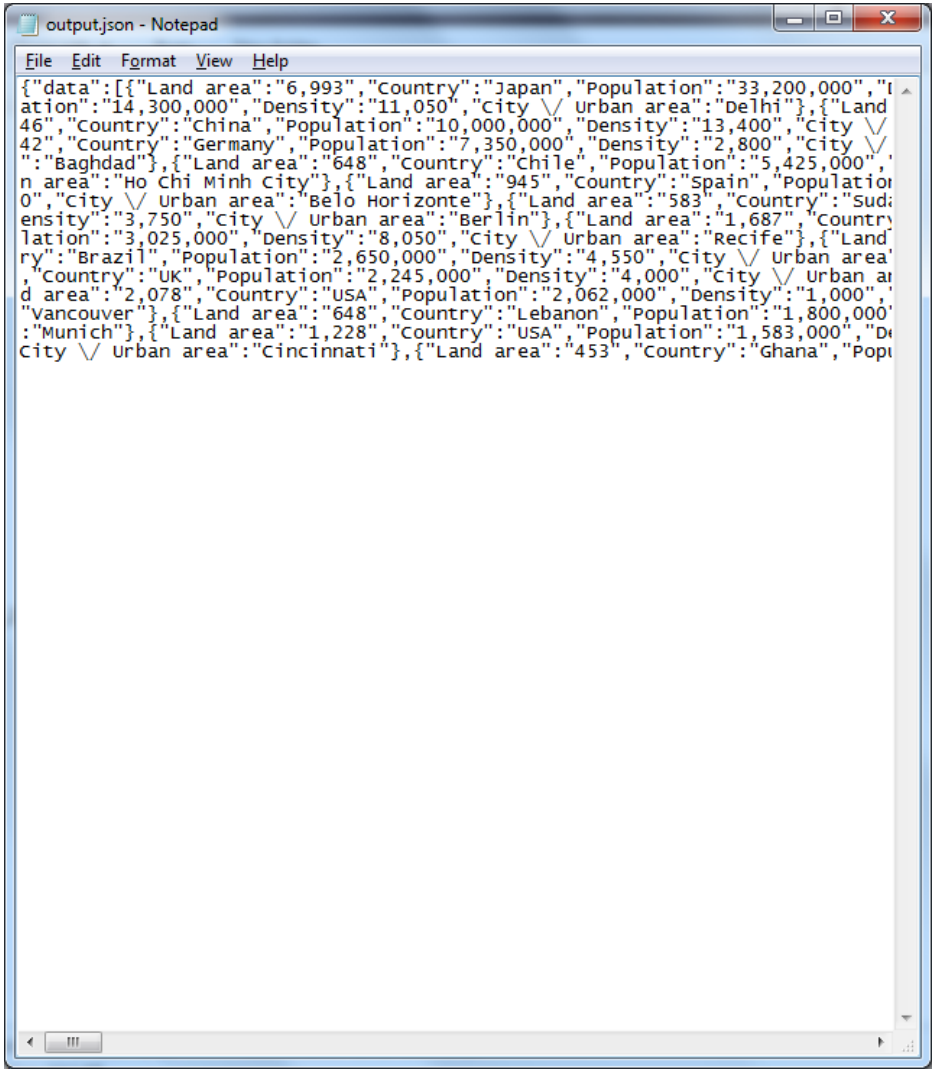


Data transformation complete

Check execution results



Now let's see the .json file



02

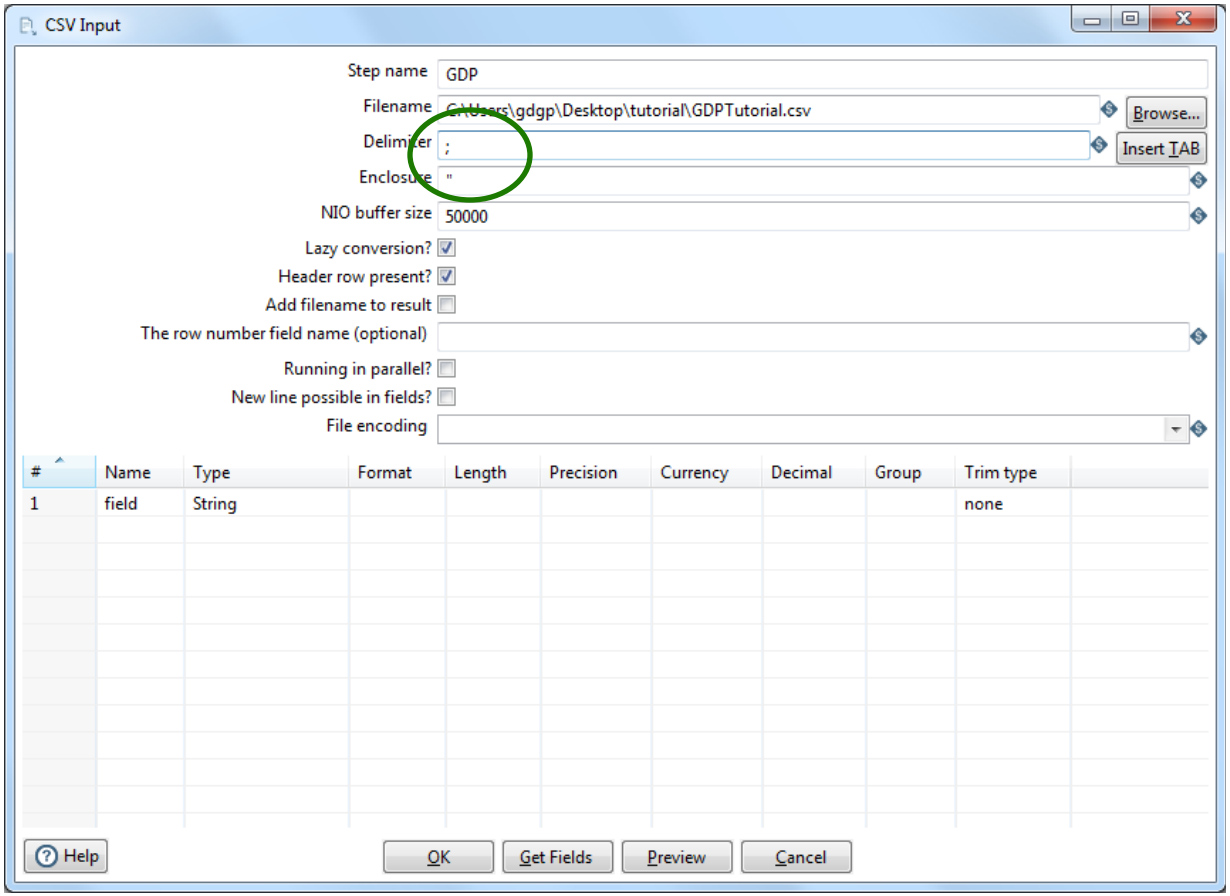
PENTAHO DATA INTEGRATION

MERGING TABLES

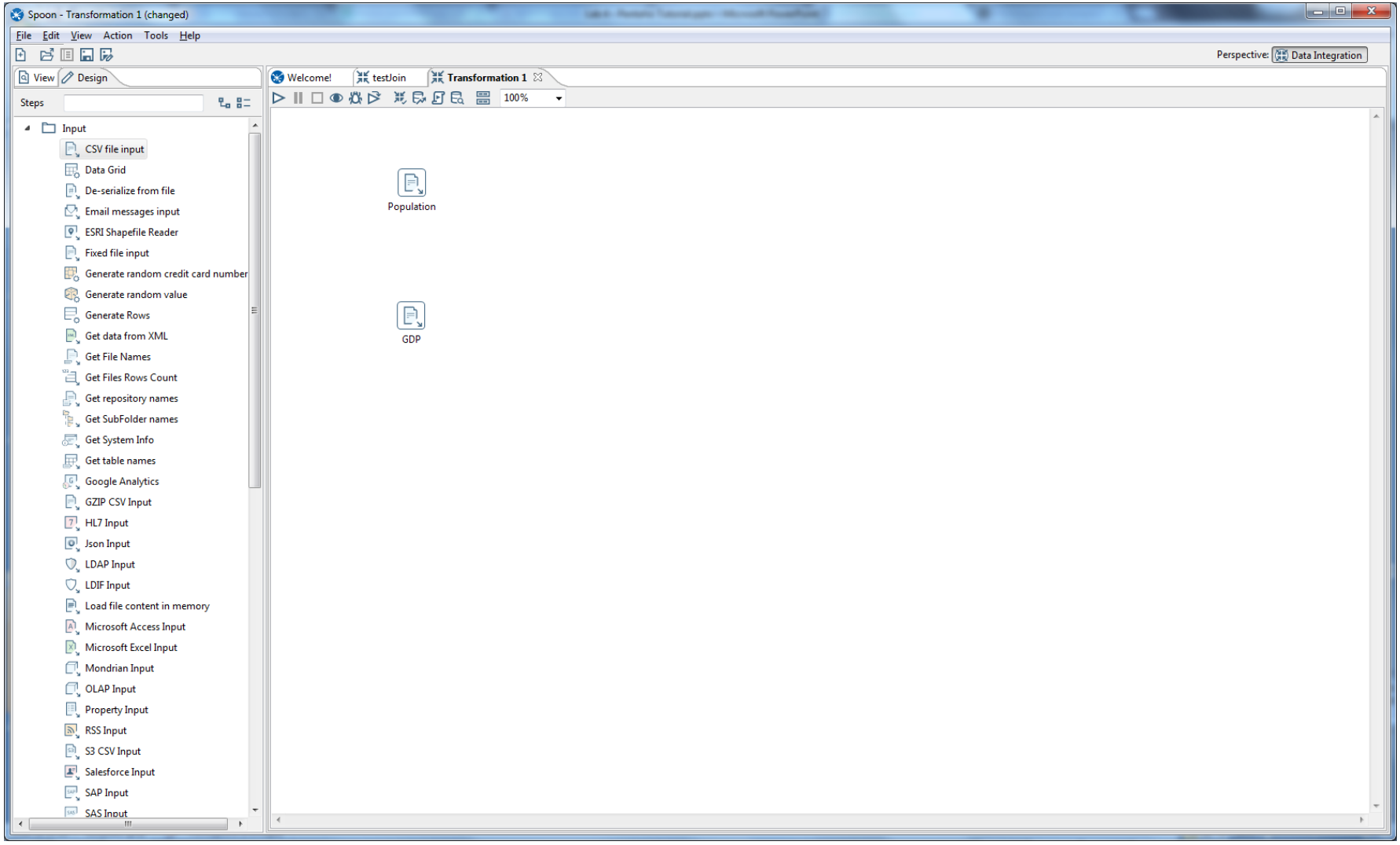
Now let's use both files as input.

Input > CSV File input > drag & drop (both files)

Attention:

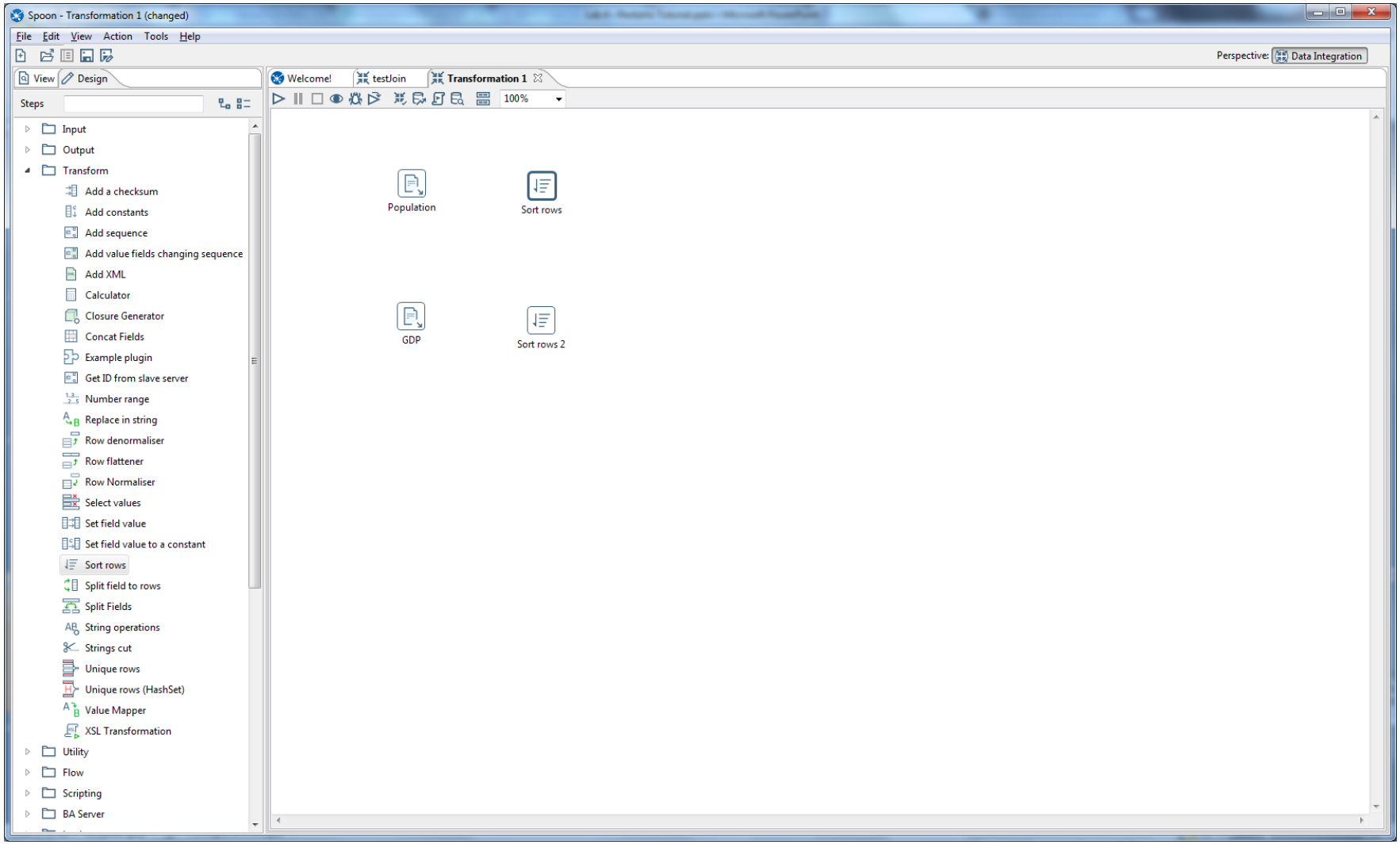


Two input files.



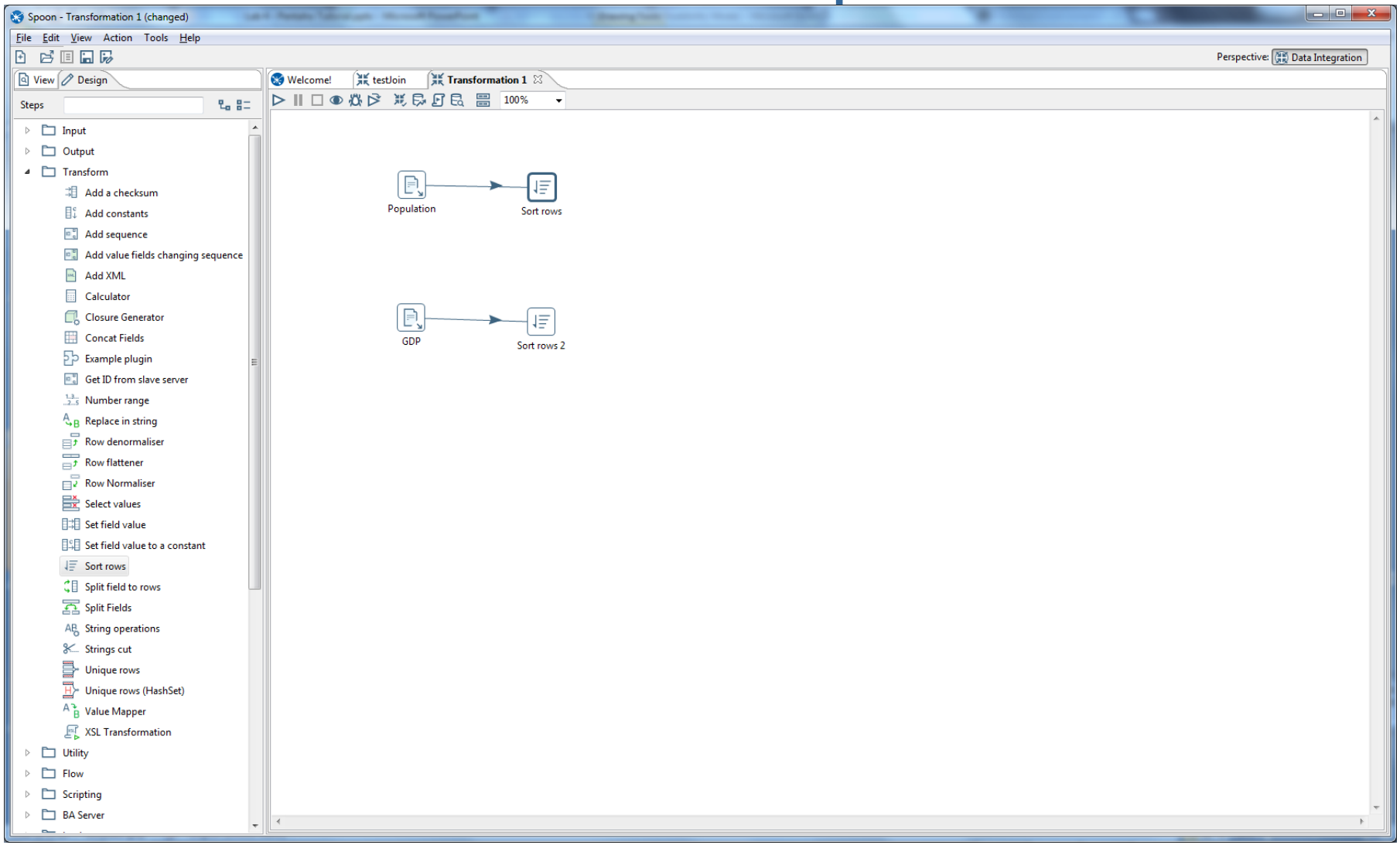
Now let's sort our data by country

Transform > sort rows > drag & drop



Now let's sort our data by country

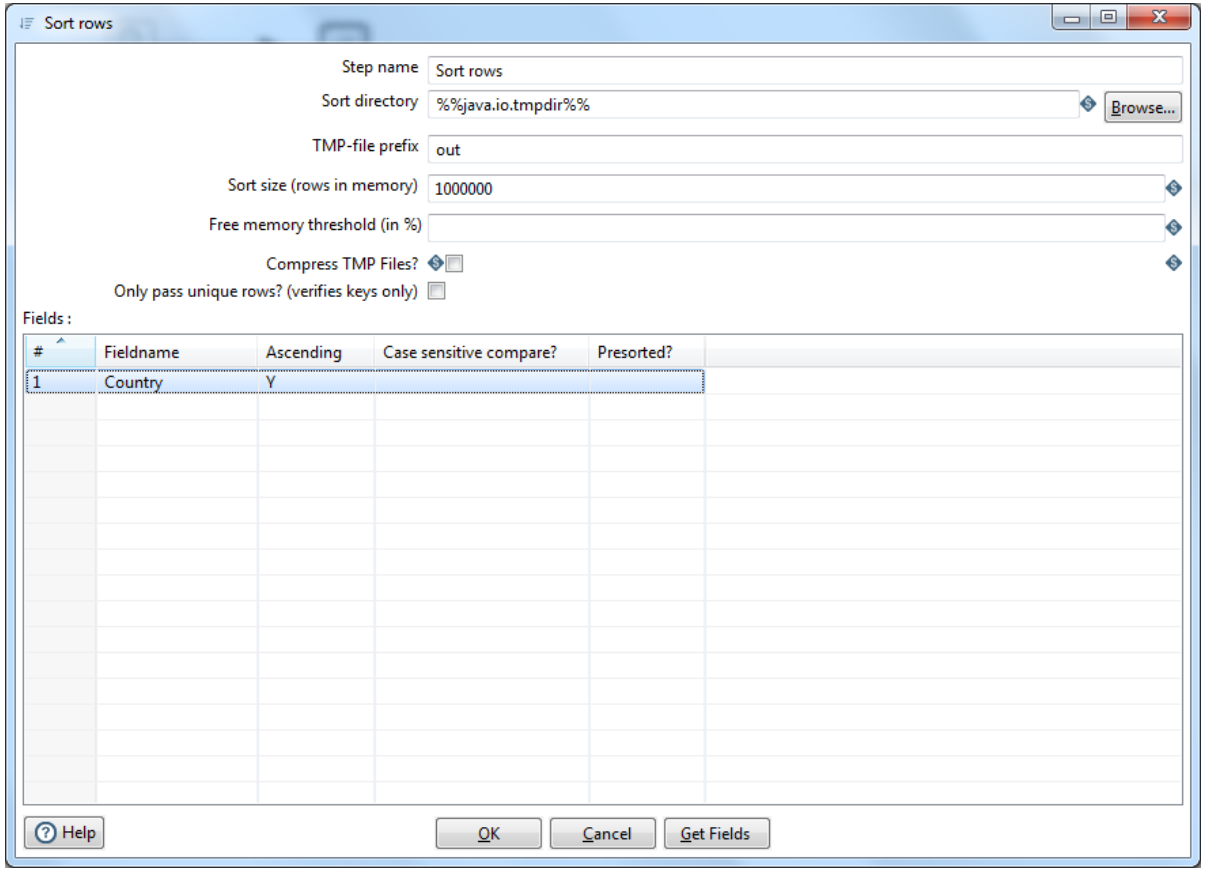
Create hops



Now let's sort our data by country

Configure sort steps (population):

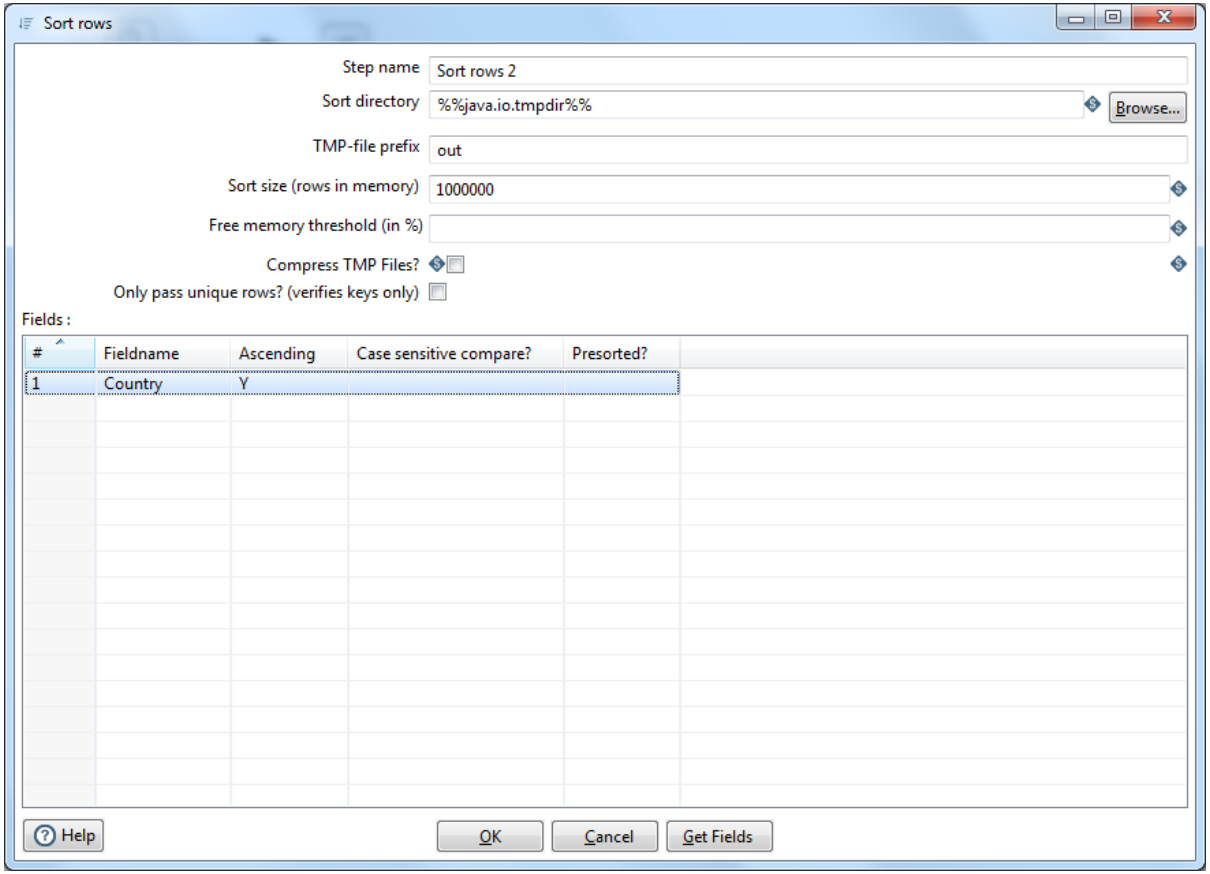
Double click > get fields > delete all but "country"



Now let's sort our data by country

Configure sort steps (GDP):

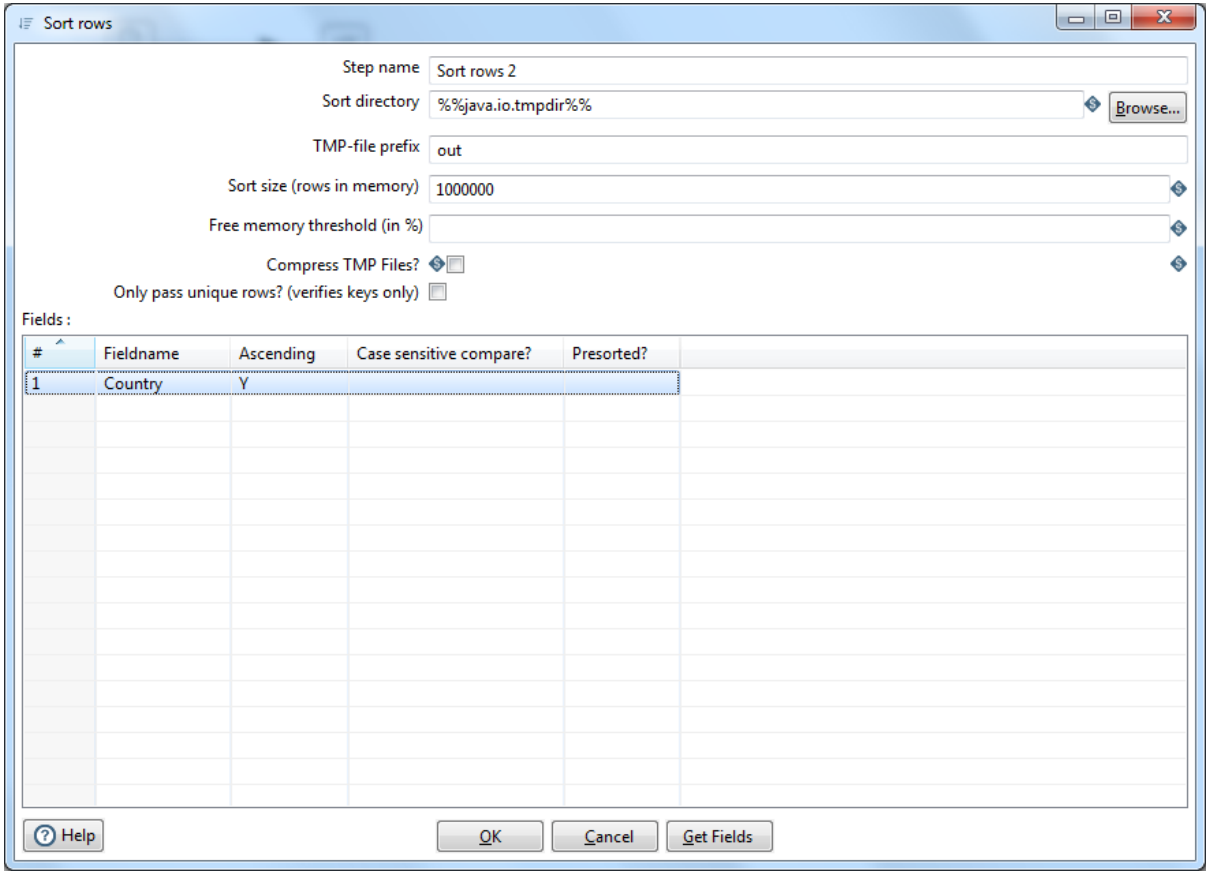
Double click > get fields > delete all but "country"



Now let's sort our data by country

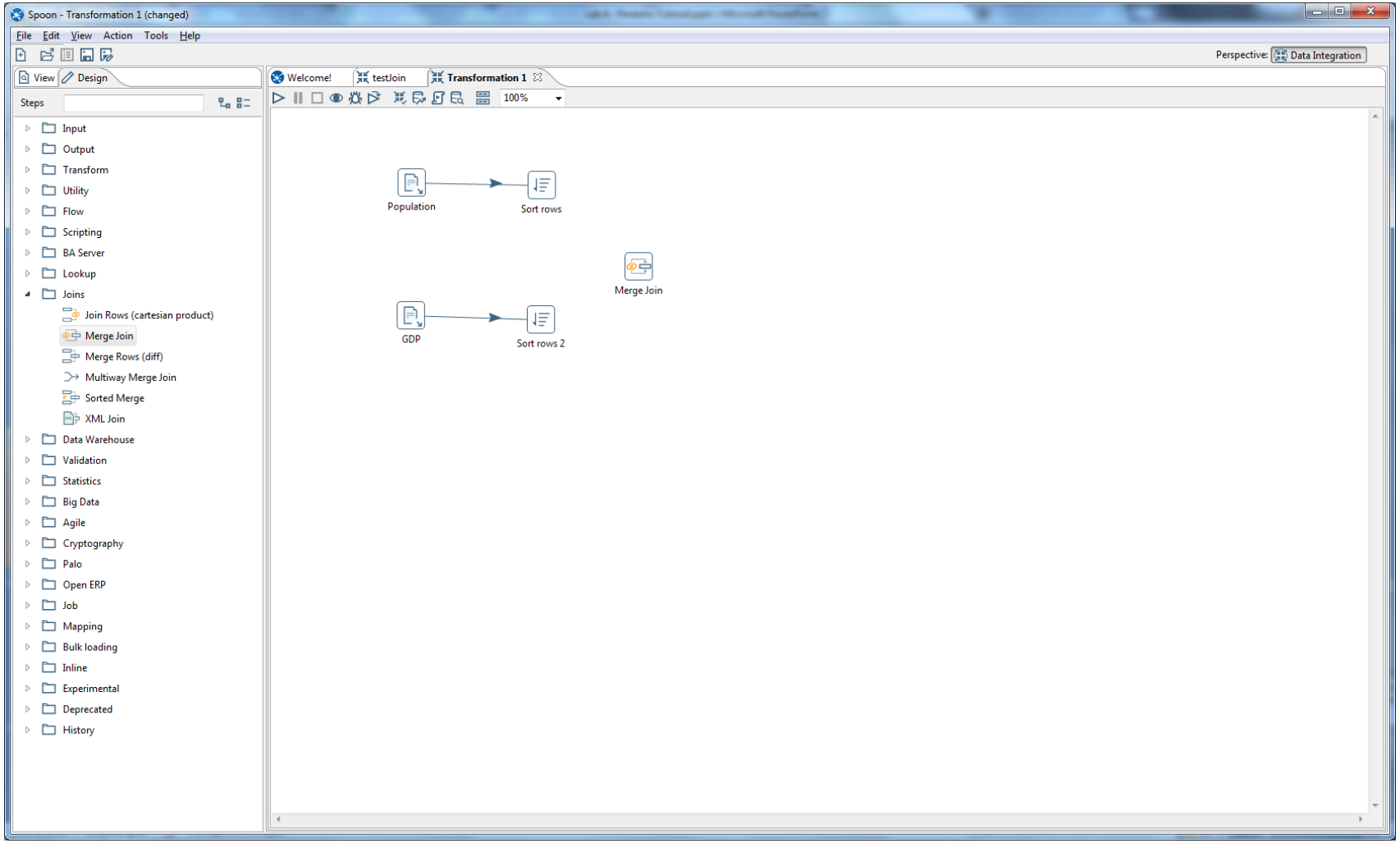
Configure sort steps (GDP):

Double click > get fields > delete all but "country"



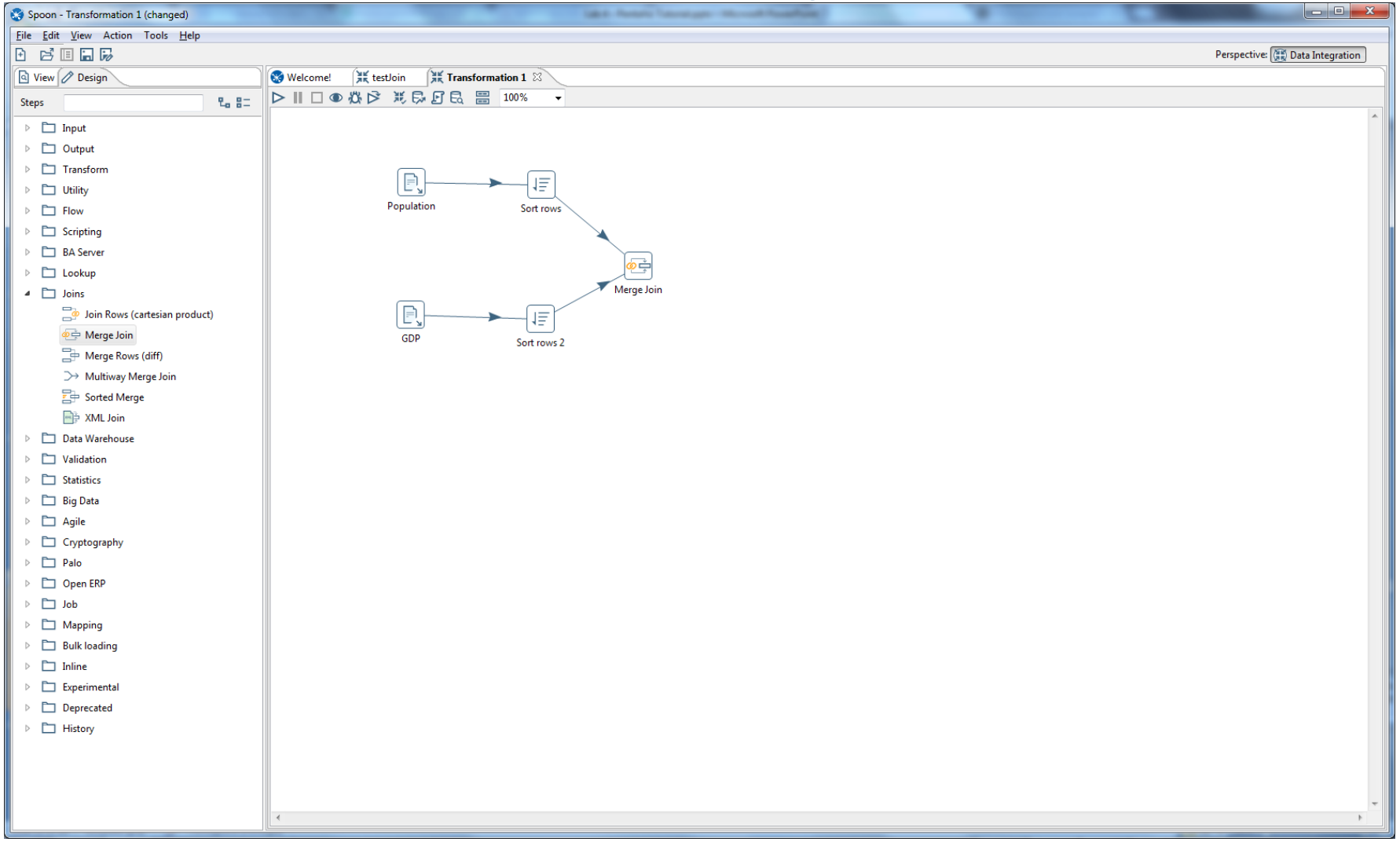
Merging by country…

Joins > Merge Join > drag&drop



Merging by country…

Create hops…

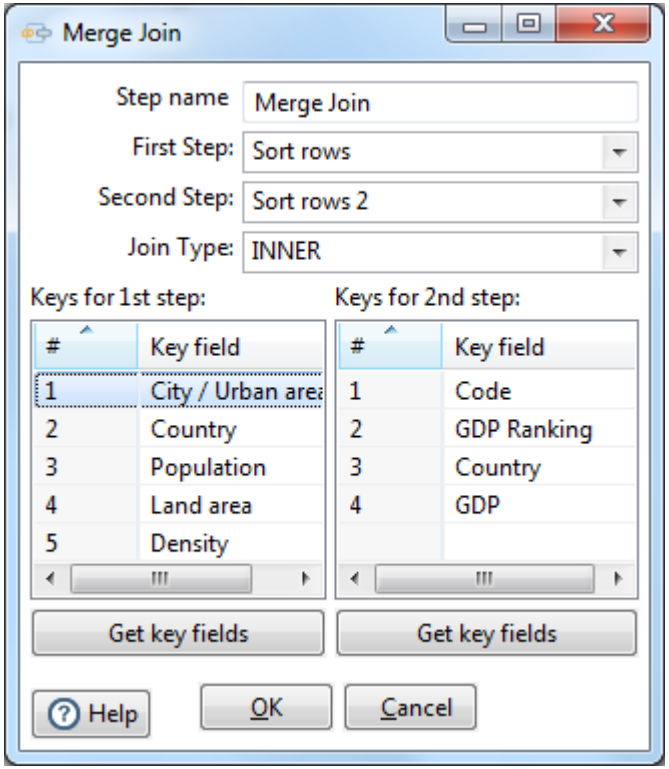


Merging by country…

Configure Merge Join step:

Double click > configure first & second step

Get key fields for both

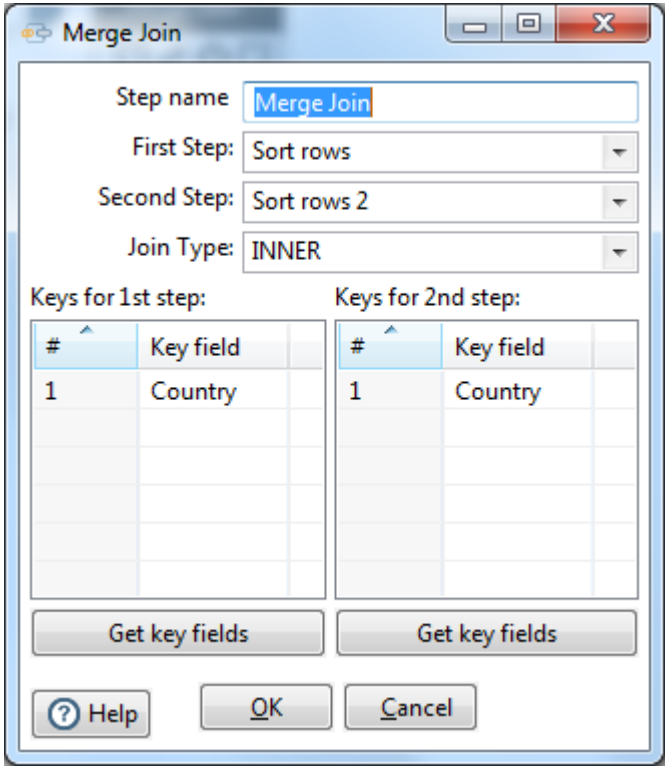


Merging by country…

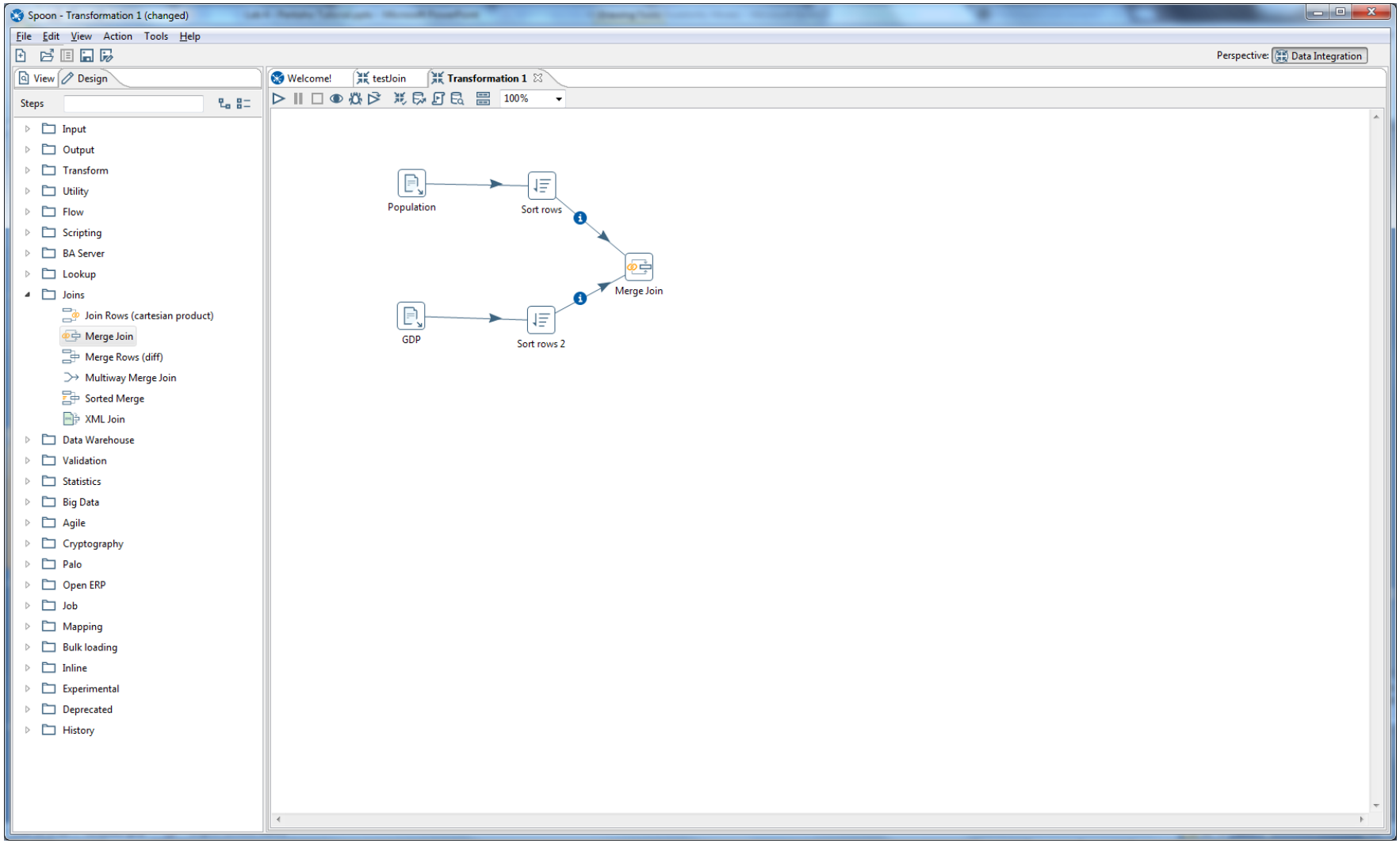
Configure Merge Join step:

Remove all keys except “Country”

Ok!



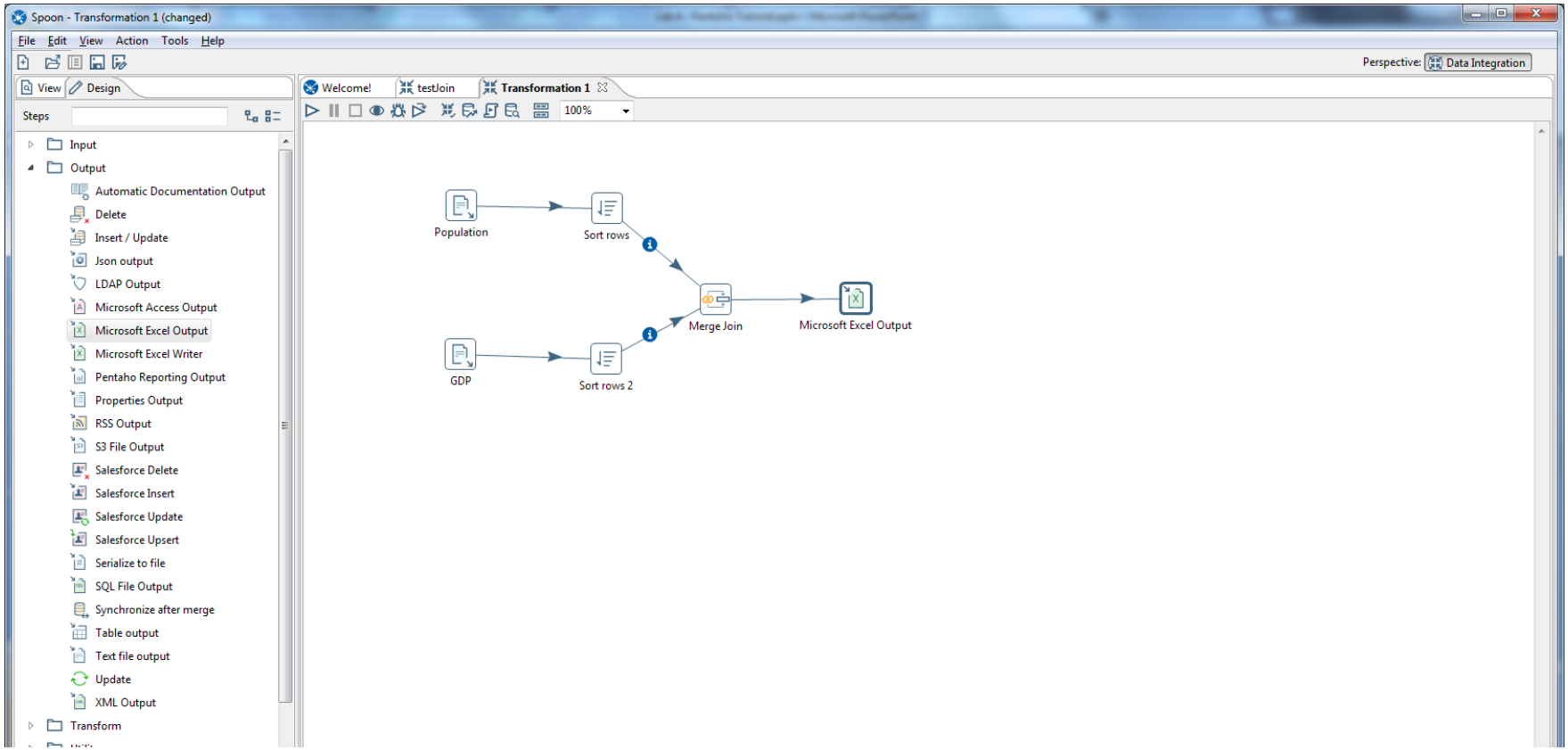
Almost there…



Let’s create an Excel file to test

Output > Microsoft Excel Output > drag&drop

Create hop

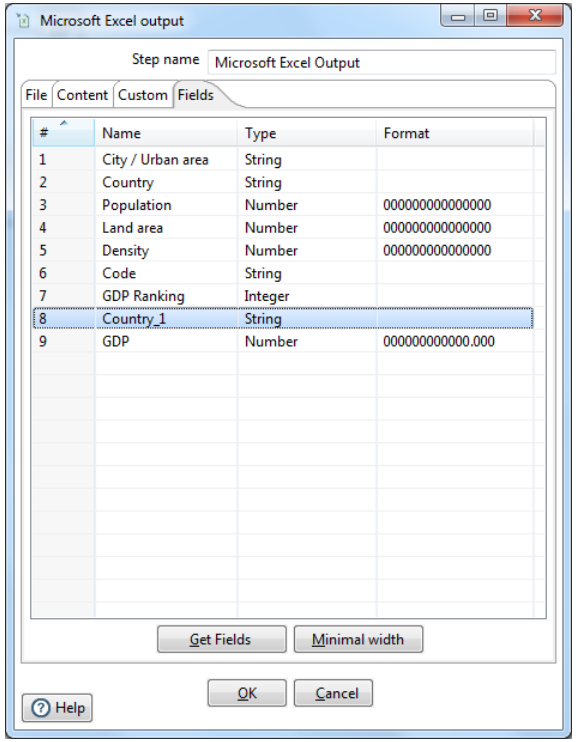


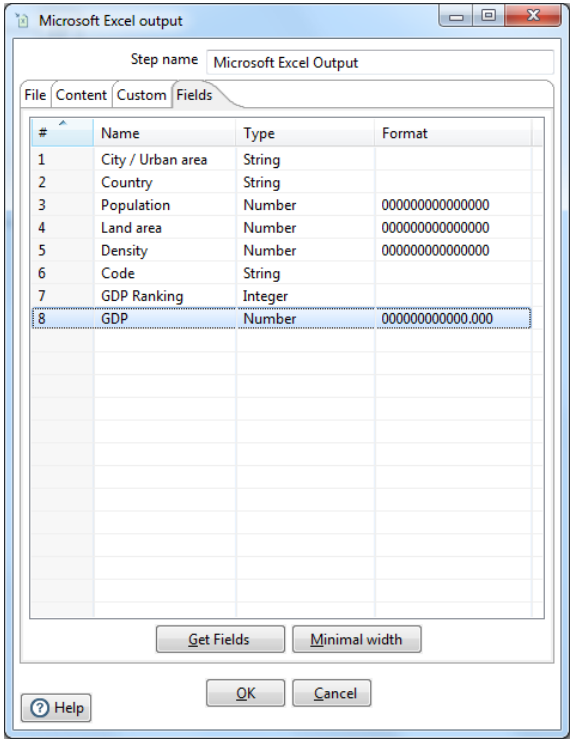
Let's create an Excel file to test

Configure Excel output:

Double click > configure filename, open "Fields" tab

Delete country_1 (the repeated field)

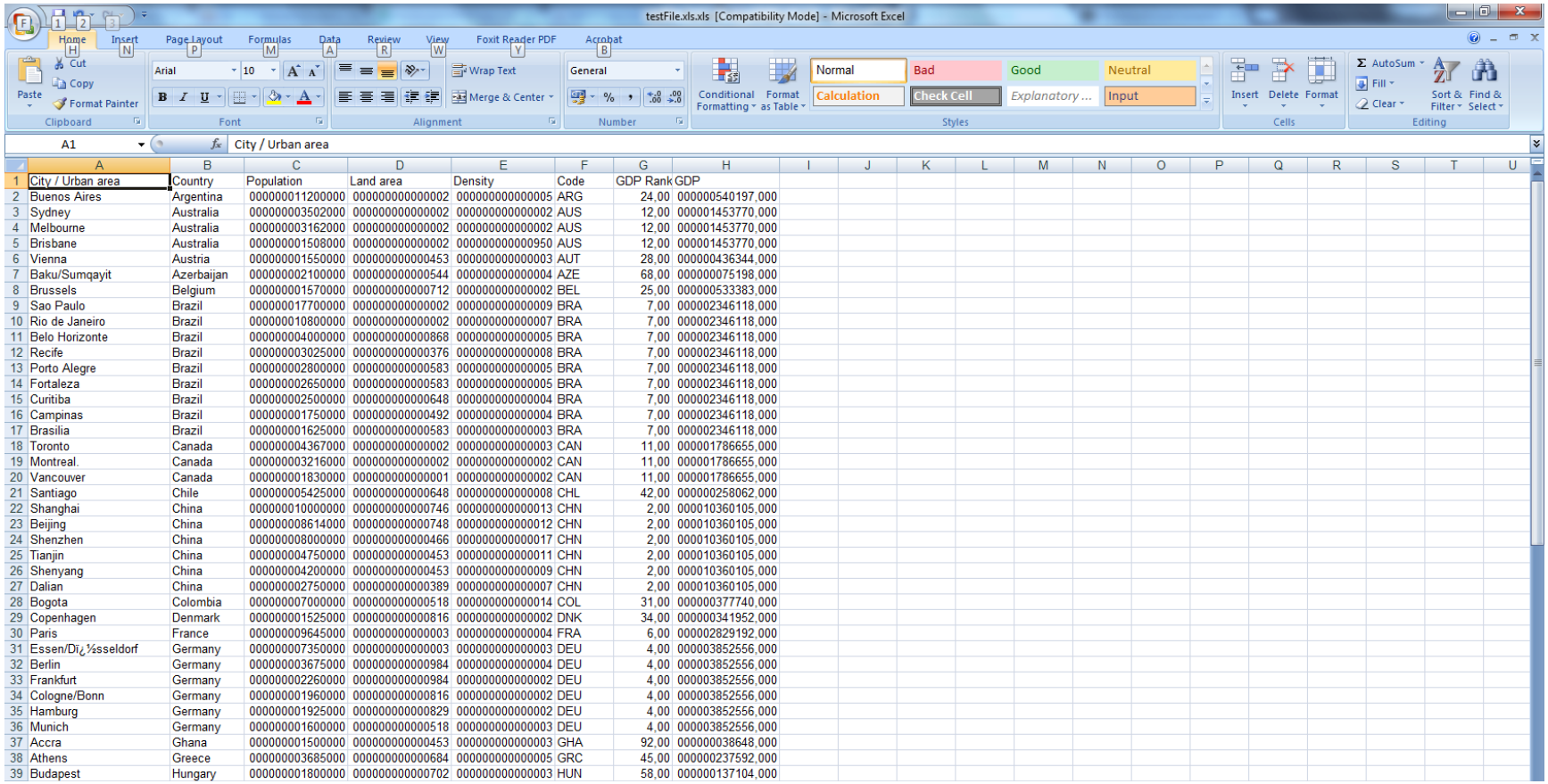




Let's create an Excel file to test

Run > Launch (save)

Open the Excel file:



	A	B	C	D	E	F	G	H
1	City / Urban area	Country	Population	Land area	Density	Code	GDP Rank	GDP
2	Buenos Aires	Argentina	000000011200000	000000000000002	000000000000005	ARG	24,00	000000540197,000
3	Sydney	Australia	000000003502000	000000000000002	000000000000002	AUS	12,00	000001453770,000
4	Melbourne	Australia	000000003162000	000000000000002	000000000000002	AUS	12,00	000001453770,000
5	Brisbane	Australia	000000001508000	000000000000002	000000000000950	AUS	12,00	000001453770,000
6	Vienna	Austria	000000001550000	000000000000453	000000000000003	AUT	28,00	000000436344,000
7	Baku/Sumqayit	Azerbaijan	000000002100000	000000000000544	000000000000004	AZE	68,00	000000075198,000
8	Brussels	Belgium	000000001570000	000000000000712	000000000000002	BEL	25,00	000000533383,000
9	Sao Paulo	Brazil	000000017700000	000000000000002	000000000000009	BRA	7,00	000002346118,000
10	Rio de Janeiro	Brazil	000000010800000	000000000000002	000000000000007	BRA	7,00	000002346118,000
11	Belo Horizonte	Brazil	000000004000000	000000000000868	000000000000005	BRA	7,00	000002346118,000
12	Recife	Brazil	000000003025000	000000000000376	000000000000008	BRA	7,00	000002346118,000
13	Porto Alegre	Brazil	000000002800000	000000000000583	000000000000005	BRA	7,00	000002346118,000
14	Fortaleza	Brazil	000000002650000	000000000000583	000000000000005	BRA	7,00	000002346118,000
15	Curitiba	Brazil	000000002500000	000000000000648	000000000000004	BRA	7,00	000002346118,000
16	Campinas	Brazil	000000001750000	000000000000492	000000000000004	BRA	7,00	000002346118,000
17	Brasilia	Brazil	000000001625000	000000000000583	000000000000003	BRA	7,00	000002346118,000
18	Toronto	Canada	000000004367000	000000000000002	000000000000003	CAN	11,00	000001786655,000
19	Montreal.	Canada	000000003216000	000000000000002	000000000000002	CAN	11,00	000001786655,000
20	Vancouver	Canada	000000001830000	000000000000001	000000000000002	CAN	11,00	000001786655,000
21	Santiago	Chile	000000005425000	000000000000648	000000000000008	CHL	42,00	000000258062,000
22	Shanghai	China	000000010000000	000000000000746	000000000000013	CHN	2,00	000010360105,000
23	Beijing	China	000000008614000	000000000000748	000000000000012	CHN	2,00	000010360105,000
24	Shenzhen	China	000000008000000	000000000000466	000000000000017	CHN	2,00	000010360105,000
25	Tianjin	China	000000004750000	000000000000453	000000000000011	CHN	2,00	000010360105,000
26	Shenyang	China	000000004200000	000000000000453	000000000000009	CHN	2,00	000010360105,000
27	Dalian	China	000000002750000	000000000000389	000000000000007	CHN	2,00	000010360105,000
28	Bogota	Colombia	000000007000000	000000000000518	000000000000014	COL	31,00	000000377740,000
29	Copenhagen	Denmark	000000001525000	000000000000816	000000000000002	DNK	34,00	000000341952,000
30	Paris	France	000000009645000	000000000000003	000000000000004	FRA	6,00	000002829192,000
31	Essen/Dï¿½sseldorf	Germany	000000007350000	000000000000003	000000000000003	DEU	4,00	000003852556,000
32	Berlin	Germany	000000003675000	000000000000984	000000000000004	DEU	4,00	000003852556,000
33	Frankfurt	Germany	000000002260000	000000000000984	000000000000002	DEU	4,00	000003852556,000
34	Cologne/Bonn	Germany	000000001960000	000000000000816	000000000000002	DEU	4,00	000003852556,000
35	Hamburg	Germany	000000001925000	000000000000829	000000000000002	DEU	4,00	000003852556,000
36	Munich	Germany	000000001600000	000000000000518	000000000000003	DEU	4,00	000003852556,000
37	Accra	Ghana	000000001500000	000000000000453	000000000000003	GHA	92,00	000000038648,000
38	Athens	Greece	000000003685000	000000000000684	000000000000005	GRC	45,00	000000237592,000
39	Budapest	Hungary	000000001800000	000000000000702	000000000000003	HUN	58,00	000000137104,000

03

PENTAHO DATA INTEGRATION:

EXERCISE: FILTERING

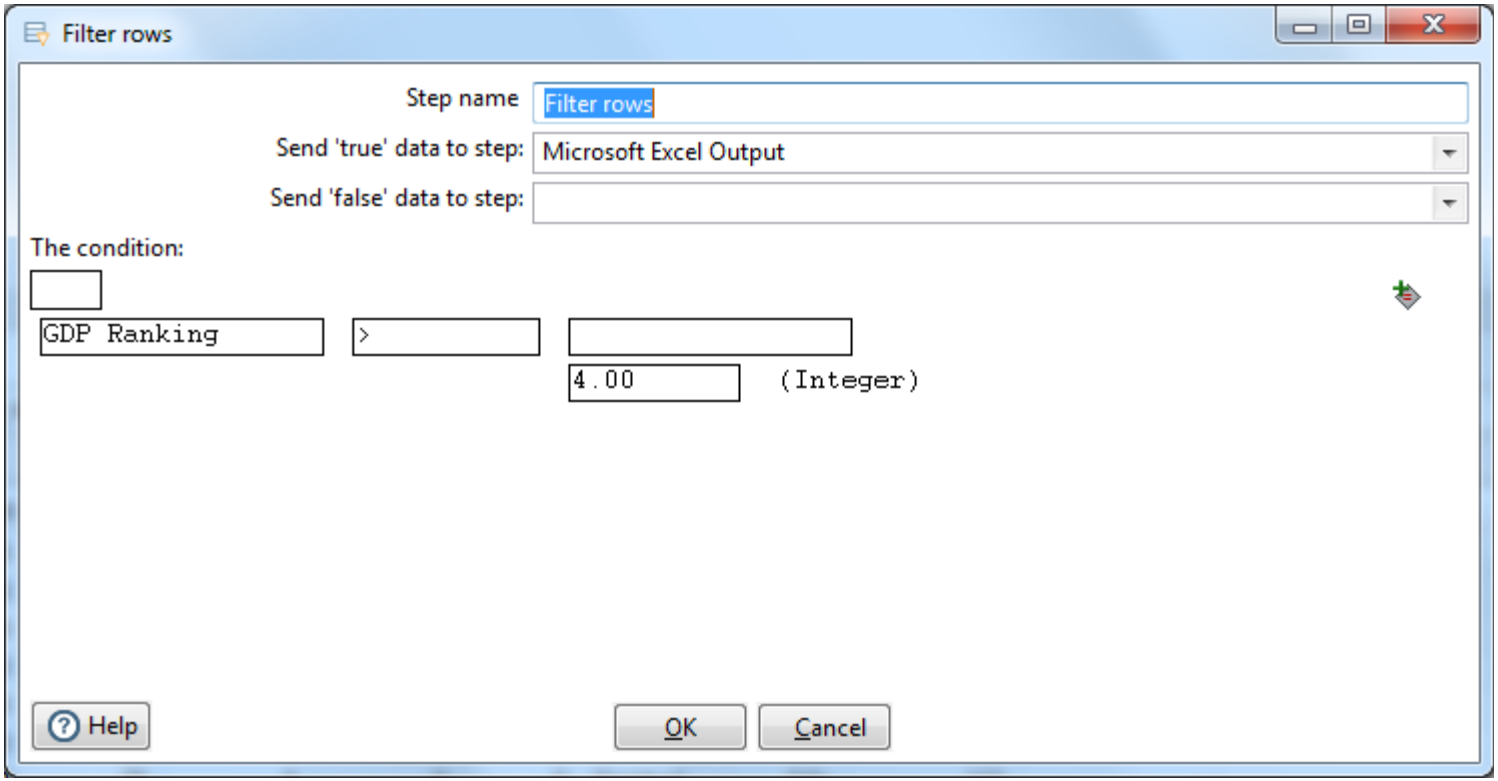
Remove the entries with GDP Ranking <= 4

Flow > Filter rows

Configure options

Don’t forget to keep hops functional!

Remove the entries with GDP Ranking <= 4



04

NEXT LAB

Data > Information

Parse your original (raw) **data** into

INFORMATION

to use as input for your visualization

QUESTIONS?