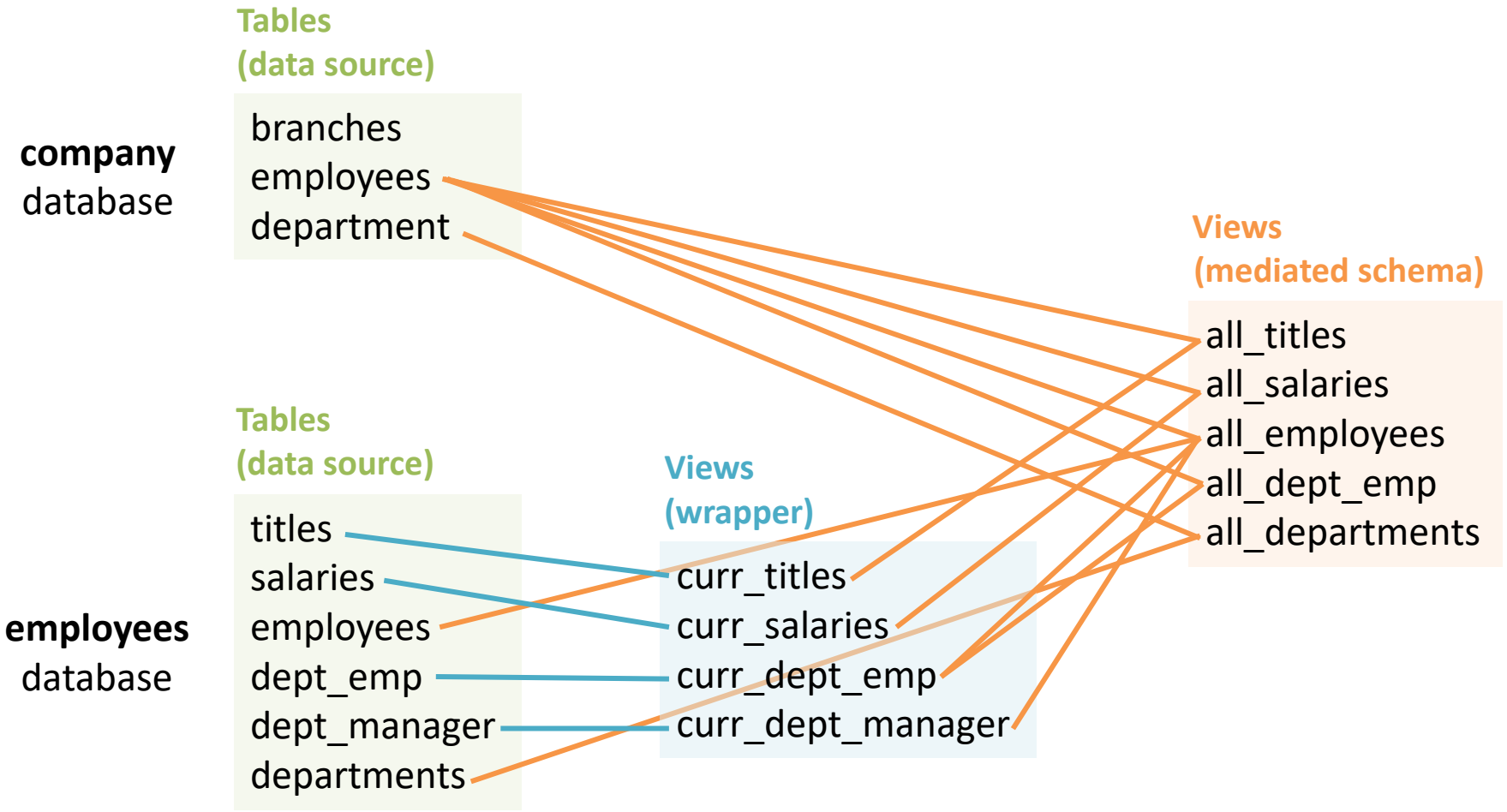
Data Analysis and Integration

Introduction to ETL tools

ETL:
Extract
Transform
e Load

Data integration using views



Data integration using views

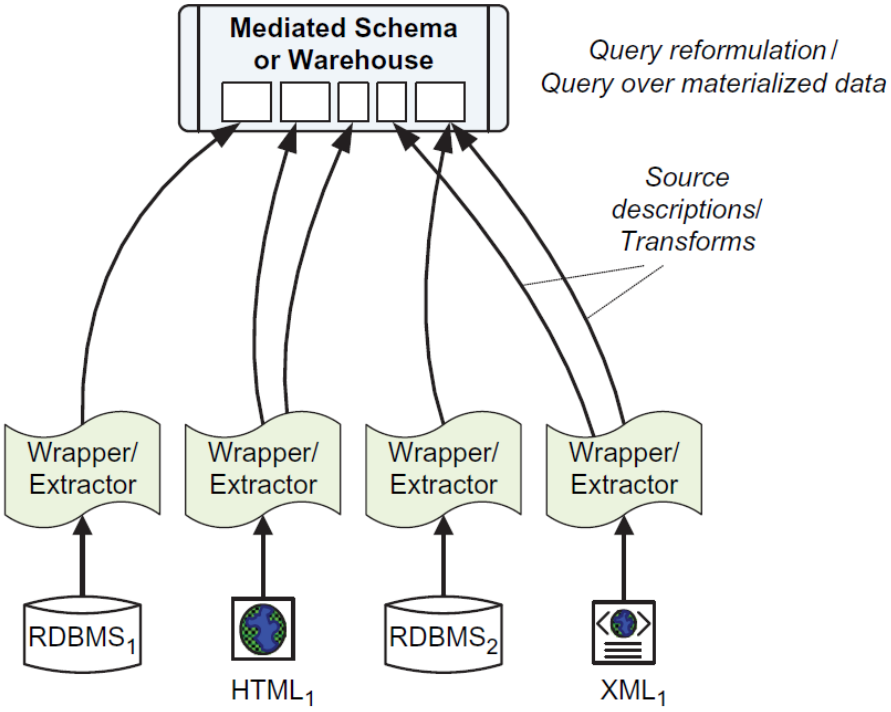
- Define the **mediated schema**
- Define the **schema mappings** between the **data sources** (or their **wrappers**) and the **mediated schema**

- Write query over **mediated schema**
- **Query unfolding** reformulates the query over the **mediated schema** and the **wrappers** as a query over the **data sources**

- Results are computed on-the-fly and are always up-to-date

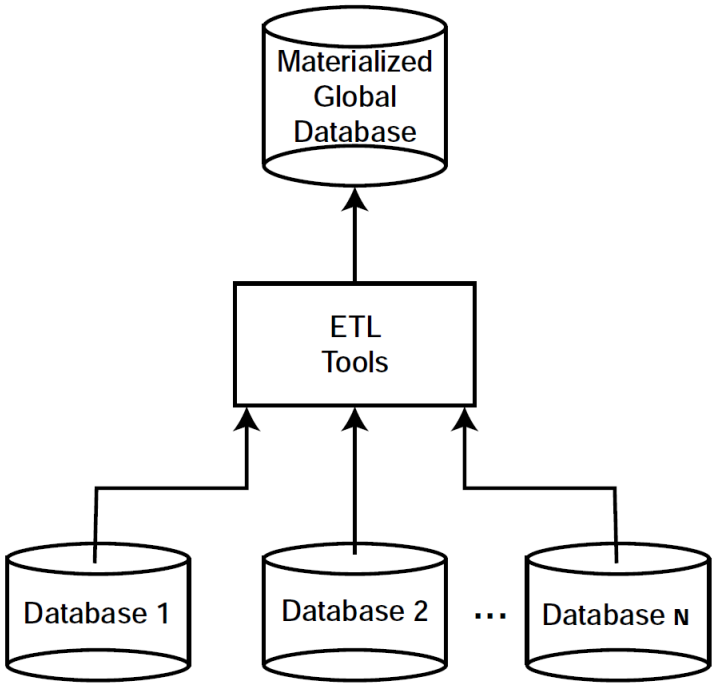
Data integration vs. data warehousing

Query unfolding IRT
Results ALWAYS up-to-date



A. Doan, A. Halevy, Z. Ives
Principles of Data Integration
Morgan Kaufmann, 2012

Queries made already
Results as up-to-date as the previous update



T. Özsu, P. Valduriez
Principles of Distributed Database Systems
Springer, 2011

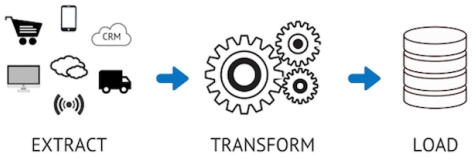
Data warehousing

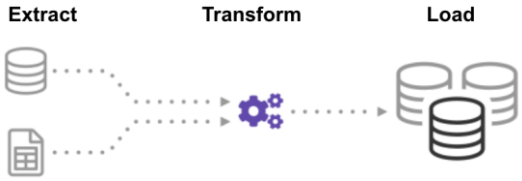
- Design a **data warehouse**
  - the data warehouse has its own schema, which is different from the data source schemas
- Implement an **extract-transform-load** (ETL) process
  - extract from data sources, transform data, store into data warehouse

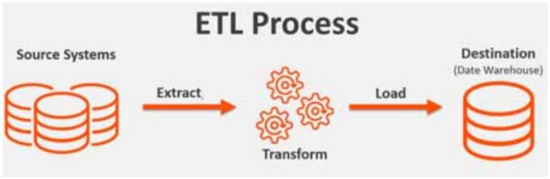
- Write query over **data warehouse**
- Retrieve results from storage
  - run ETL process regularly to keep **data warehouse** up-to-date

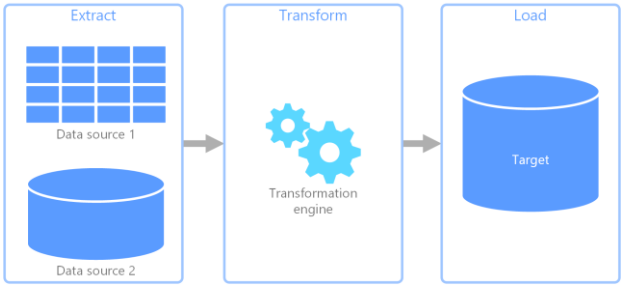
ETL process (Pentaho is an ETL tool)

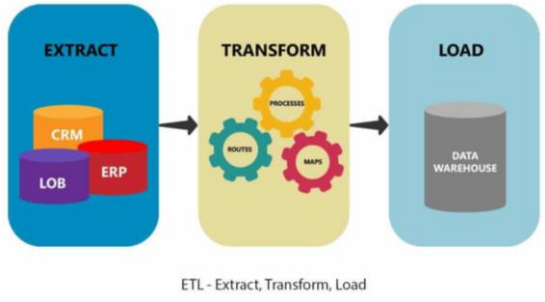
- Extract, transform, load

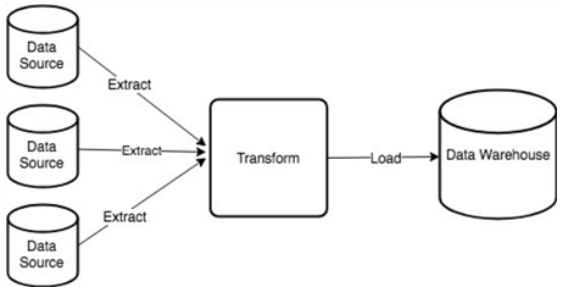






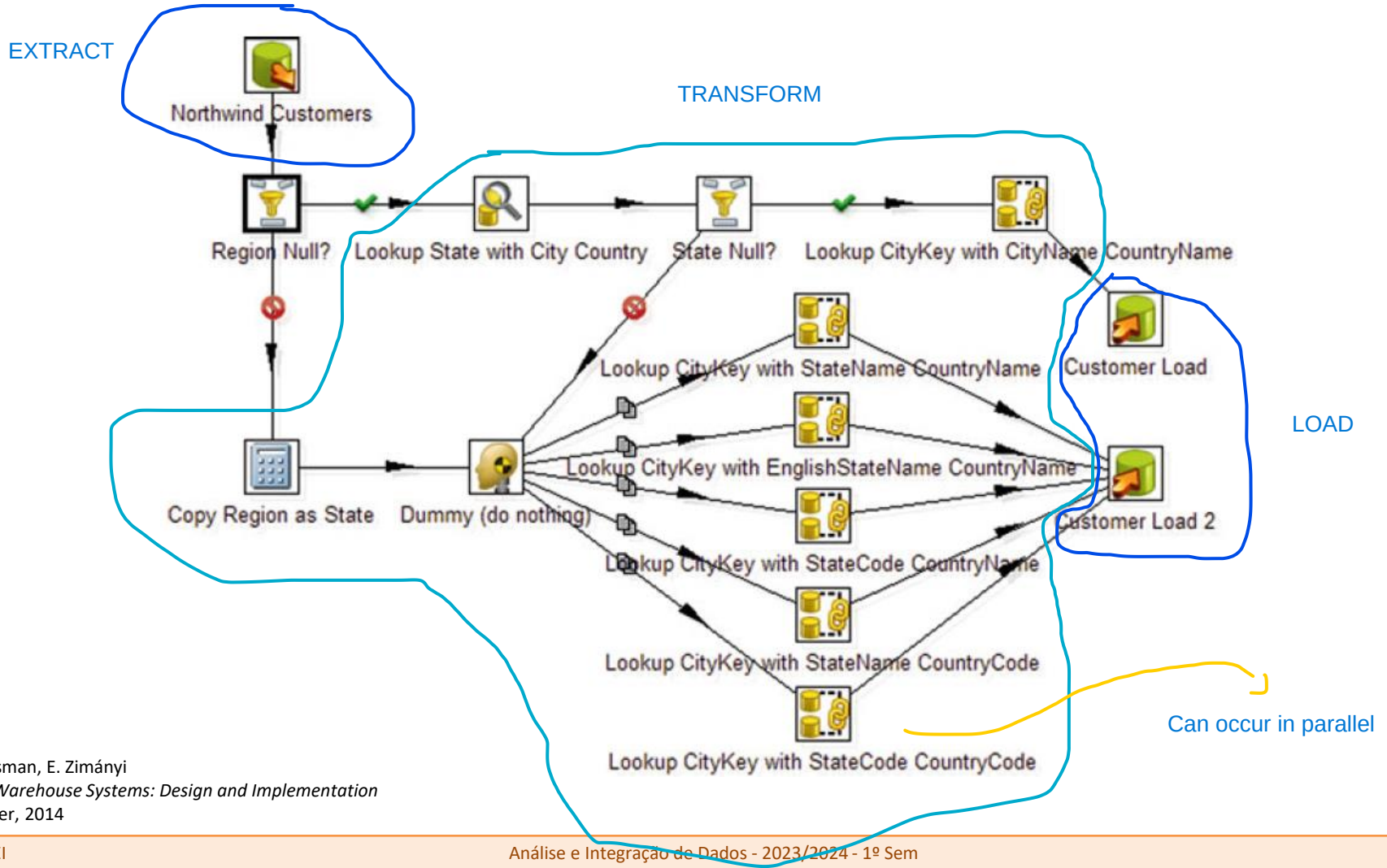






ETL process

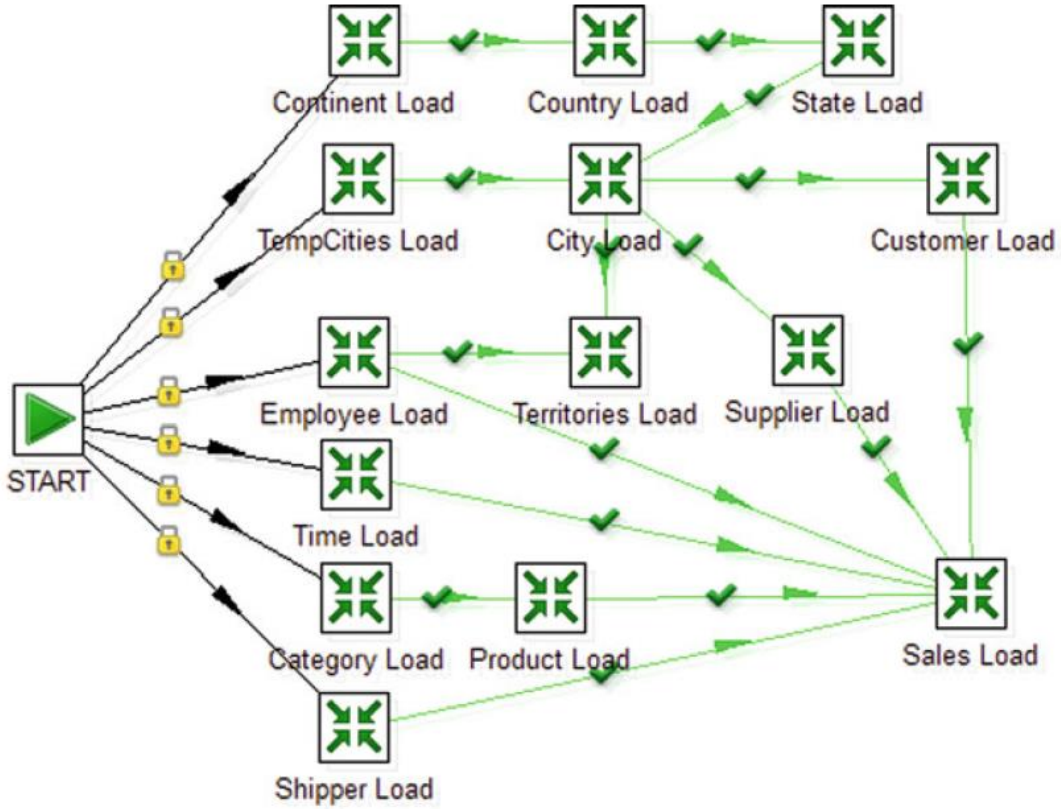
- Extracting, transforming and loading customers



A. Vaisman, E. Zimányi
Data Warehouse Systems: Design and Implementation
Springer, 2014

ETL process

- An ETL process comprises many such transformations



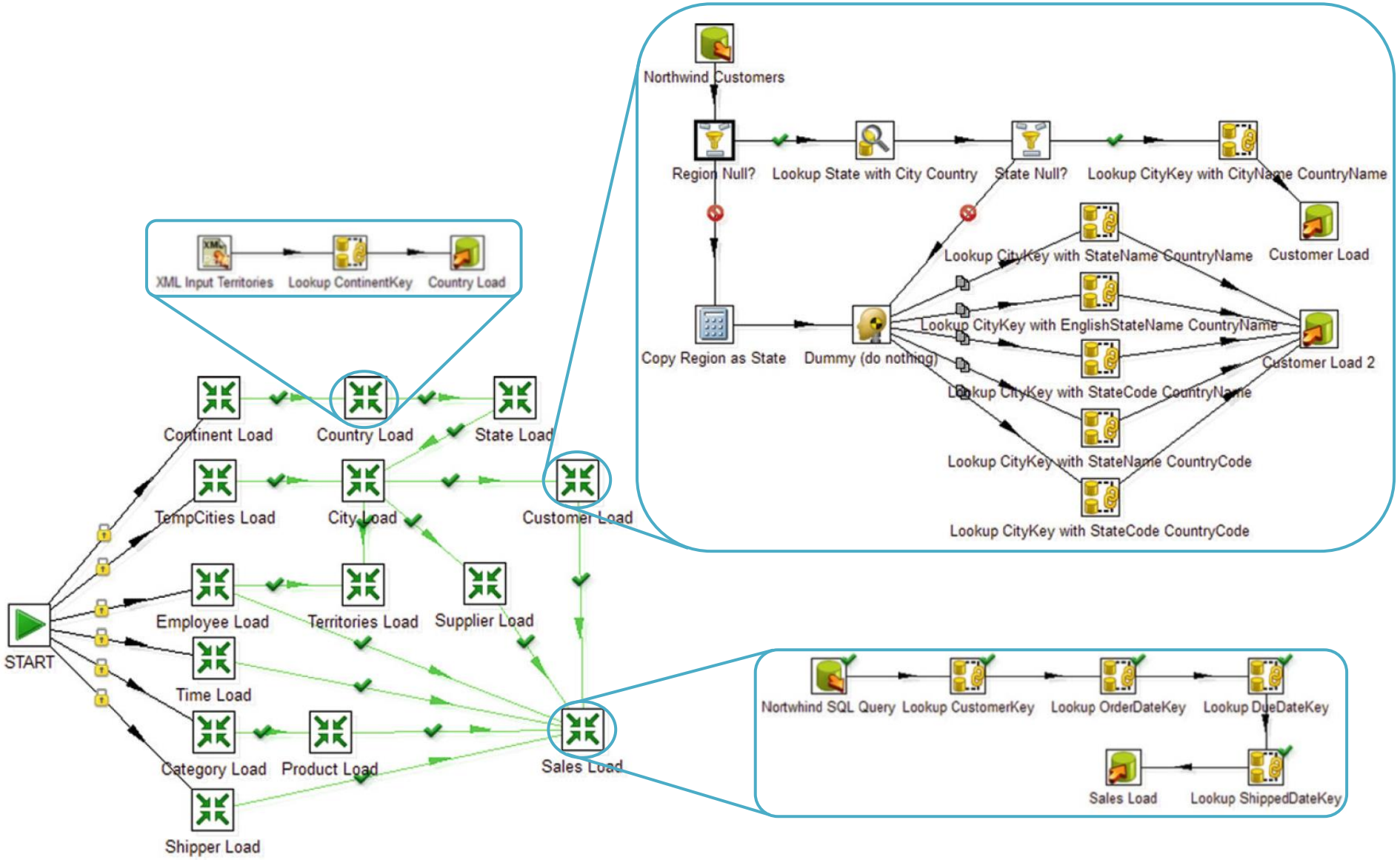
We might need to transform terabytes of data!

Not many computers can handle that
So! We pipeline/stream this system

All steps may occur at the same time

A. Vaisman, E. Zimányi
Data Warehouse Systems: Design and Implementation
Springer, 2014

ETL process


Northwind Customers
Region Null?
Lookup State with City Country
State Null?
Lookup CityKey with CityName CountryName
Customer Load
Lookup CityKey with StateName CountryName
Lookup CityKey with EnglishStateName CountryName
Lookup CityKey with StateCode CountryName
Lookup CityKey with StateName CountryCode
Lookup CityKey with StateCode CountryCode
Customer Load 2
Copy Region as State
Dummy (do nothing)
XML Input Territories
Lookup ContinentKey
Country Load
Continent Load
Country Load
State Load
TempCities Load
City Load
Customer Load
START
Employee Load
Territories Load
Supplier Load
Time Load
Category Load
Product Load
Shipper Load
Sales Load
Northwind SQL Query
Lookup CustomerKey
Lookup OrderDateKey
Lookup DueDateKey
Sales Load
Lookup ShippedDateKey

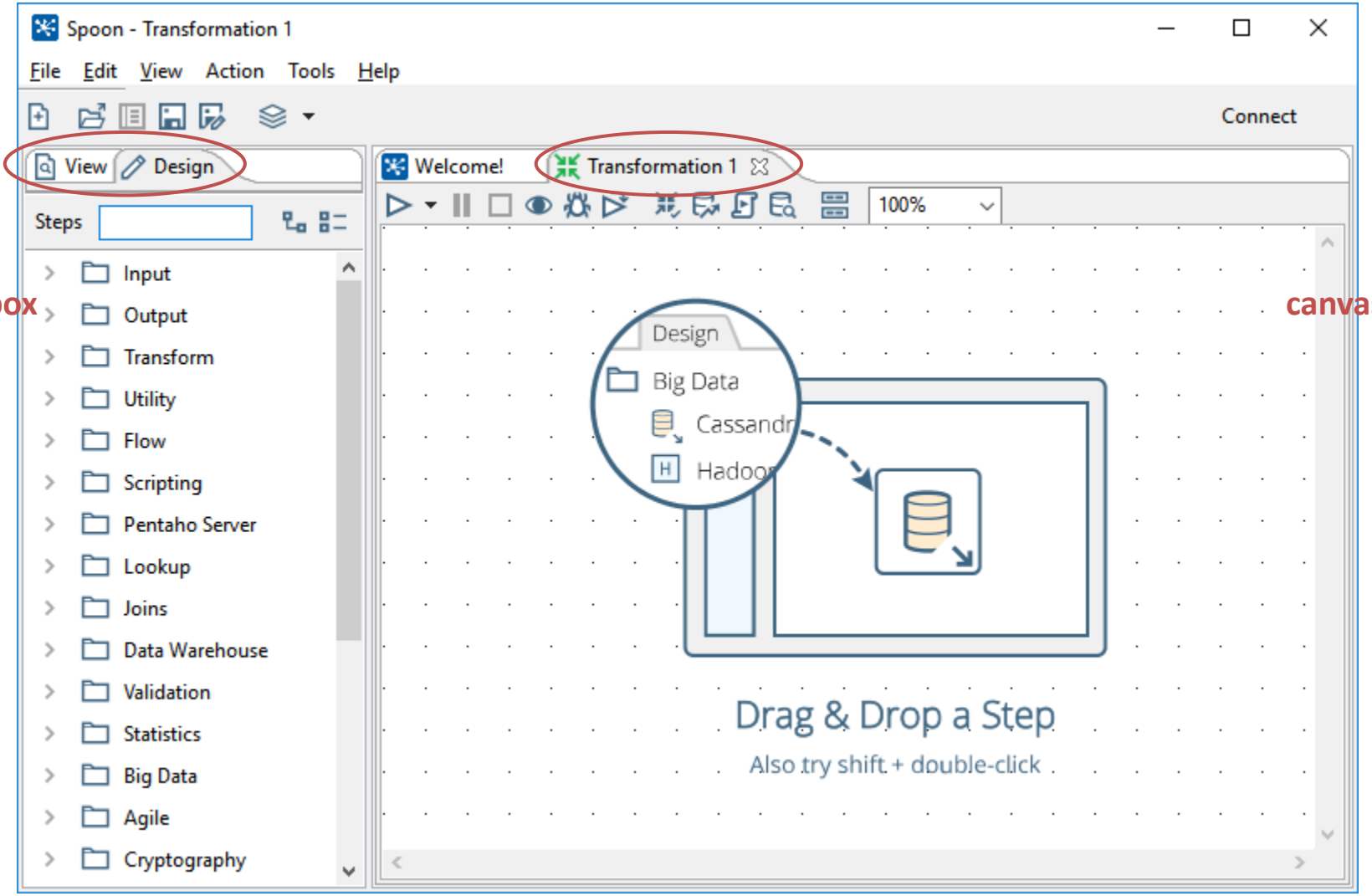
ETL tools

- ETL tools
  - what do they do?
  - how do they work?
  - how can we use them?

- The ETL tool that we will be using:
  - Pentaho Data Integration
  - also known as PDI, Kettle, or Spoon
  - competing products (e.g. SQL Server Integration Services)

Workspace

Every step in the tool must be configures

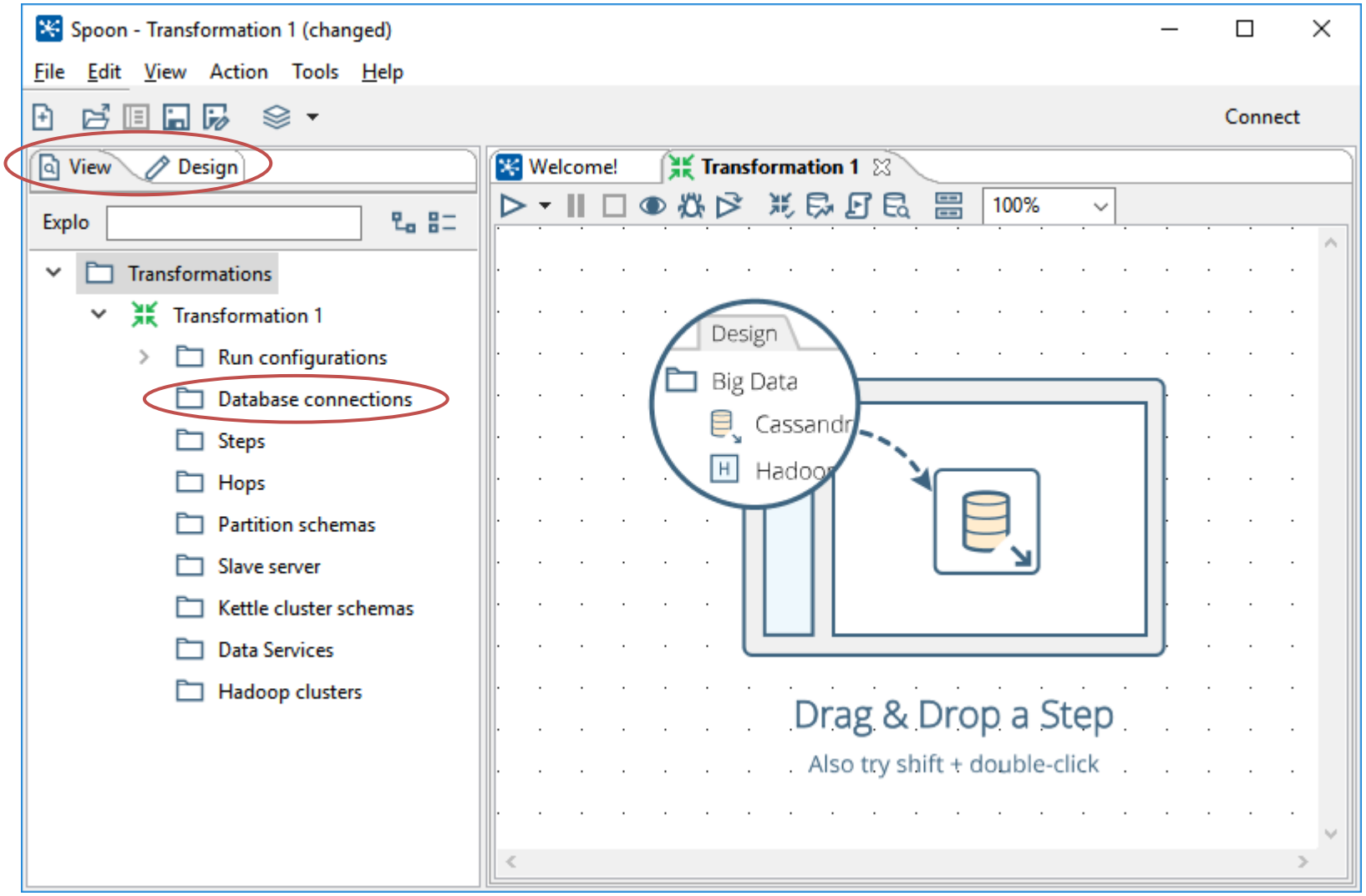


toolbox

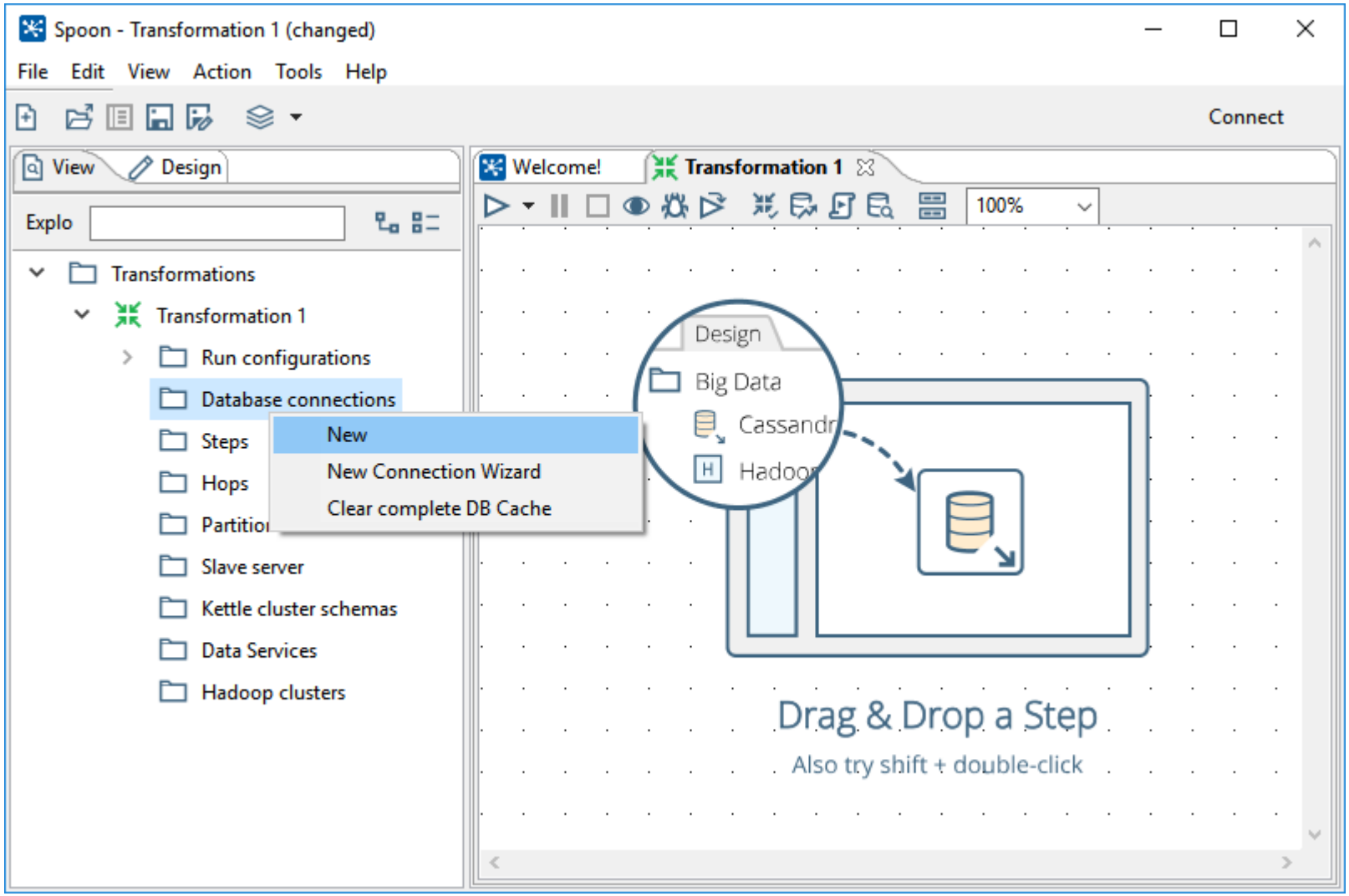
canvas

To do anything with pentaho, we need a database connection

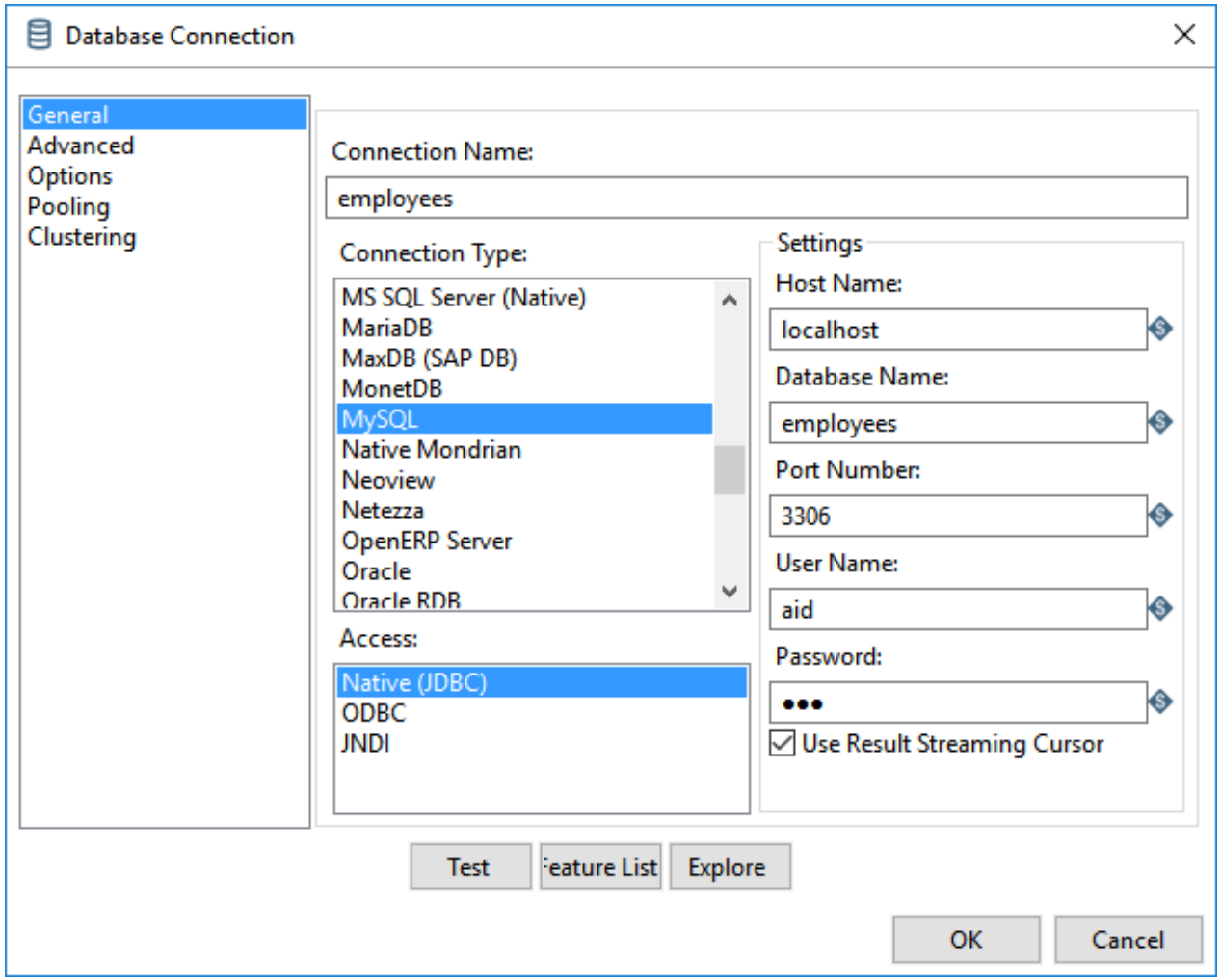
Database connections



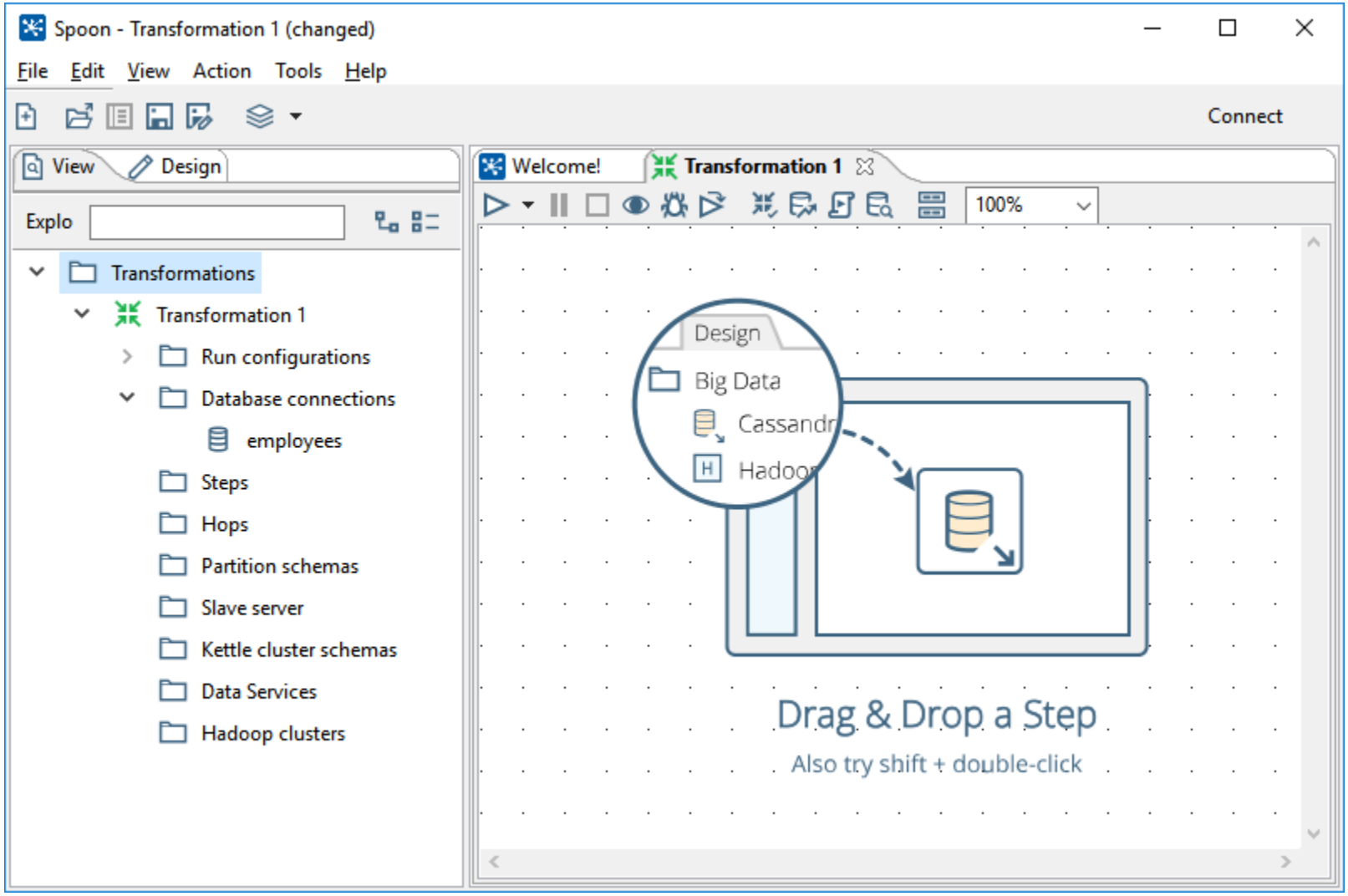
New database connection



New database connection

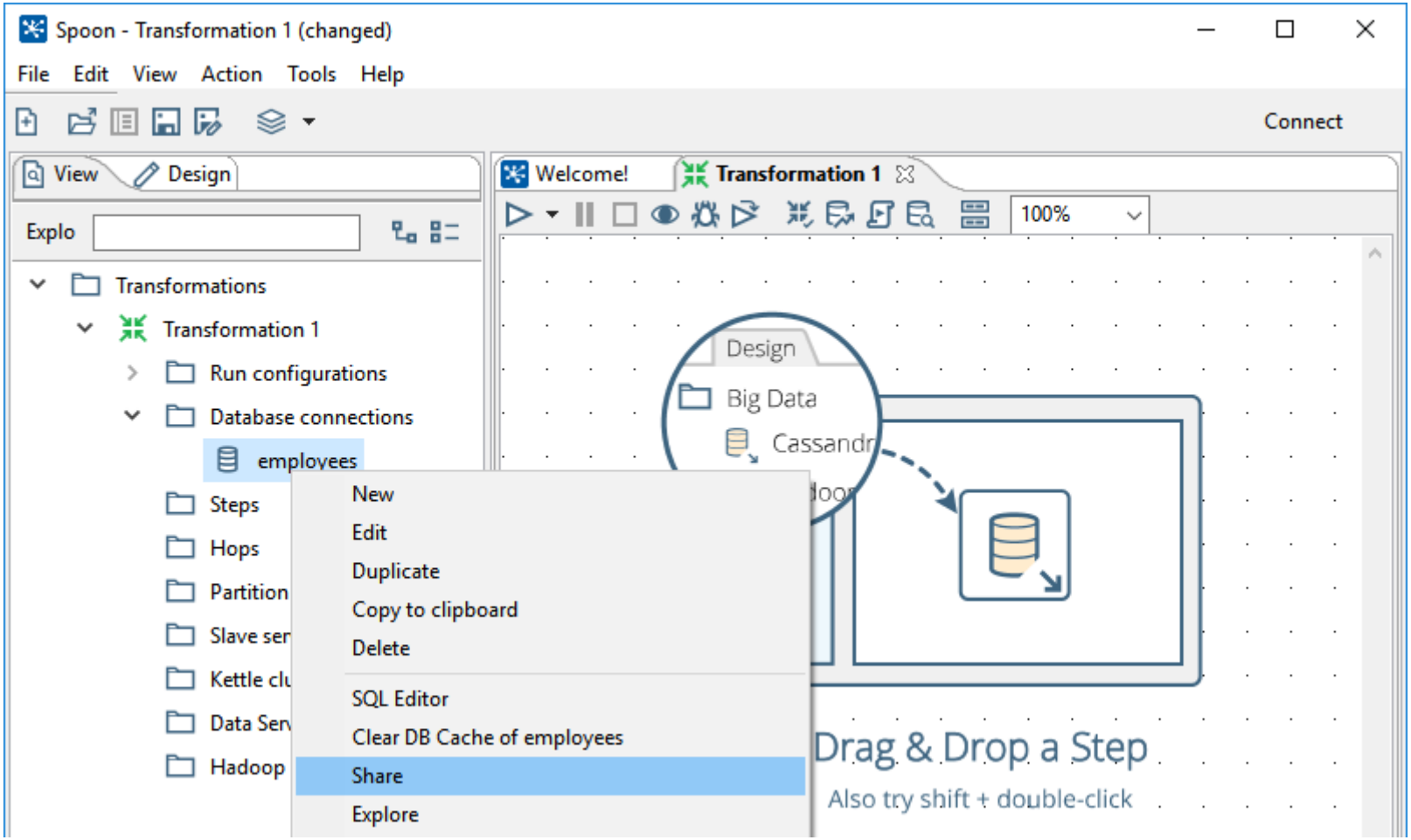


New database connection



New database connection

- Using the same connection in multiple transformations



New database connection

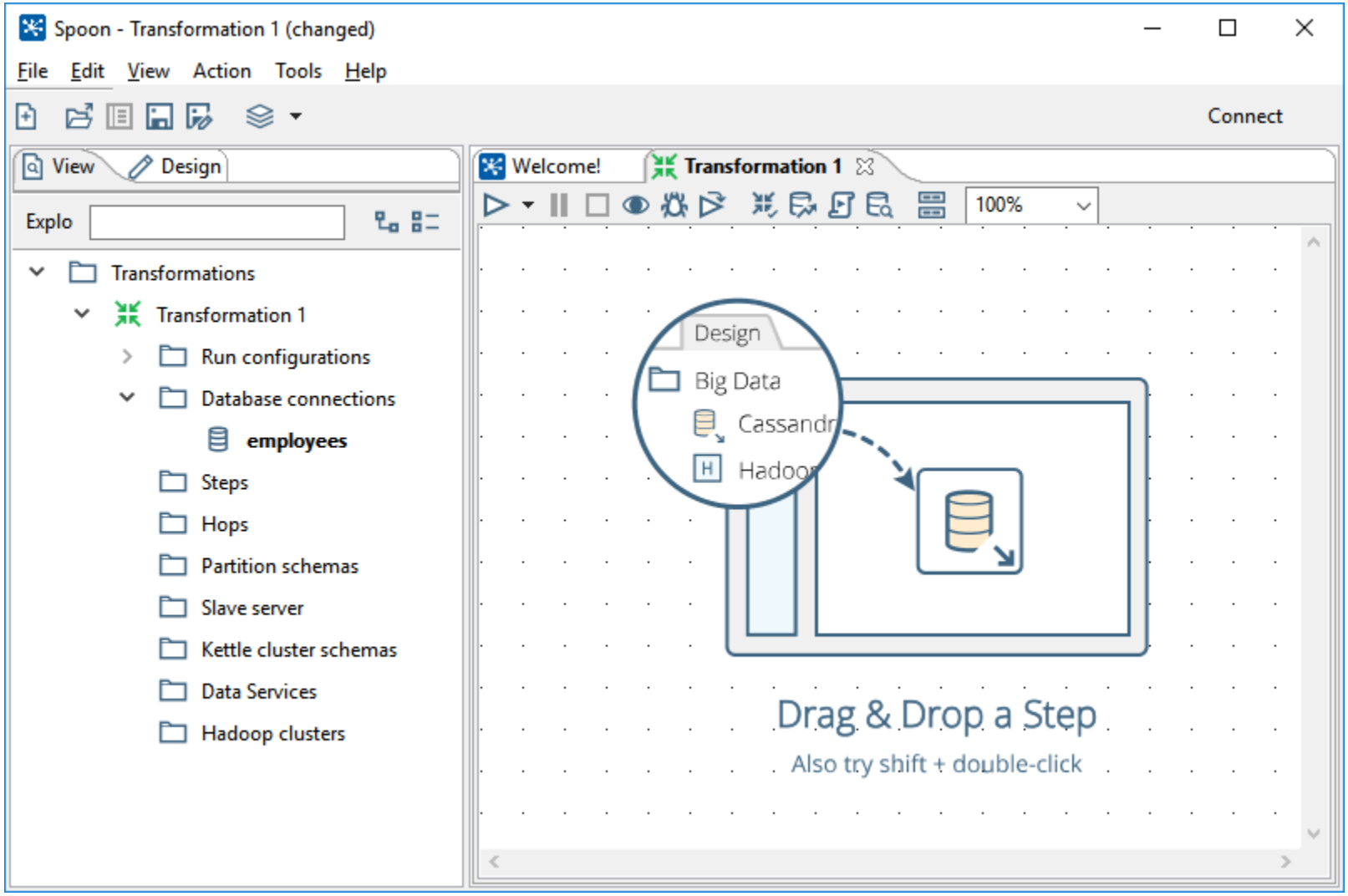


Table input

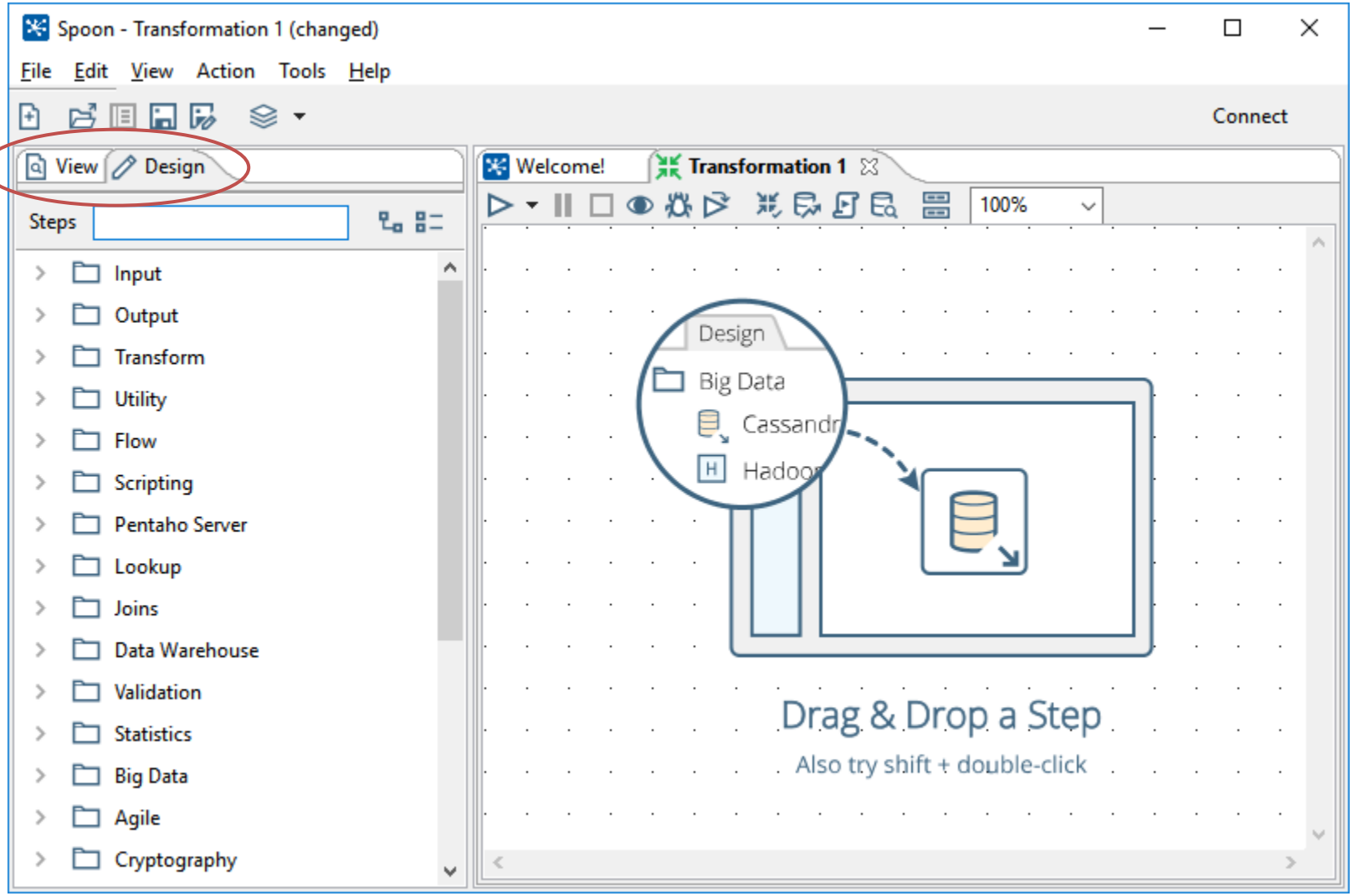

Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
Input
Output
Transform
Utility
Flow
Scripting
Pentaho Server
Lookup
Joins
Data Warehouse
Validation
Statistics
Big Data
Agile
Cryptography
Welcome!
Transformation 1
100%
Design
Big Data
Drag & Drop a Step
Also try shift + double-click

Table input

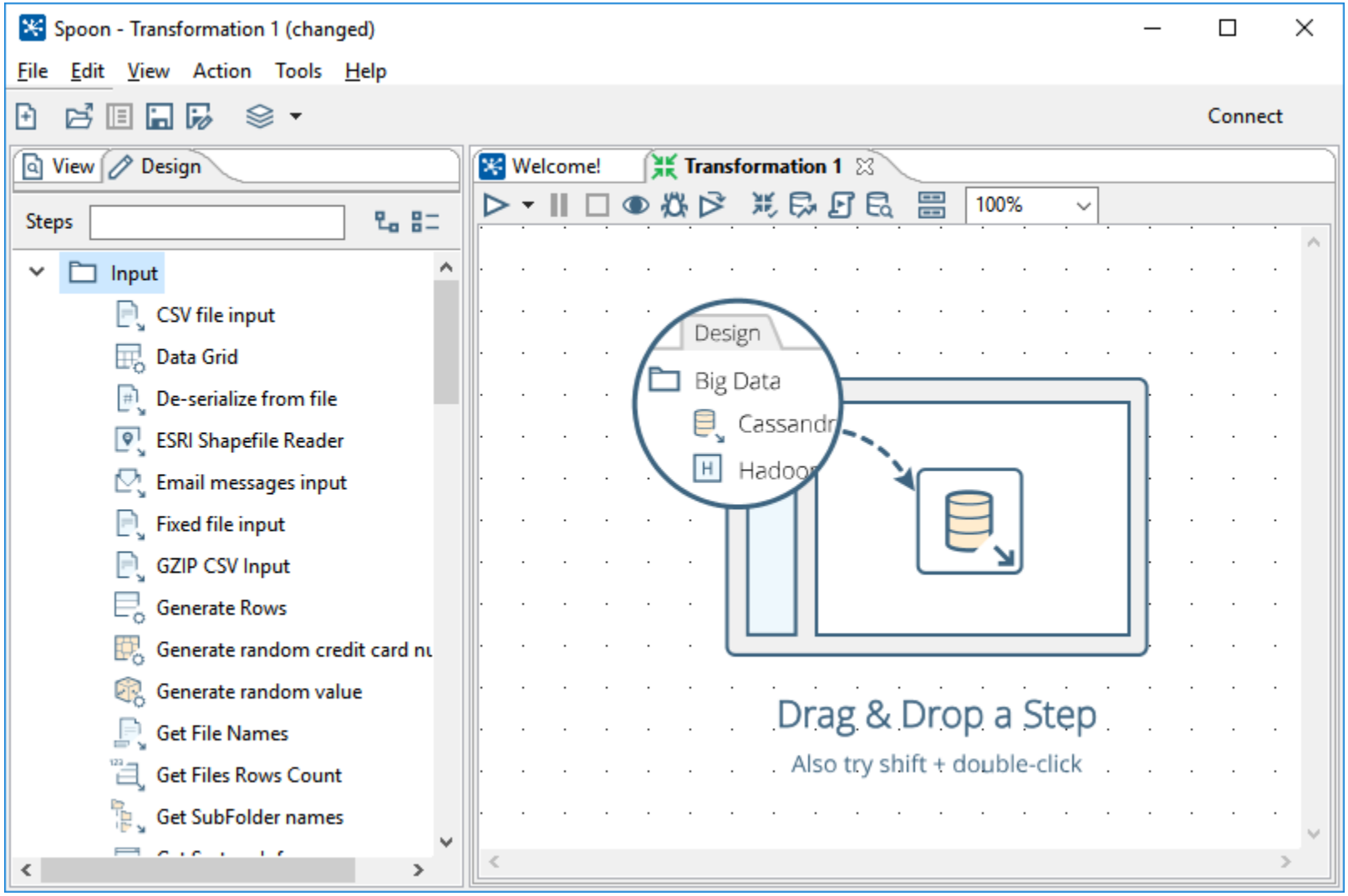

Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
Input
CSV file input
Data Grid
De-serialize from file
ESRI Shapefile Reader
Email messages input
Fixed file input
GZIP CSV Input
Generate Rows
Generate random credit card nu
Generate random value
Get File Names
Get Files Rows Count
Get SubFolder names
Welcome!
Transformation 1
100%
Design
Big Data
Drag & Drop a Step
Also try shift + double-click

Table input

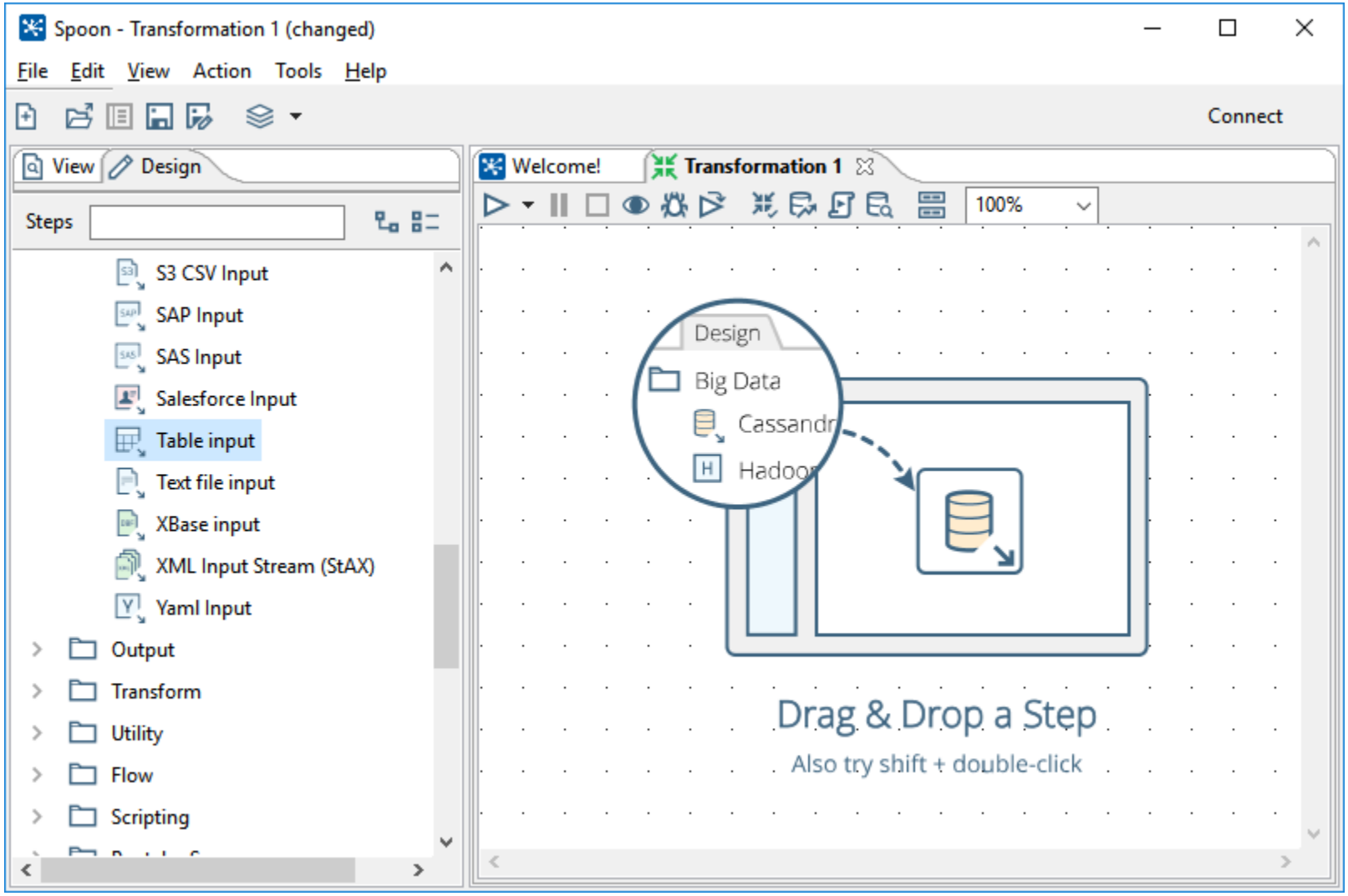

Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
S3 CSV Input
SAP Input
SAS Input
Salesforce Input
Table input
Text file input
XBase input
XML Input Stream (StAX)
Yaml Input
Output
Transform
Utility
Flow
Scripting
Welcome!
Transformation 1
100%
Design
Big Data
Drag & Drop a Step
Also try shift + double-click

Table input

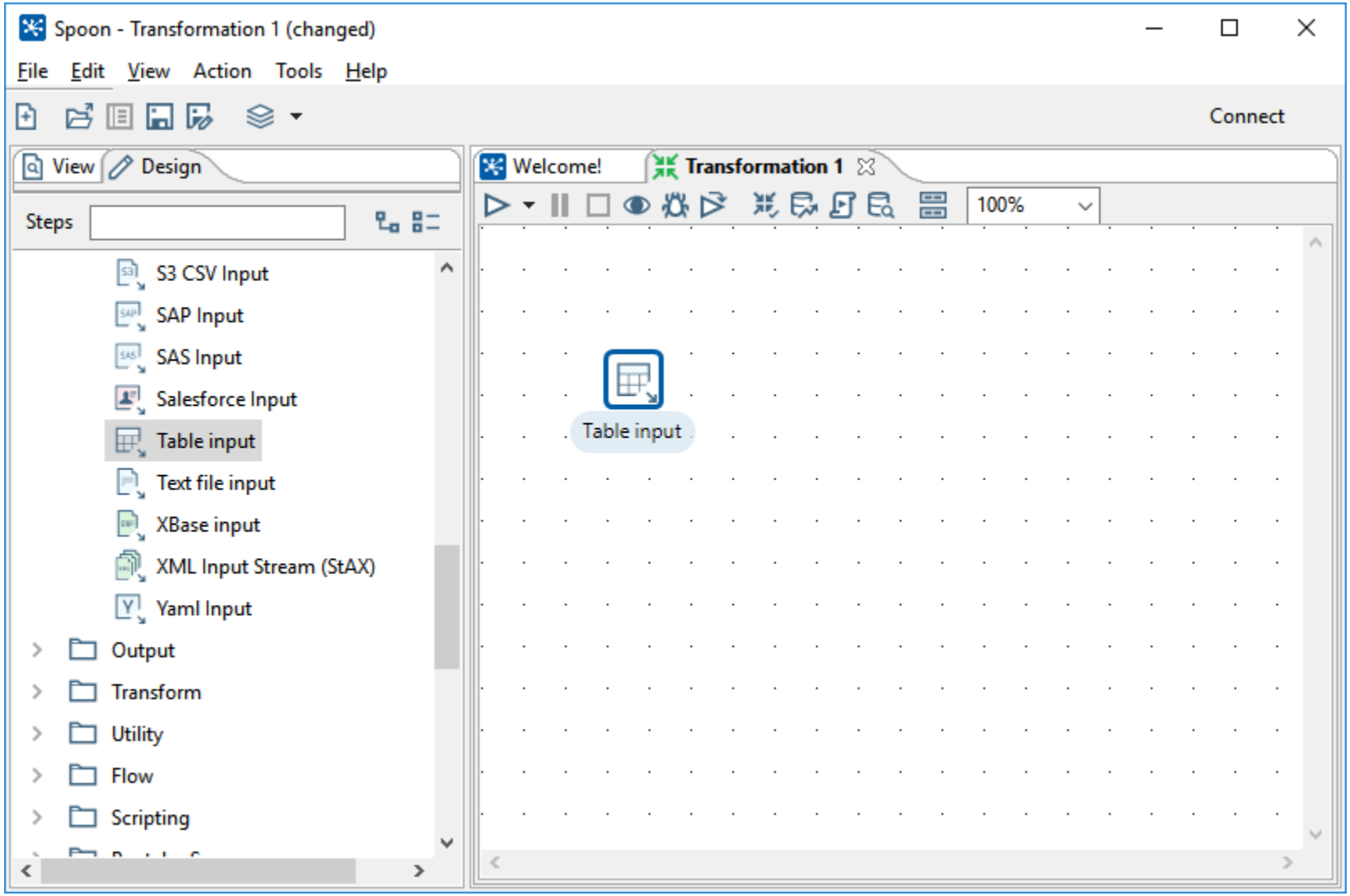

Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
S3 CSV Input
SAP Input
SAS Input
Salesforce Input
Table input
Text file input
XBase input
XML Input Stream (StAX)
Yaml Input
Output
Transform
Utility
Flow
Scripting
Welcome!
Transformation 1
100%
Table input

Table input

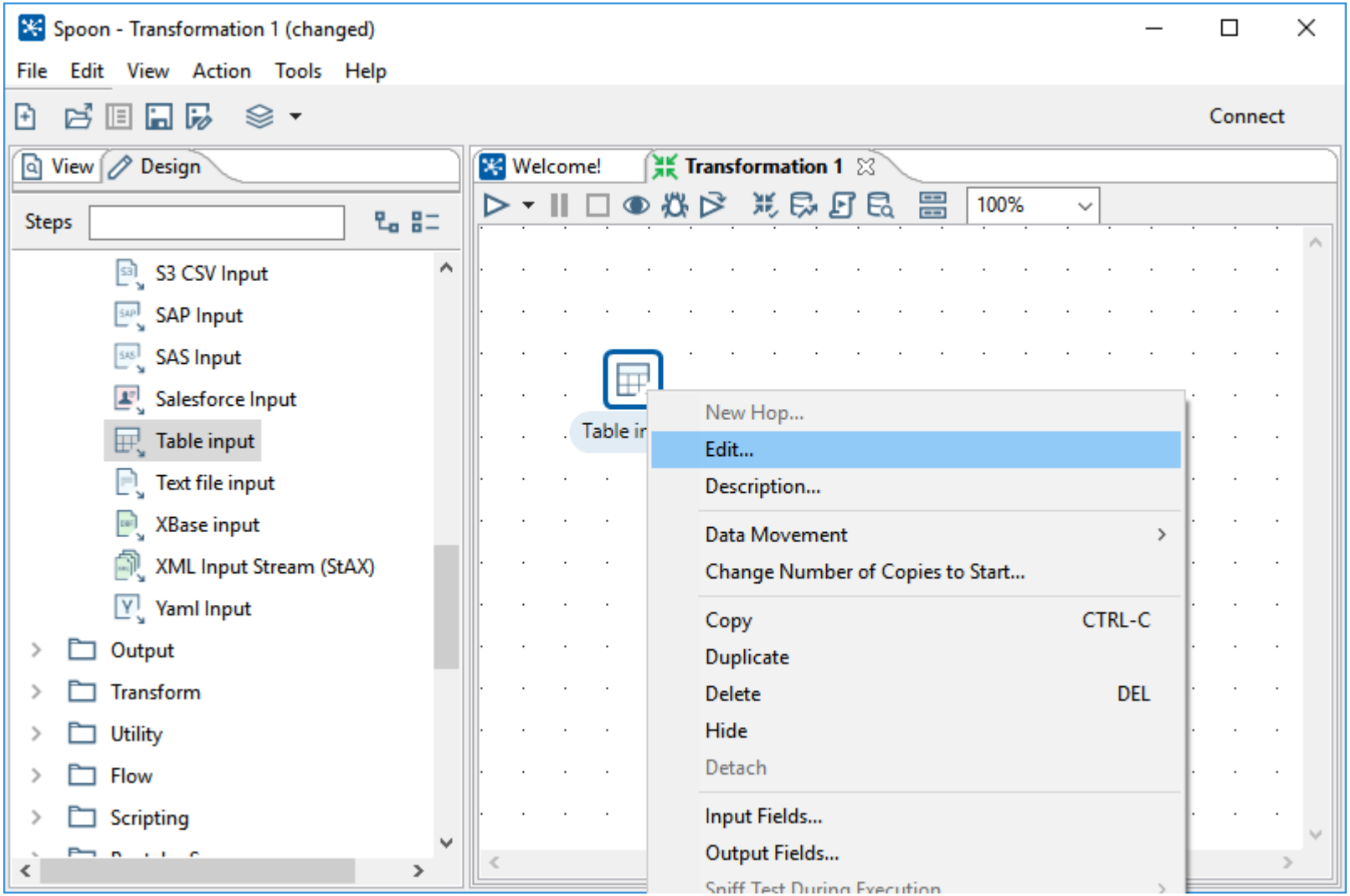

Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
S3 CSV Input
SAP Input
SAS Input
Salesforce Input
Table input
Text file input
XBase input
XML Input Stream (StAX)
Yaml Input
Output
Transform
Utility
Flow
Scripting
Welcome!
Transformation 1
100%
Table ir
New Hop...
Edit...
Description...
Data Movement
Change Number of Copies to Start...
Copy CTRL-C
Duplicate
Delete DEL
Hide
Detach
Input Fields...
Output Fields...

Table input

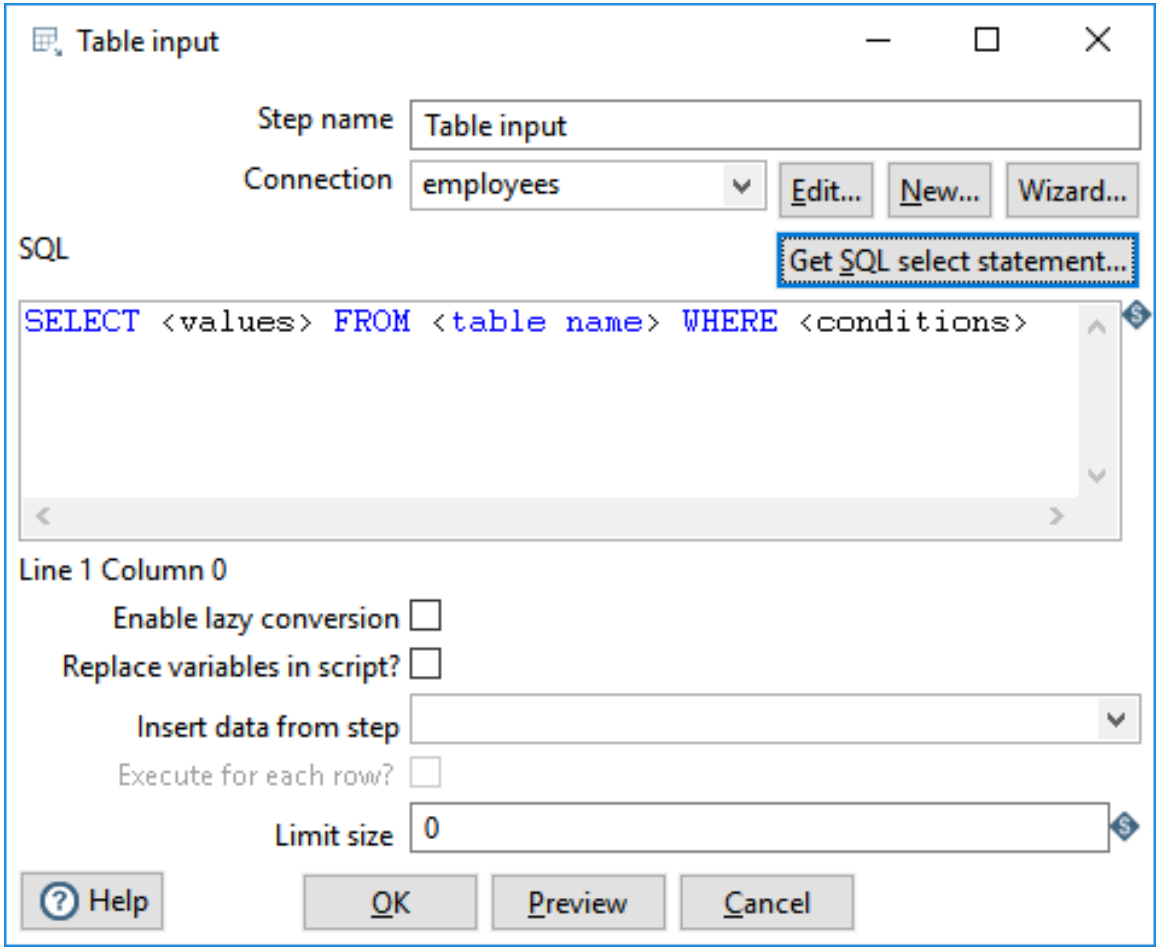


Table input
Step name
Table input
Connection
employees
Edit...
New...
Wizard...
SQL
Get SQL select statement...
SELECT <values> FROM <table name> WHERE <conditions>
Line 1 Column 0
Enable lazy conversion
Replace variables in script?
Insert data from step
Execute for each row?
Limit size
0
Help
OK
Preview
Cancel

Table input

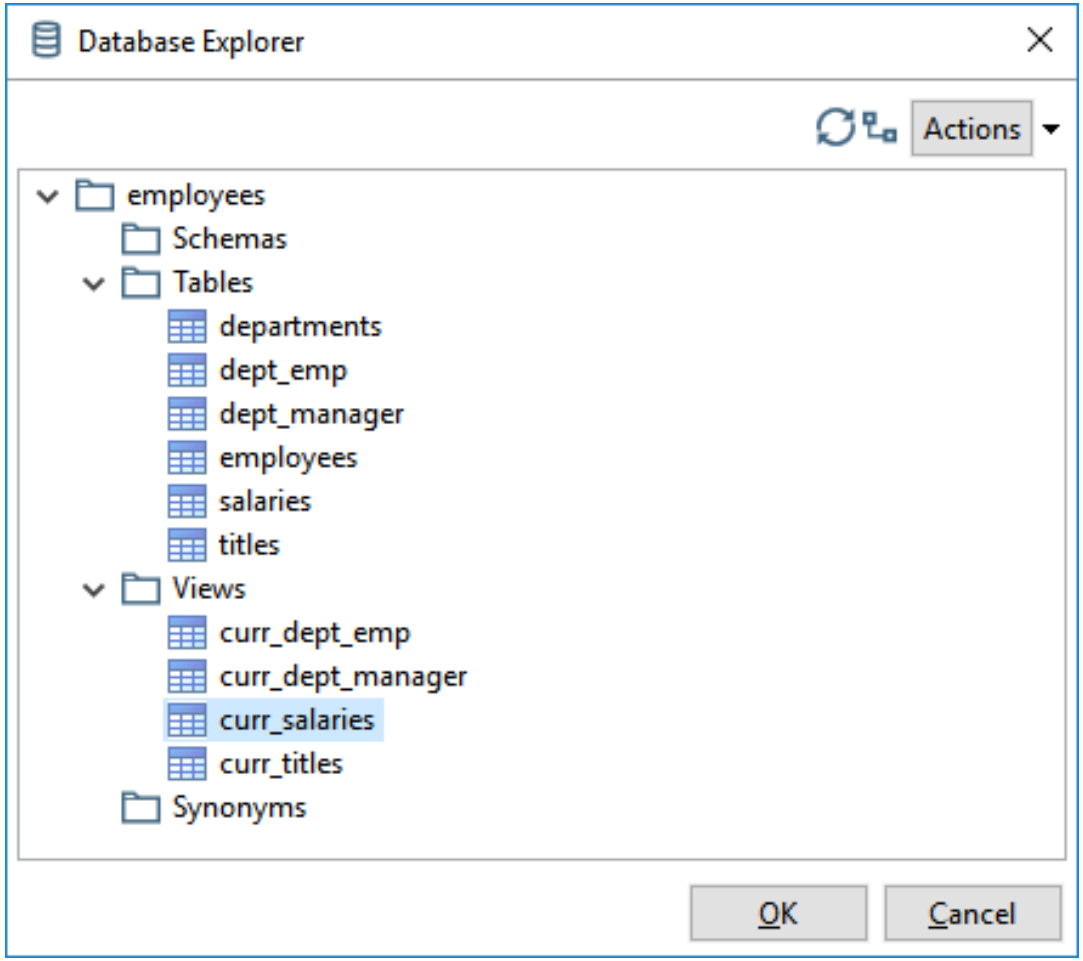

Database Explorer
Actions
employees
Schemas
Tables
departments
dept_emp
dept_manager
employees
salaries
titles
Views
curr_dept_emp
curr_dept_manager
curr_salaries
curr_titles
Synonyms
OK
Cancel

Table input

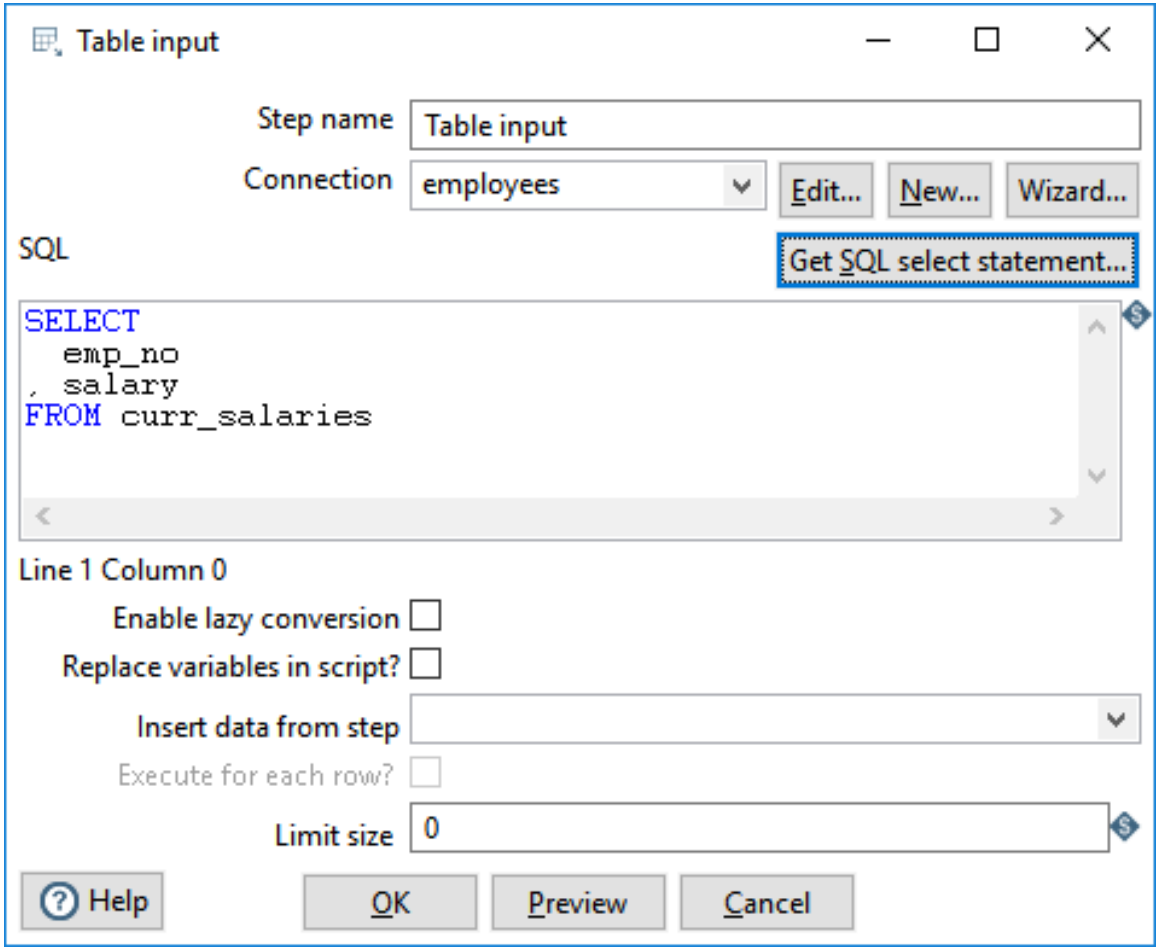


Table input
Step name
Table input
Connection
employees
Edit...
New...
Wizard...
SQL
Get SQL select statement...
SELECT
emp_no
, salary
FROM curr_salaries
Line 1 Column 0
Enable lazy conversion
Replace variables in script?
Insert data from step
Execute for each row?
Limit size
0
Help
OK
Preview
Cancel

Table input

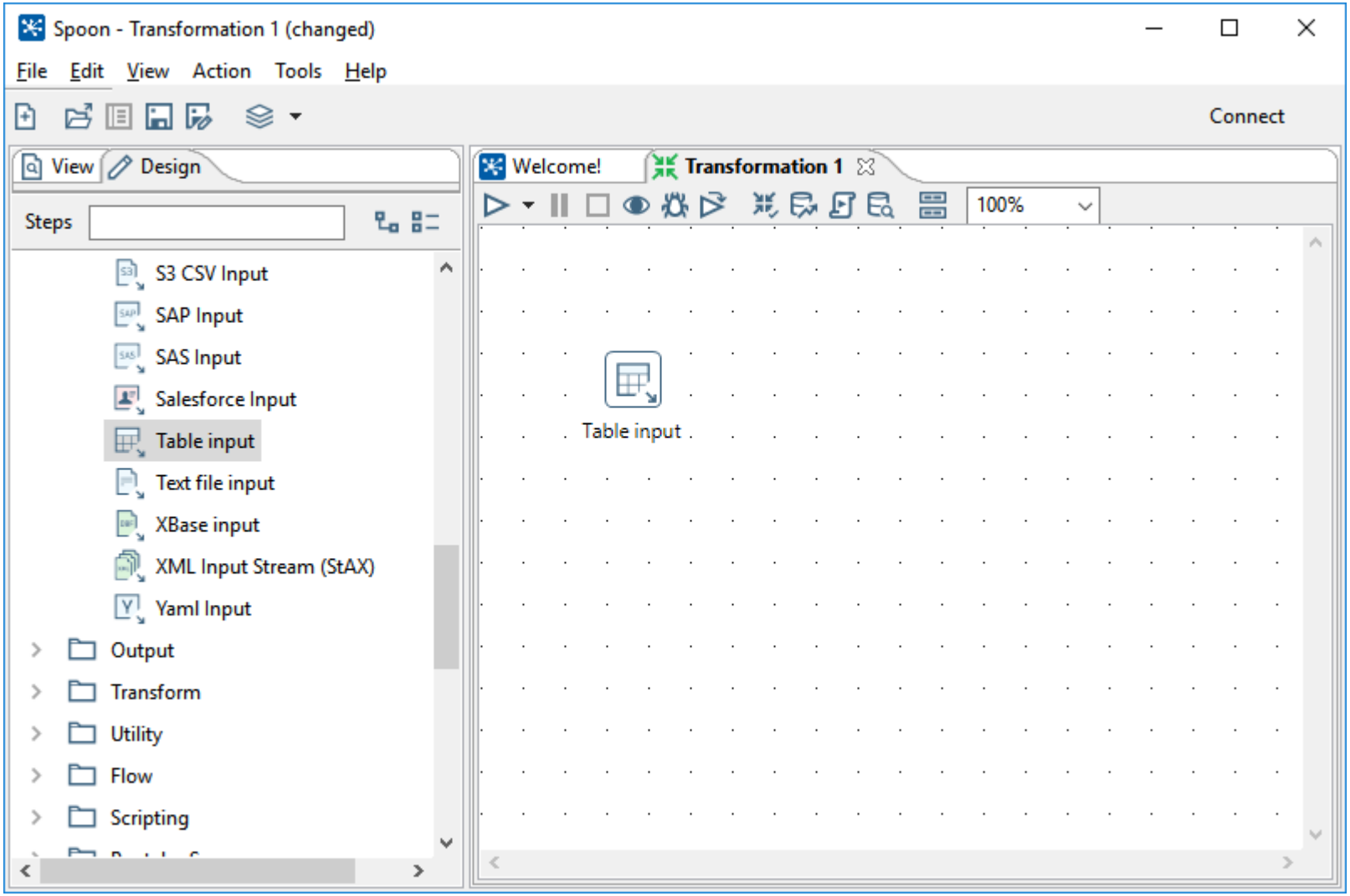

Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
S3 CSV Input
SAP Input
SAS Input
Salesforce Input
Table input
Text file input
XBase input
XML Input Stream (StAX)
Yaml Input
Output
Transform
Utility
Flow
Scripting
Welcome!
Transformation 1
100%
Table input

Table input

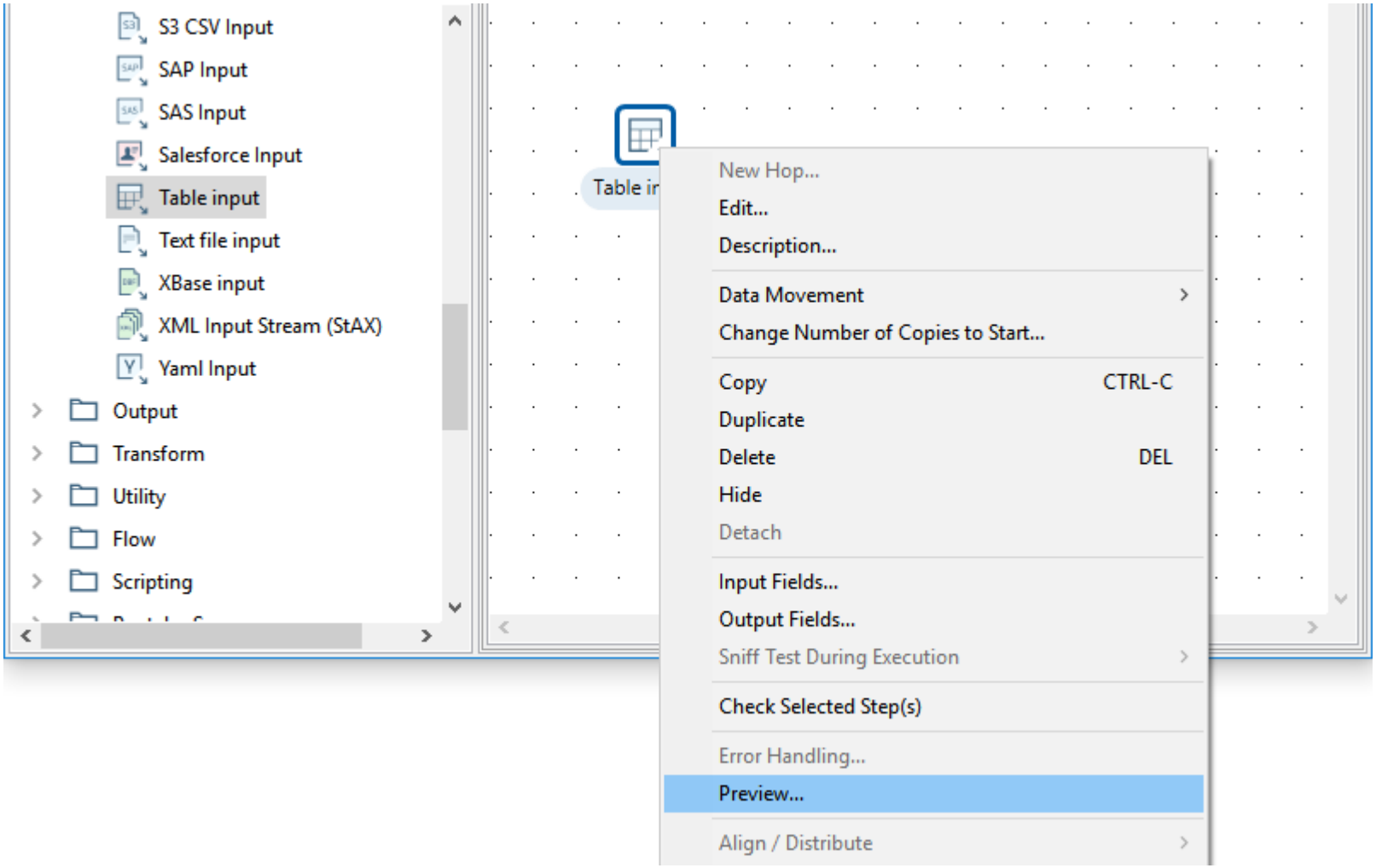

S3 CSV Input
SAP Input
SAS Input
Salesforce Input
Table input
Text file input
XBase input
XML Input Stream (StAX)
Yaml Input
Output
Transform
Utility
Flow
Scripting
Table ir
New Hop...
Edit...
Description...
Data Movement
Change Number of Copies to Start...
Copy CTRL-C
Duplicate
Delete DEL
Hide
Detach
Input Fields...
Output Fields...
Sniff Test During Execution
Check Selected Step(s)
Error Handling...
Preview...
Align / Distribute

Table input

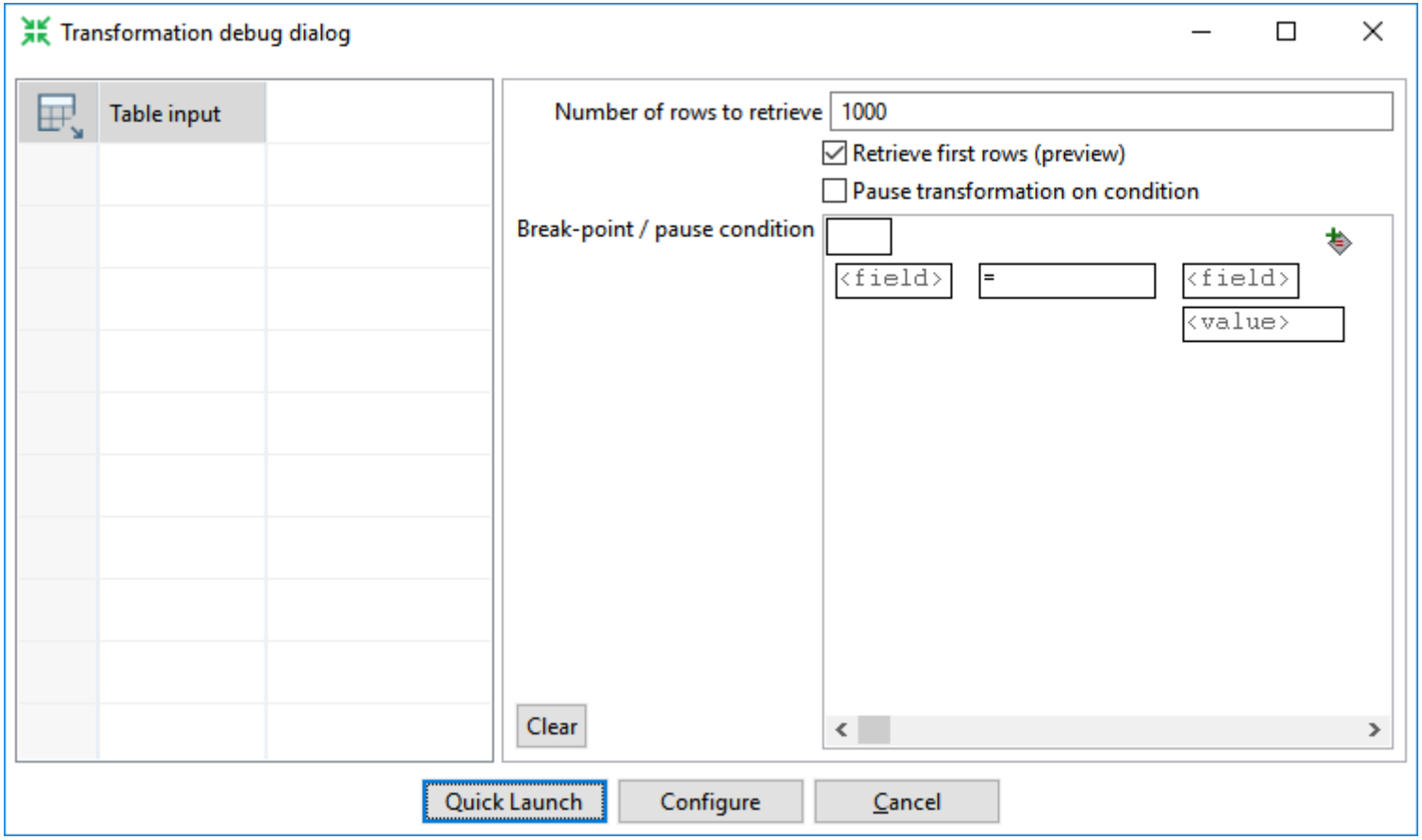
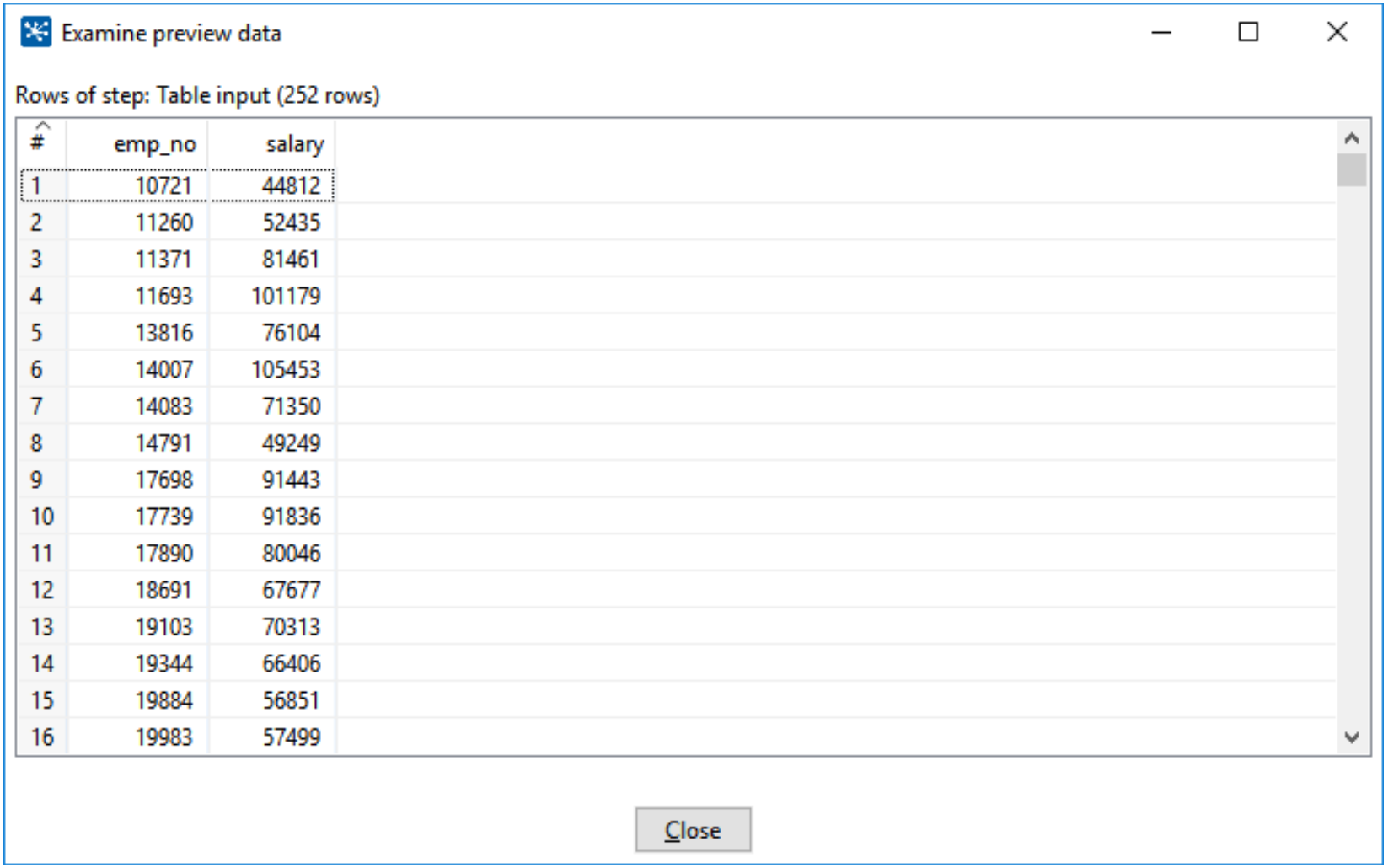
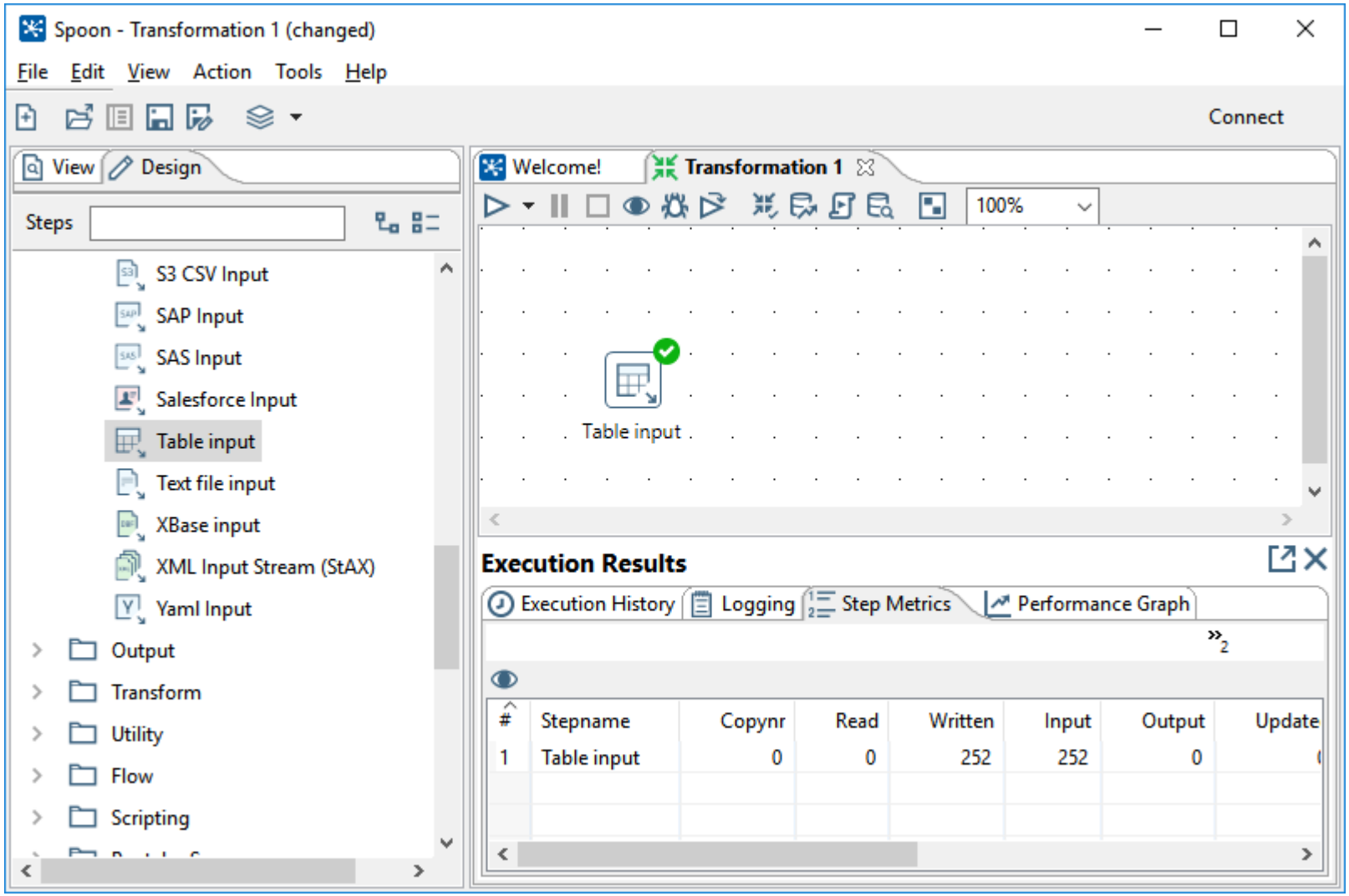

Transformation debug dialog
Table input
Number of rows to retrieve
1000
Retrieve first rows (preview)
Pause transformation on condition
Break-point / pause condition
<field>
=
<field>
<value>
Clear
Quick Launch
Configure
Cancel

Table input

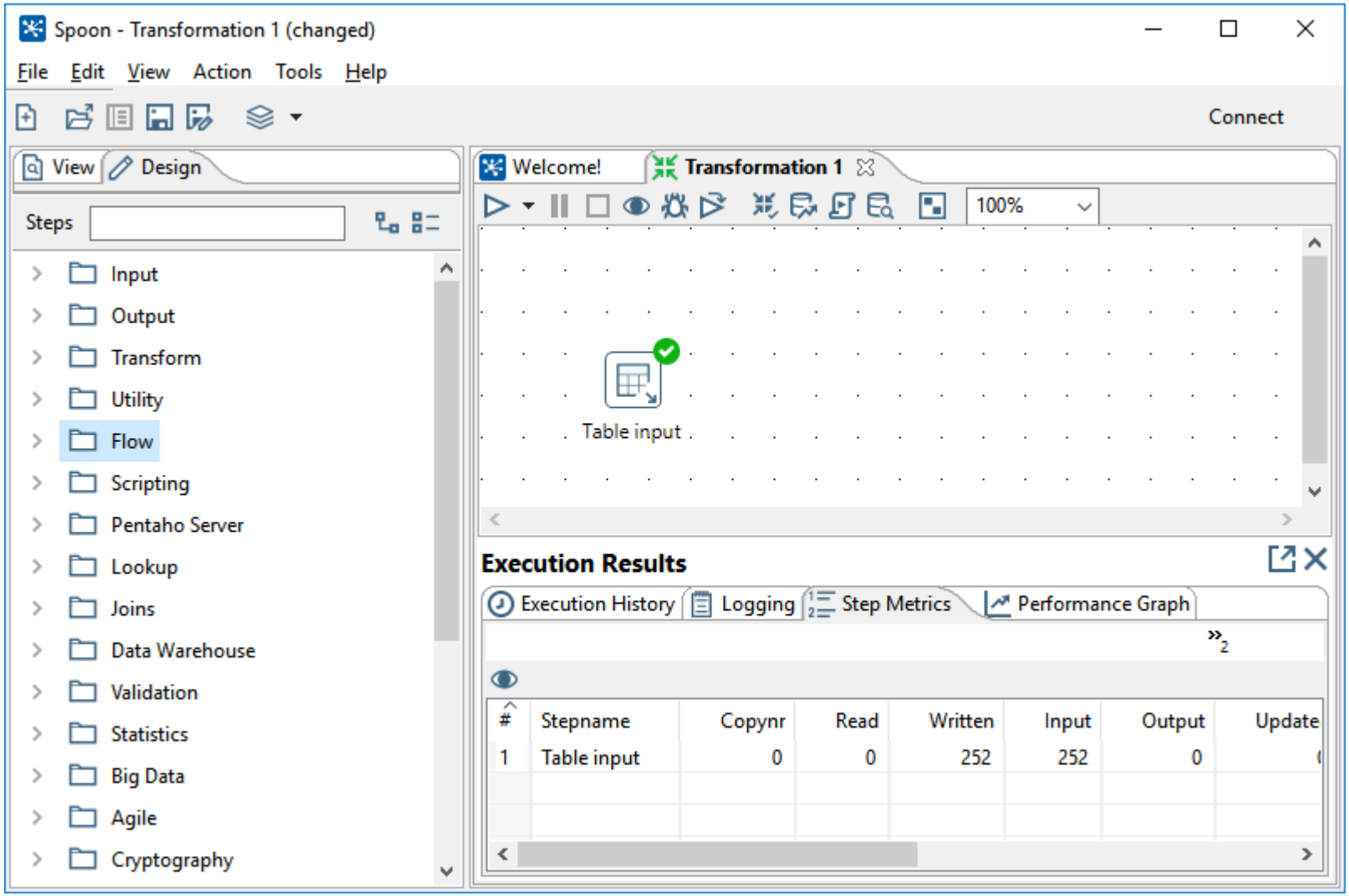


#	emp_no	salary
1	10721	44812
2	11260	52435
3	11371	81461
4	11693	101179
5	13816	76104
6	14007	105453
7	14083	71350
8	14791	49249
9	17698	91443
10	17739	91836
11	17890	80046
12	18691	67677
13	19103	70313
14	19344	66406
15	19884	56851
16	19983	57499

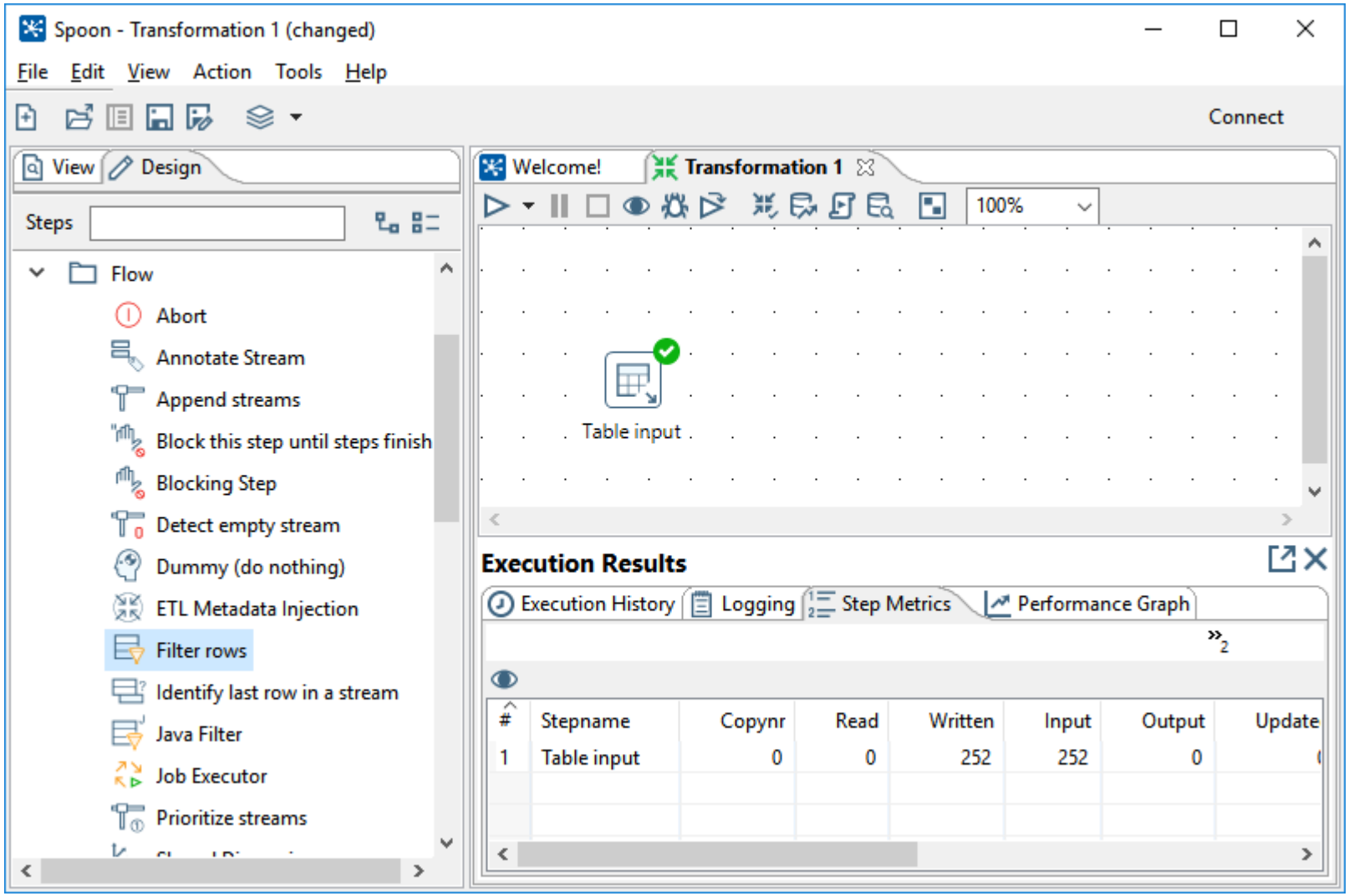
Table input


Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
S3 CSV Input
SAP Input
SAS Input
Salesforce Input
Table input
Text file input
XBase input
XML Input Stream (StAX)
Yaml Input
Output
Transform
Utility
Flow
Scripting
Welcome!
Transformation 1
100%
Table input
Execution Results
Execution History
Logging
Step Metrics
Performance Graph
Stepname Copynr Read Written Input Output Update
1 Table input 0 0 252 252 0

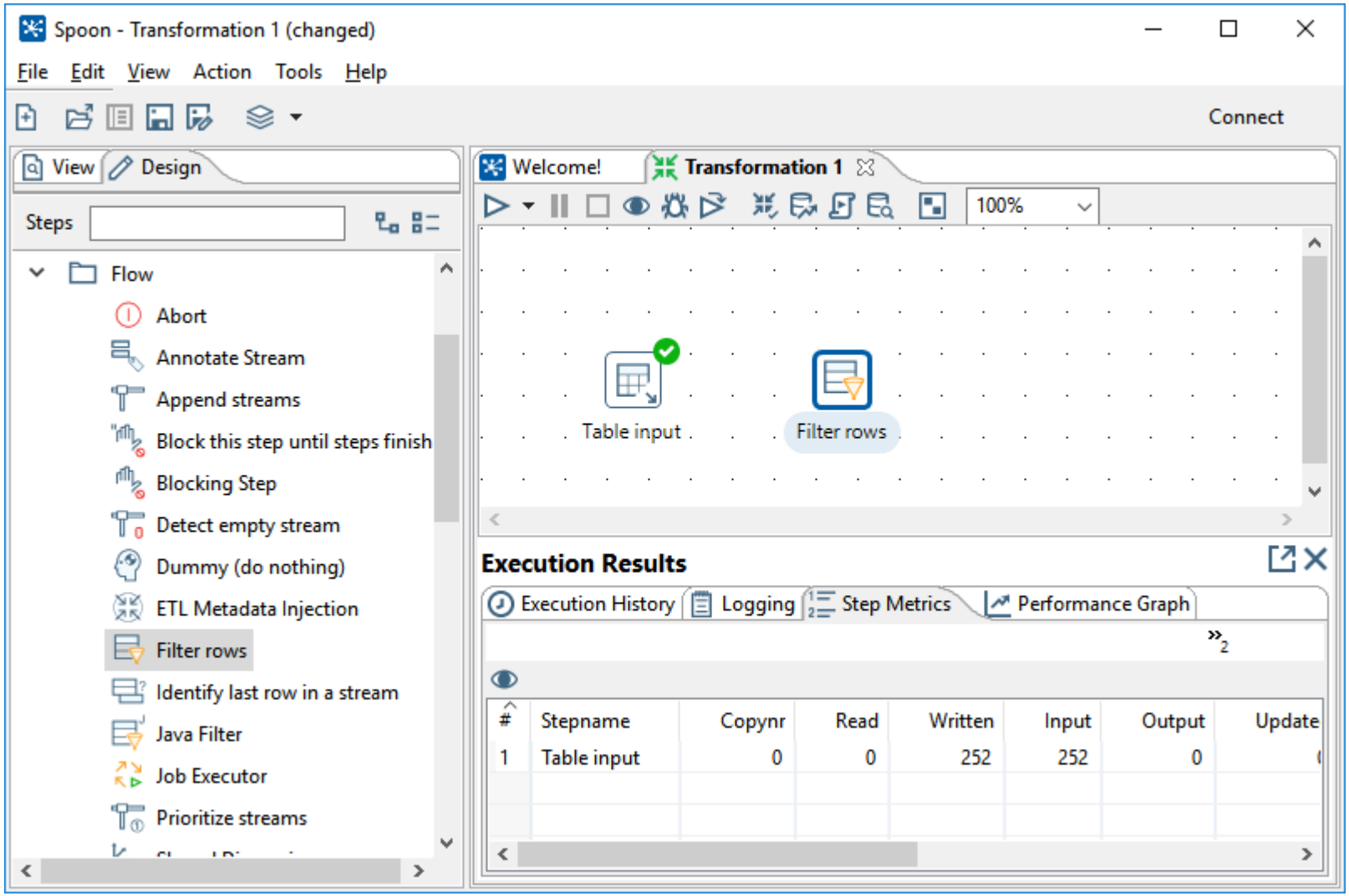
Filter rows


Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
Input
Output
Transform
Utility
Flow
Scripting
Pentaho Server
Lookup
Joins
Data Warehouse
Validation
Statistics
Big Data
Agile
Cryptography
Welcome!
Transformation 1
100%
Table input
Execution Results
Execution History
Logging
Step Metrics
Performance Graph
Stepname Copynr Read Written Input Output Update
1 Table input 0 0 252 252 0

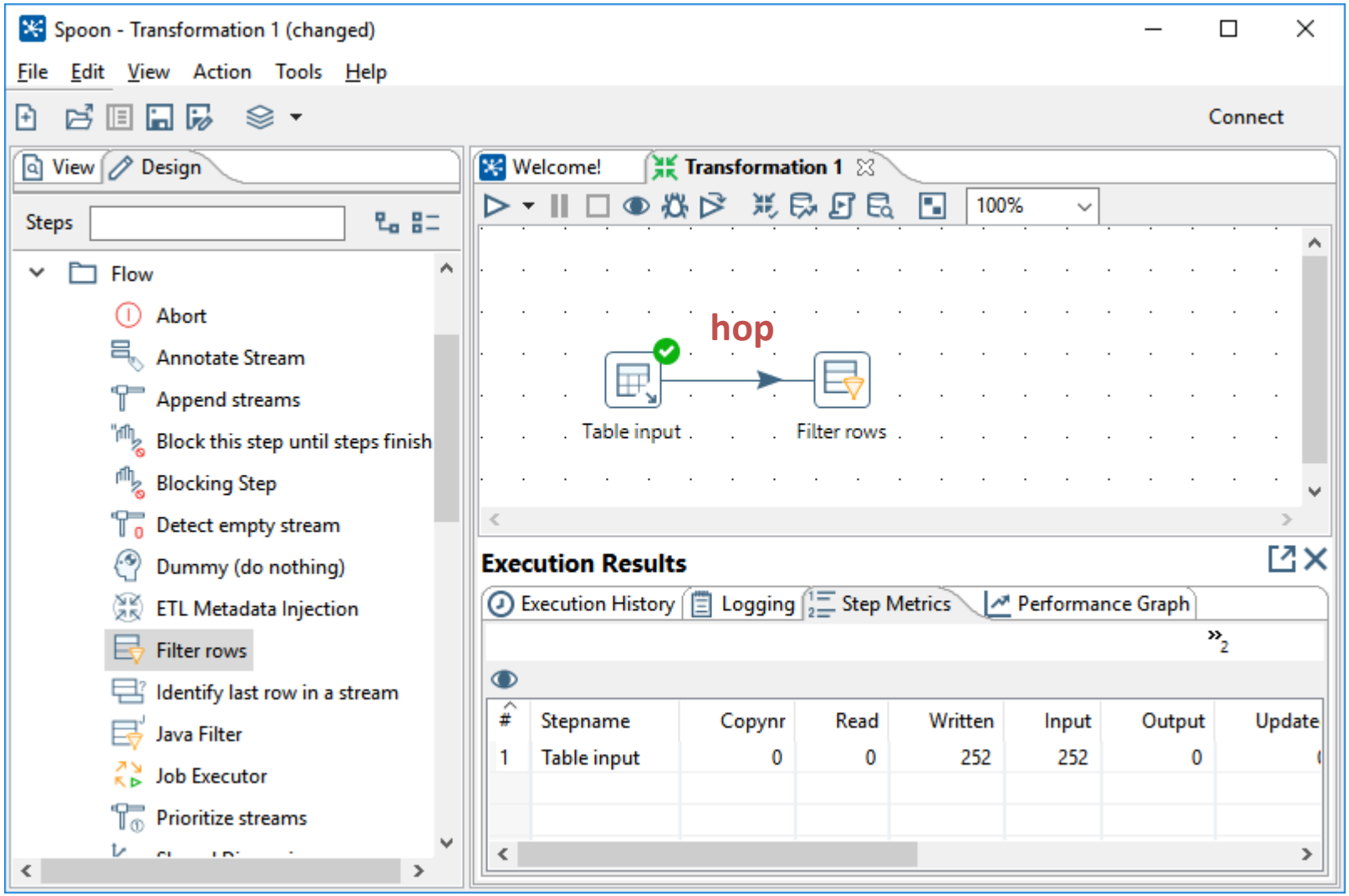
Filter rows


Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
Flow
Abort
Annotate Stream
Append streams
Block this step until steps finish
Blocking Step
Detect empty stream
Dummy (do nothing)
ETL Metadata Injection
Filter rows
Identify last row in a stream
Java Filter
Job Executor
Prioritize streams
Welcome!
Transformation 1
100%
Table input
Execution Results
Execution History
Logging
Step Metrics
Performance Graph
Stepname Copynr Read Written Input Output Update
1 Table input 0 0 252 252 0

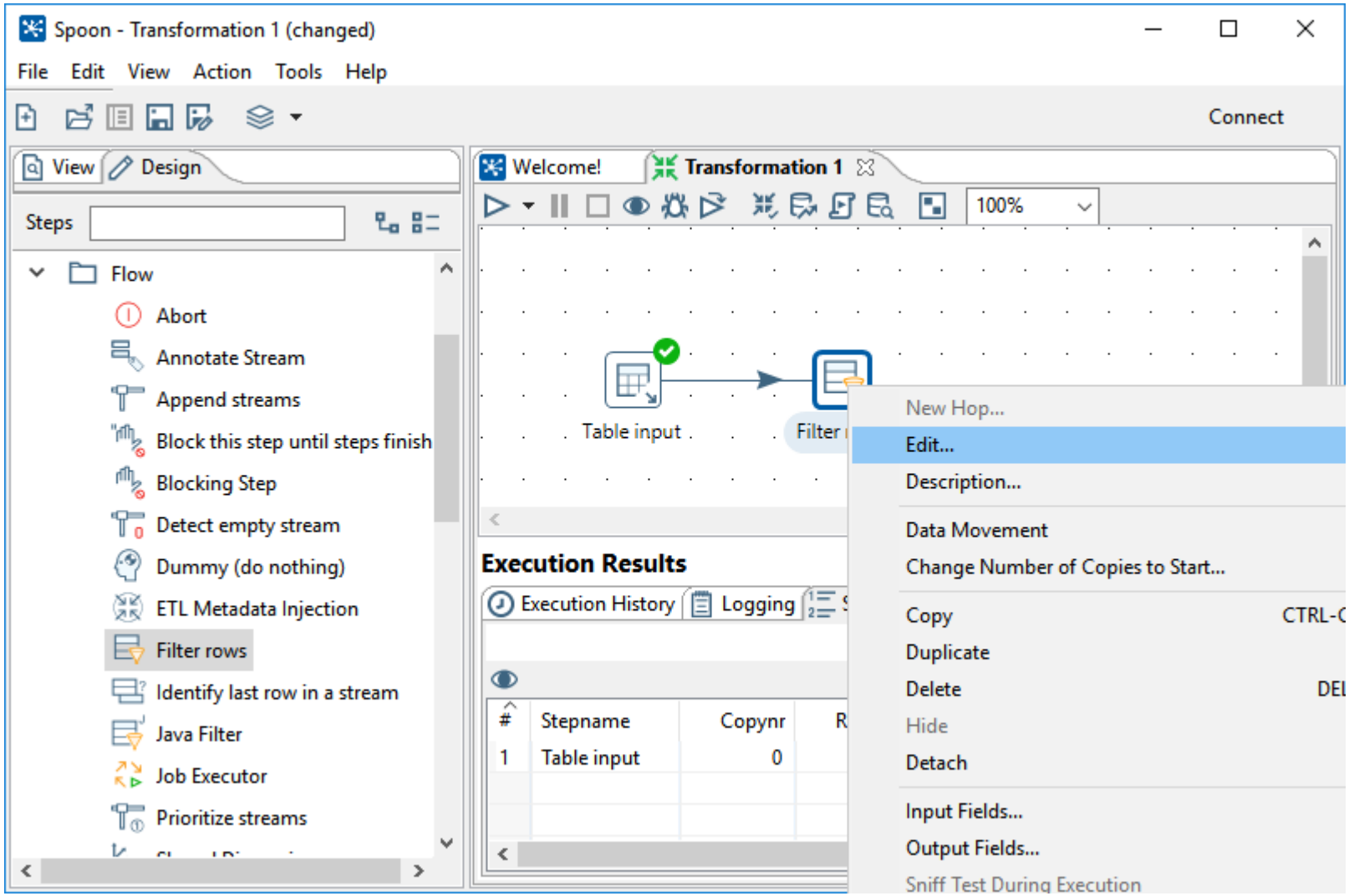
Filter rows


Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
Flow
Abort
Annotate Stream
Append streams
Block this step until steps finish
Blocking Step
Detect empty stream
Dummy (do nothing)
ETL Metadata Injection
Filter rows
Identify last row in a stream
Java Filter
Job Executor
Prioritize streams
Welcome!
Transformation 1
100%
Table input
Filter rows
Execution Results
Execution History
Logging
Step Metrics
Performance Graph
Stepname Copynr Read Written Input Output Update
1 Table input 0 0 252 252 0

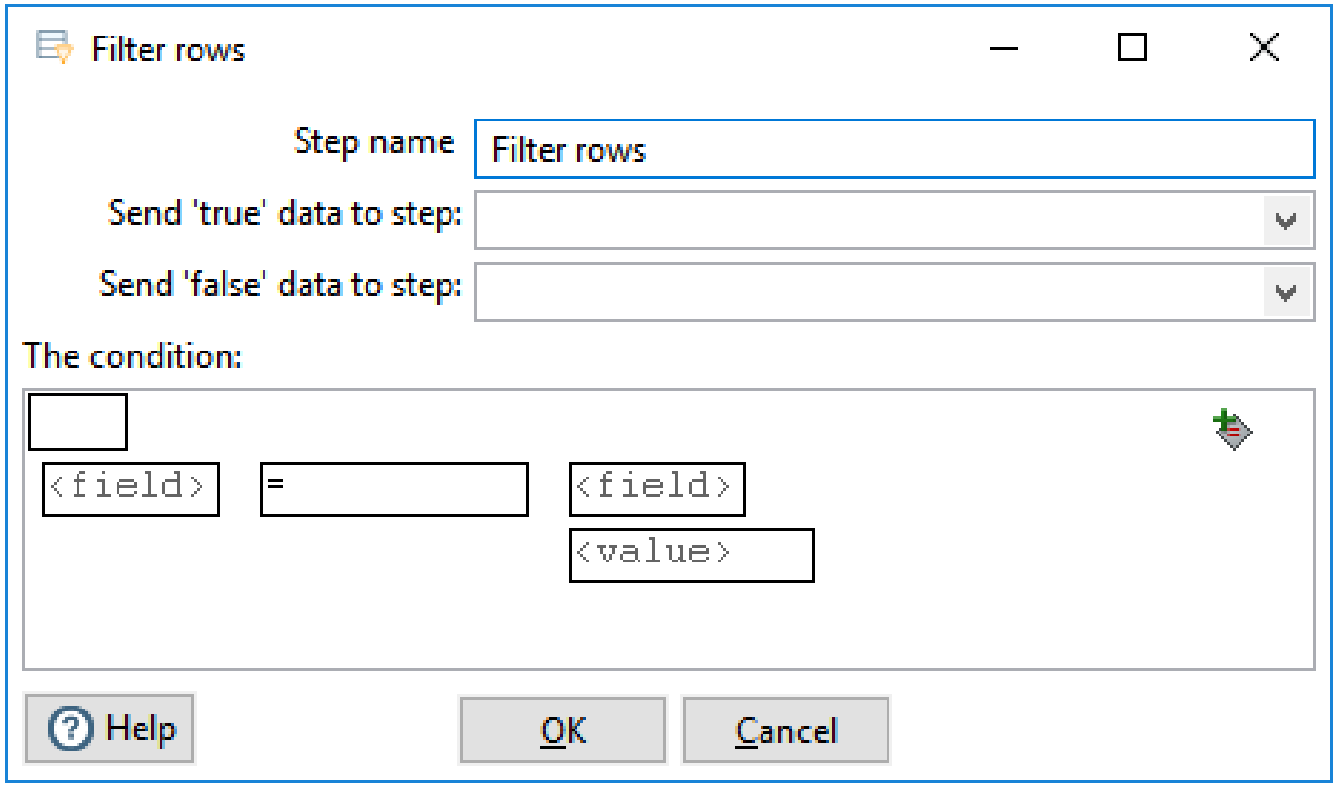
Filter rows


Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
Flow
Abort
Annotate Stream
Append streams
Block this step until steps finish
Blocking Step
Detect empty stream
Dummy (do nothing)
ETL Metadata Injection
Filter rows
Identify last row in a stream
Java Filter
Job Executor
Prioritize streams
Welcome!
Transformation 1
100%
hop
Table input
Filter rows
Execution Results
Execution History
Logging
Step Metrics
Performance Graph
Stepname Copynr Read Written Input Output Update
1 Table input 0 0 252 252 0

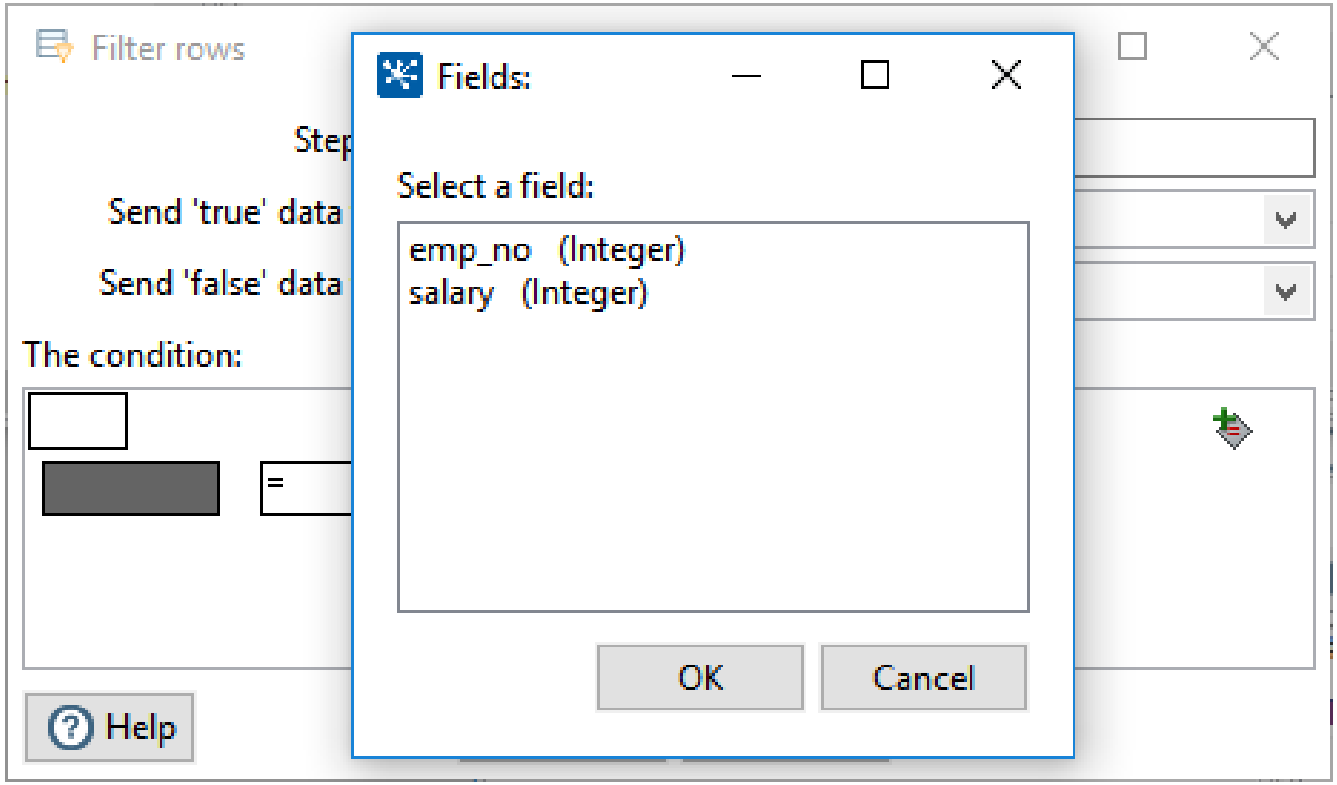
Filter rows


Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
Flow
Abort
Annotate Stream
Append streams
Block this step until steps finish
Blocking Step
Detect empty stream
Dummy (do nothing)
ETL Metadata Injection
Filter rows
Identify last row in a stream
Java Filter
Job Executor
Prioritize streams
Welcome!
Transformation 1
100%
Table input
Filter
Execution Results
Execution History
Logging
Stepname Copynr
1 Table input 0
New Hop...
Edit...
Description...
Data Movement
Change Number of Copies to Start...
Copy CTRL-C
Duplicate
Delete DEL
Hide
Detach
Input Fields...
Output Fields...
Sniff Test During Execution

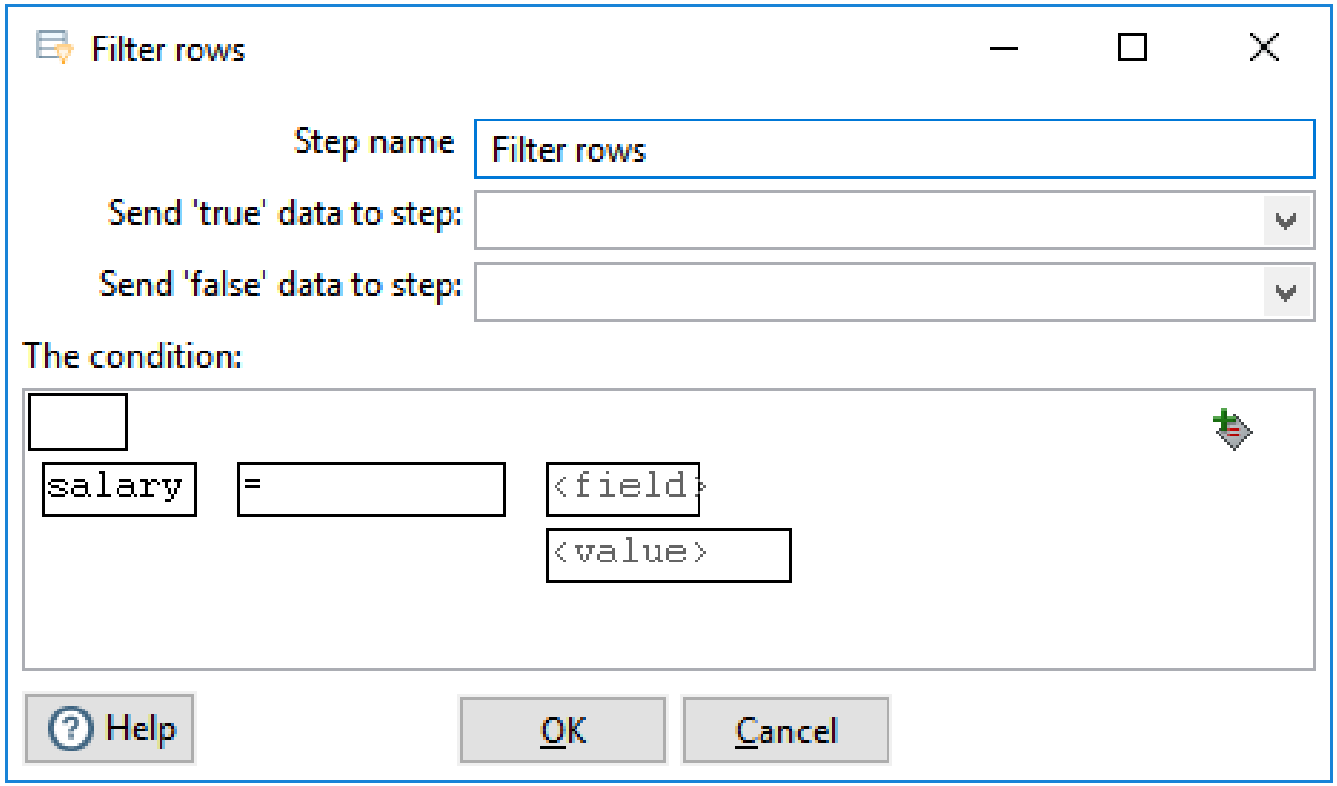
Filter rows


Filter rows
Step name
Filter rows
Send 'true' data to step:
Send 'false' data to step:
The condition:
<field>
=
<field>
<value>
Help
OK
Cancel

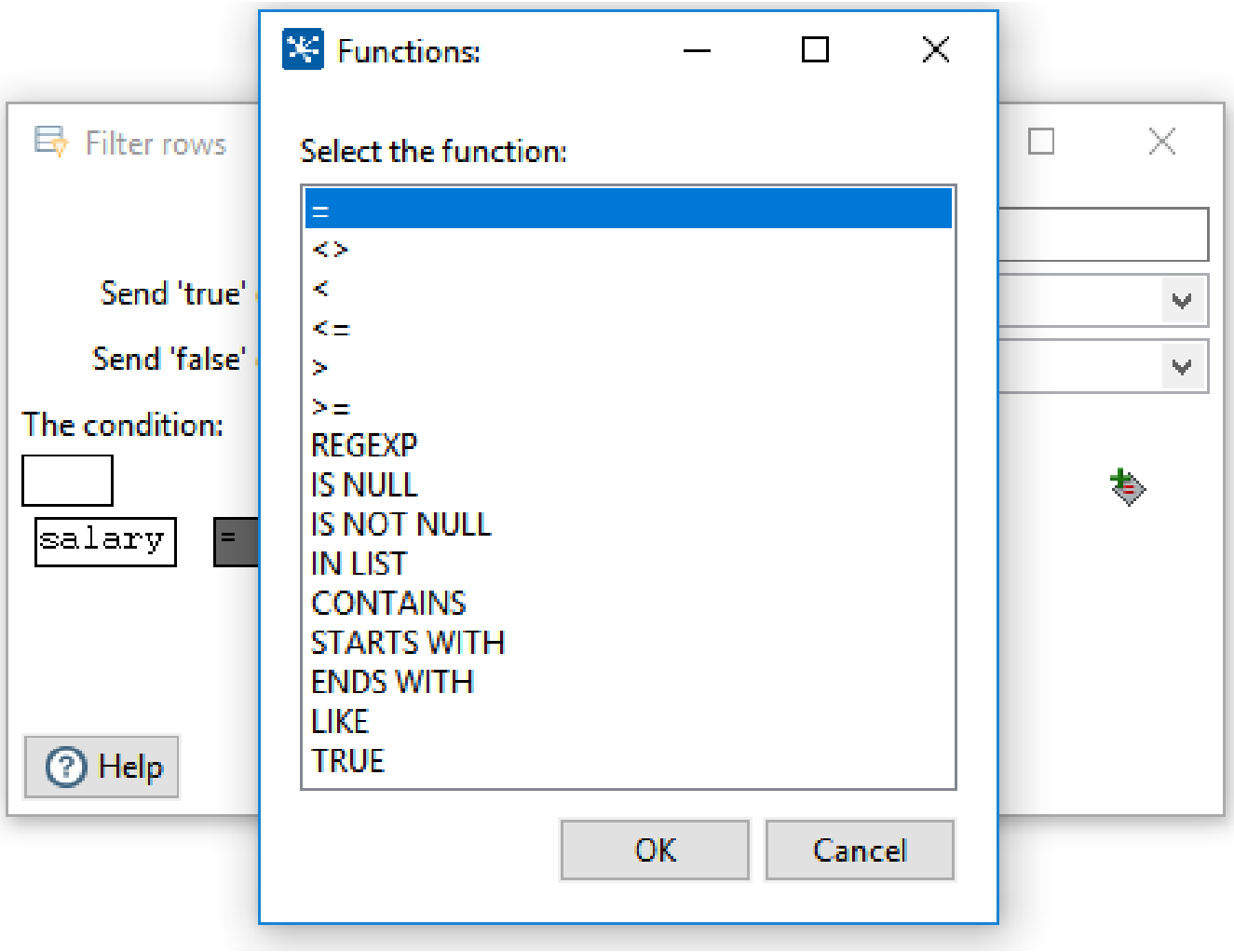
Filter rows


Filter rows
Step
Send 'true' data
Send 'false' data
The condition:
=
Help
Fields:
Select a field:
emp_no (Integer)
salary (Integer)
OK
Cancel

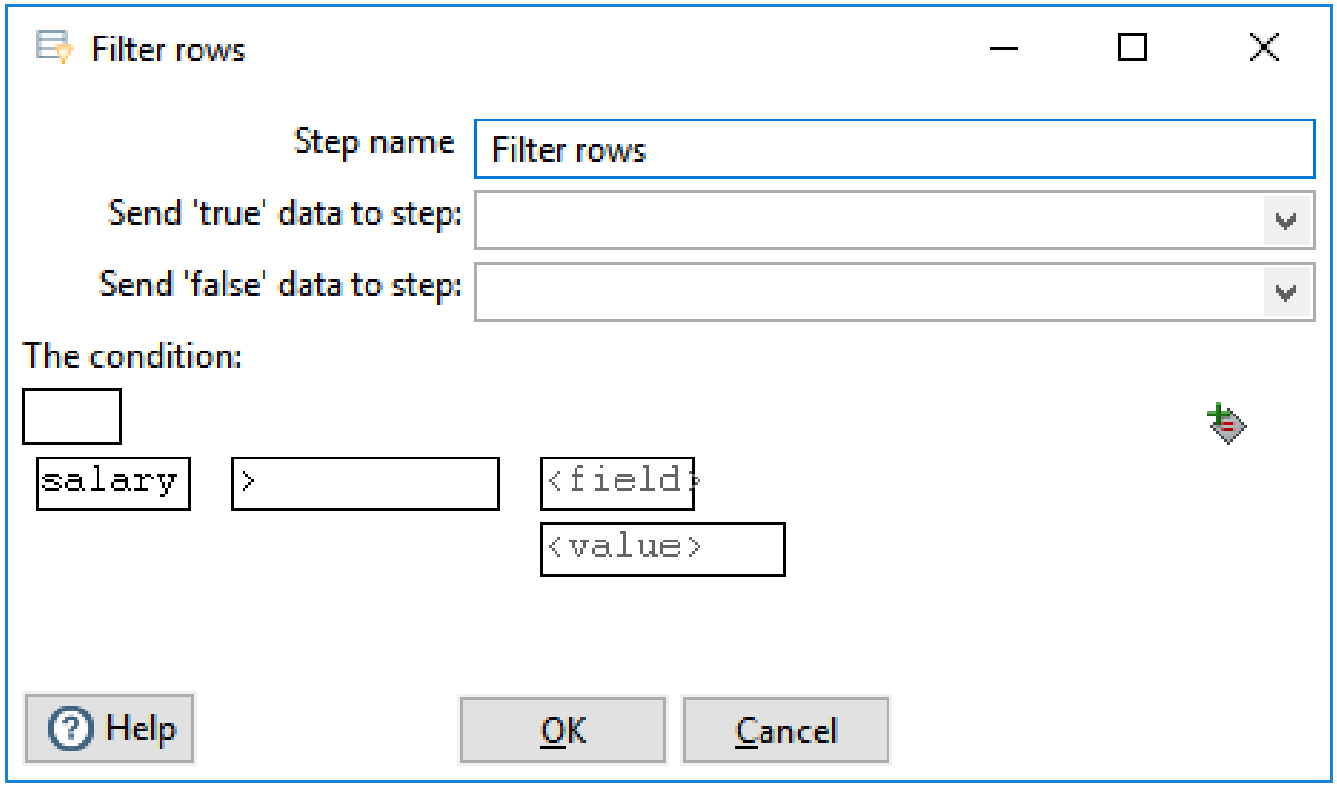
Filter rows


Filter rows
Step name
Filter rows
Send 'true' data to step:
Send 'false' data to step:
The condition:
salary
=
<field>
<value>
Help
OK
Cancel

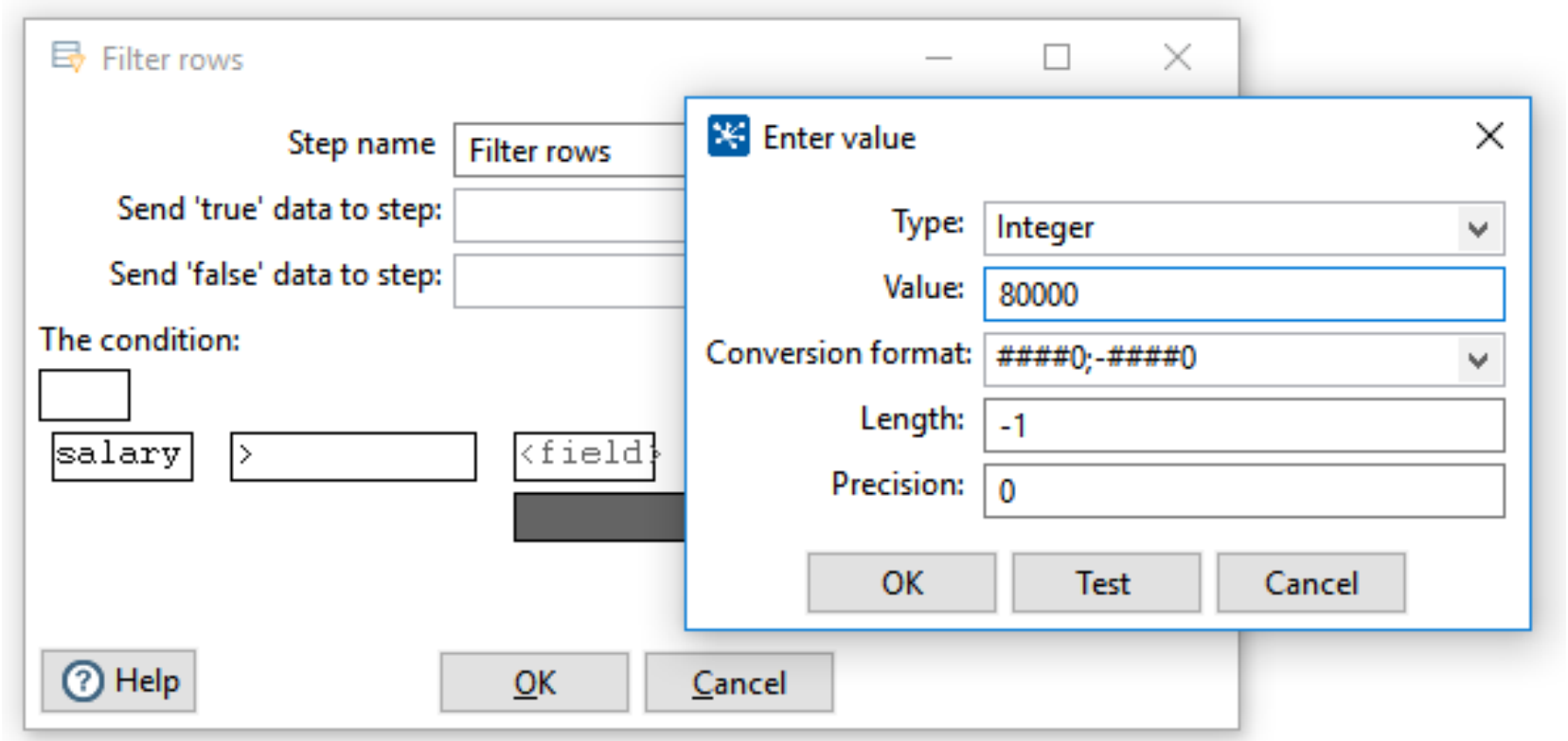
Filter rows


Filter rows
Send 'true'
Send 'false'
The condition:
salary
=
Help
Functions:
Select the function:
=
<>
<
<=
>
>=
REGEXP
IS NULL
IS NOT NULL
IN LIST
CONTAINS
STARTS WITH
ENDS WITH
LIKE
TRUE
OK
Cancel

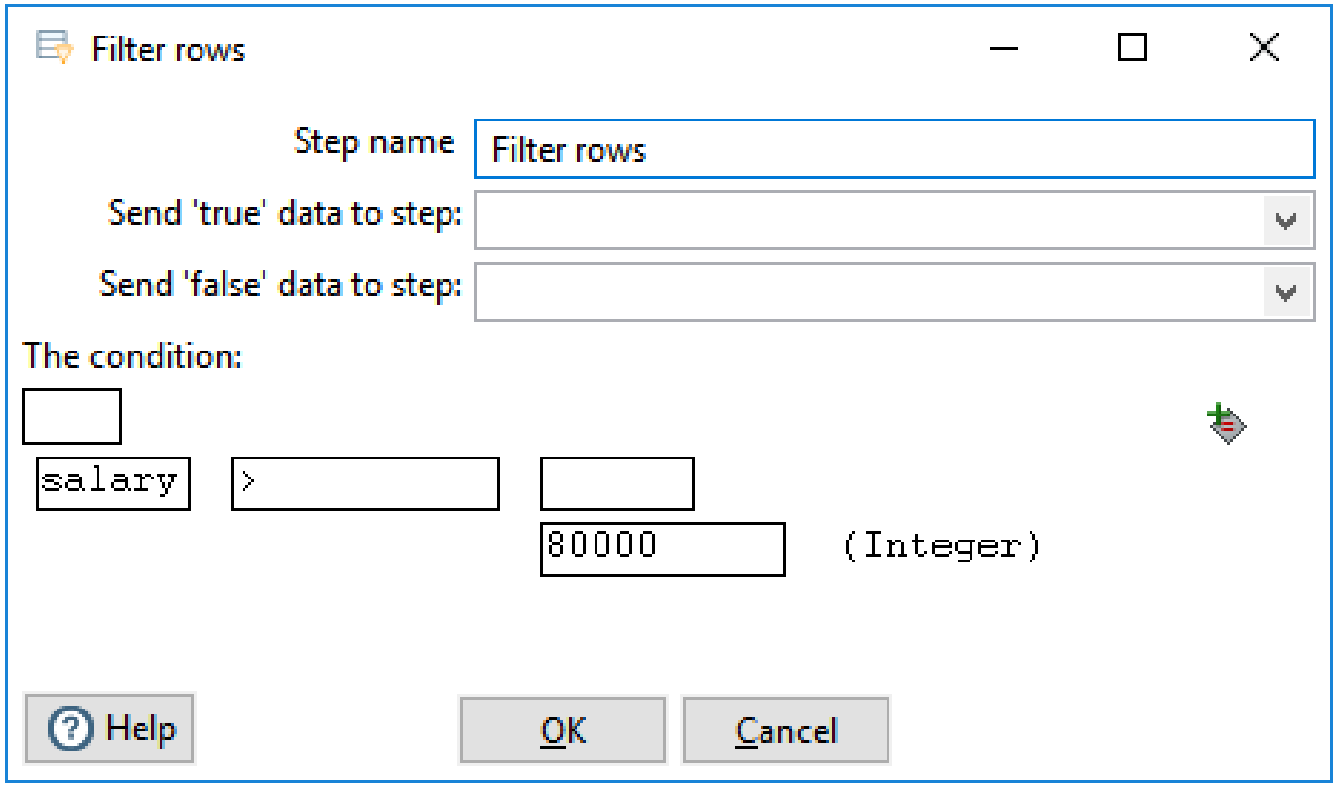
Filter rows


Filter rows
Step name
Filter rows
Send 'true' data to step:
Send 'false' data to step:
The condition:
salary
>
<field>
<value>
Help
OK
Cancel

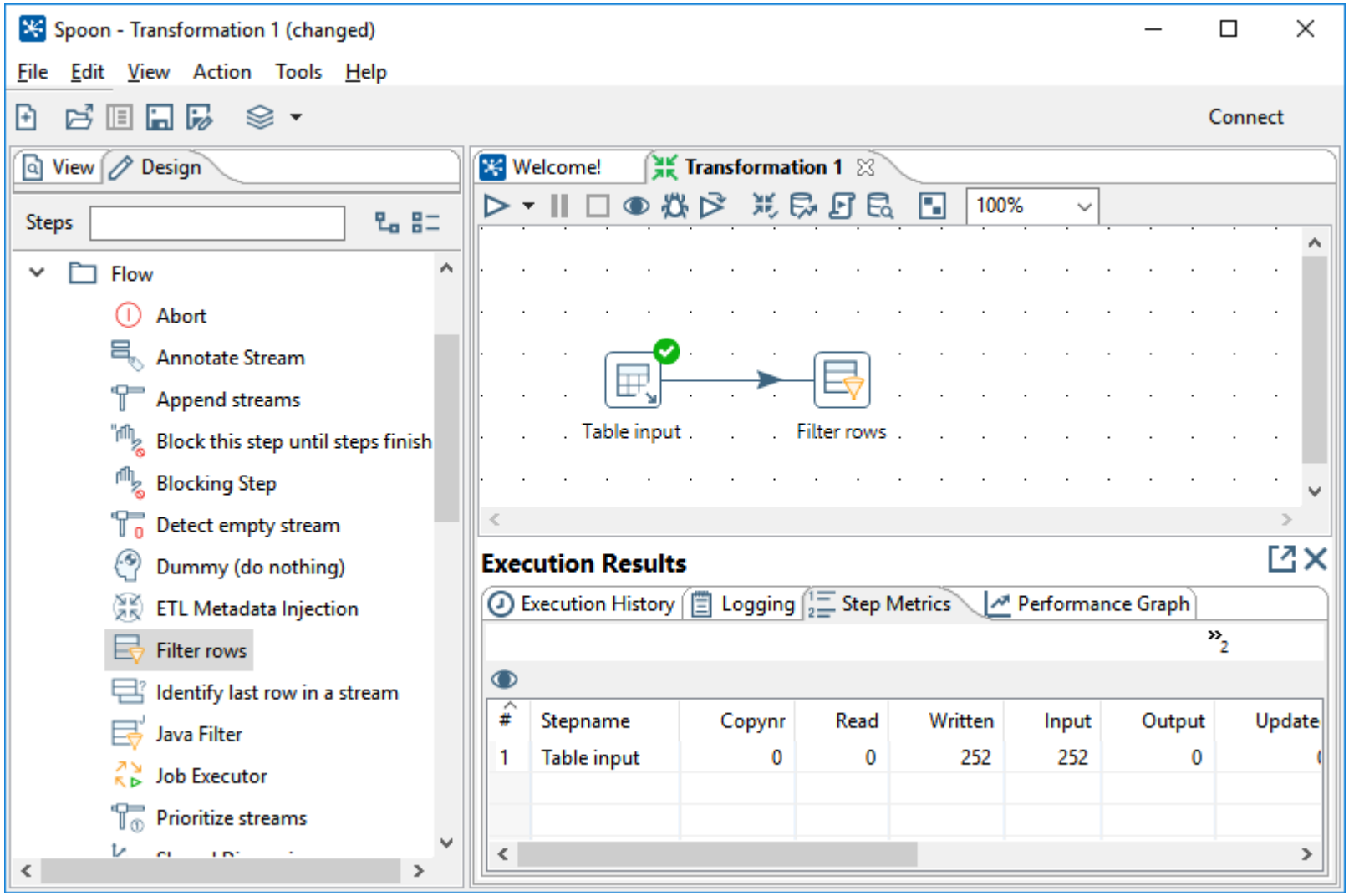
Filter rows


Filter rows
Step name
Filter rows
Send 'true' data to step:
Send 'false' data to step:
The condition:
salary
>
<field>
Help
OK
Cancel
Enter value
Type:
Integer
Value:
80000
Conversion format:
####0;-####0
Length:
-1
Precision:
0
OK
Test
Cancel

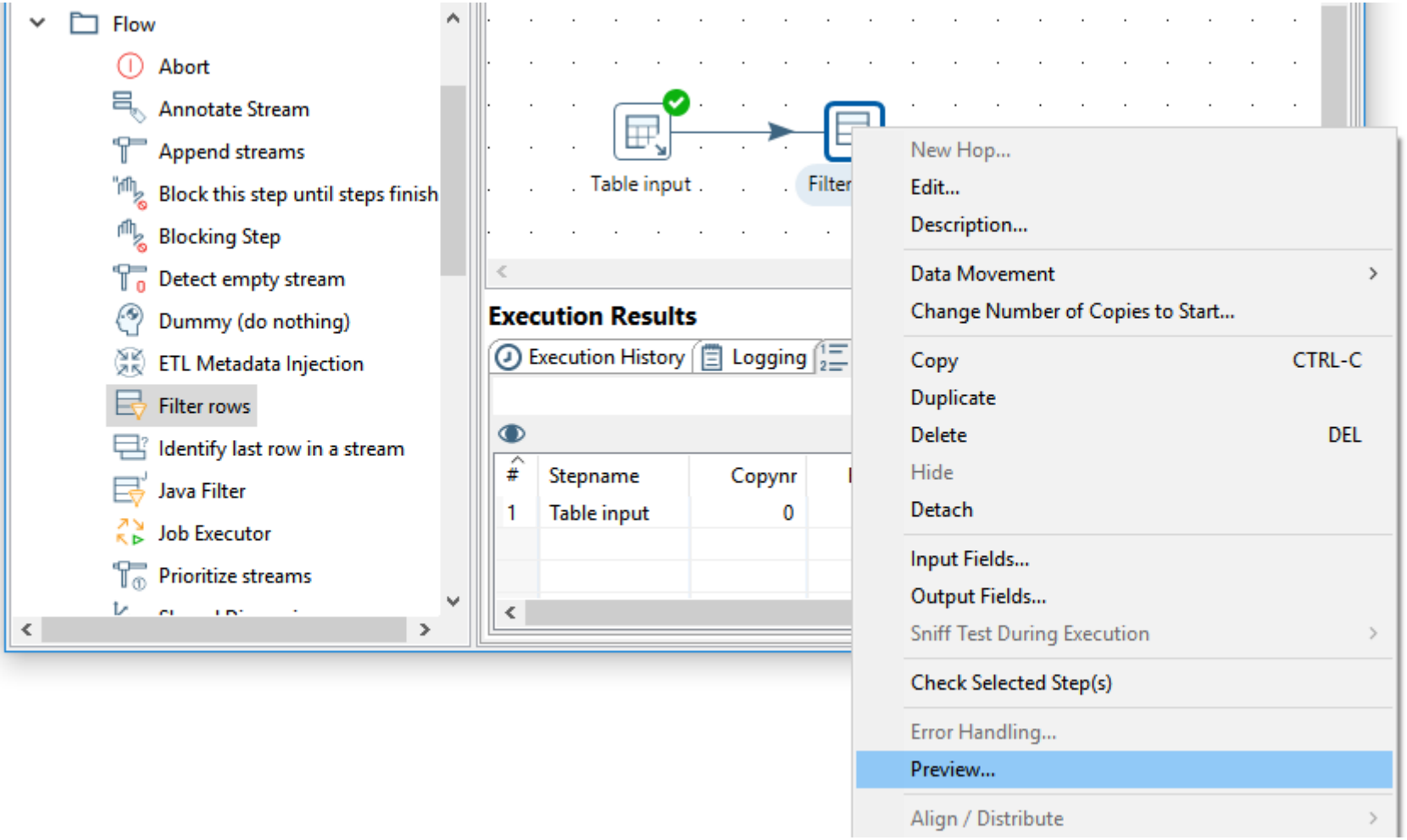
Filter rows


Filter rows
Step name
Filter rows
Send 'true' data to step:
Send 'false' data to step:
The condition:
salary
>
80000
(Integer)
Help
OK
Cancel

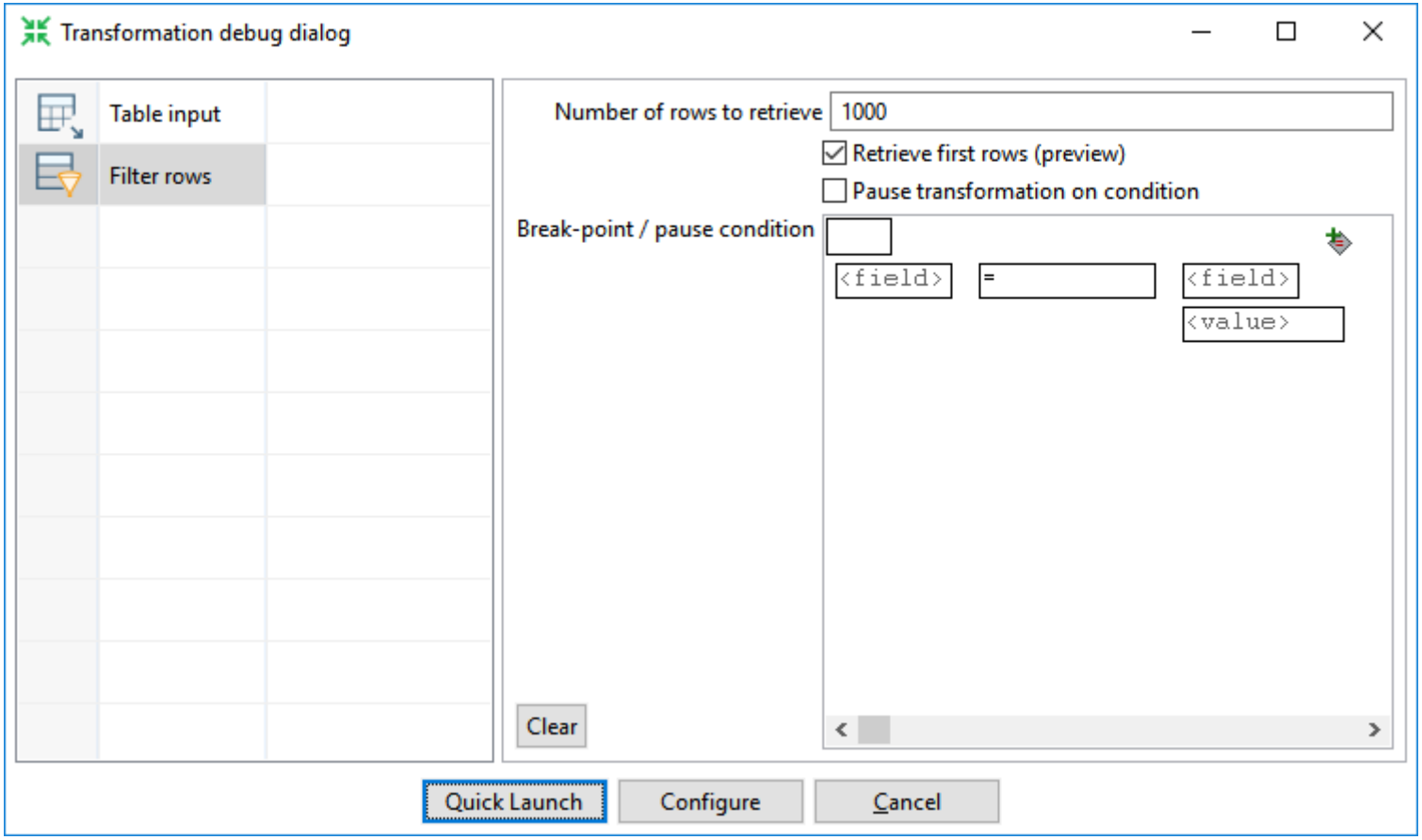
Filter rows


Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
Flow
Abort
Annotate Stream
Append streams
Block this step until steps finish
Blocking Step
Detect empty stream
Dummy (do nothing)
ETL Metadata Injection
Filter rows
Identify last row in a stream
Java Filter
Job Executor
Prioritize streams
Welcome!
Transformation 1
100%
Table input
Filter rows
Execution Results
Execution History
Logging
Step Metrics
Performance Graph
Stepname Copynr Read Written Input Output Update
1 Table input 0 0 252 252 0

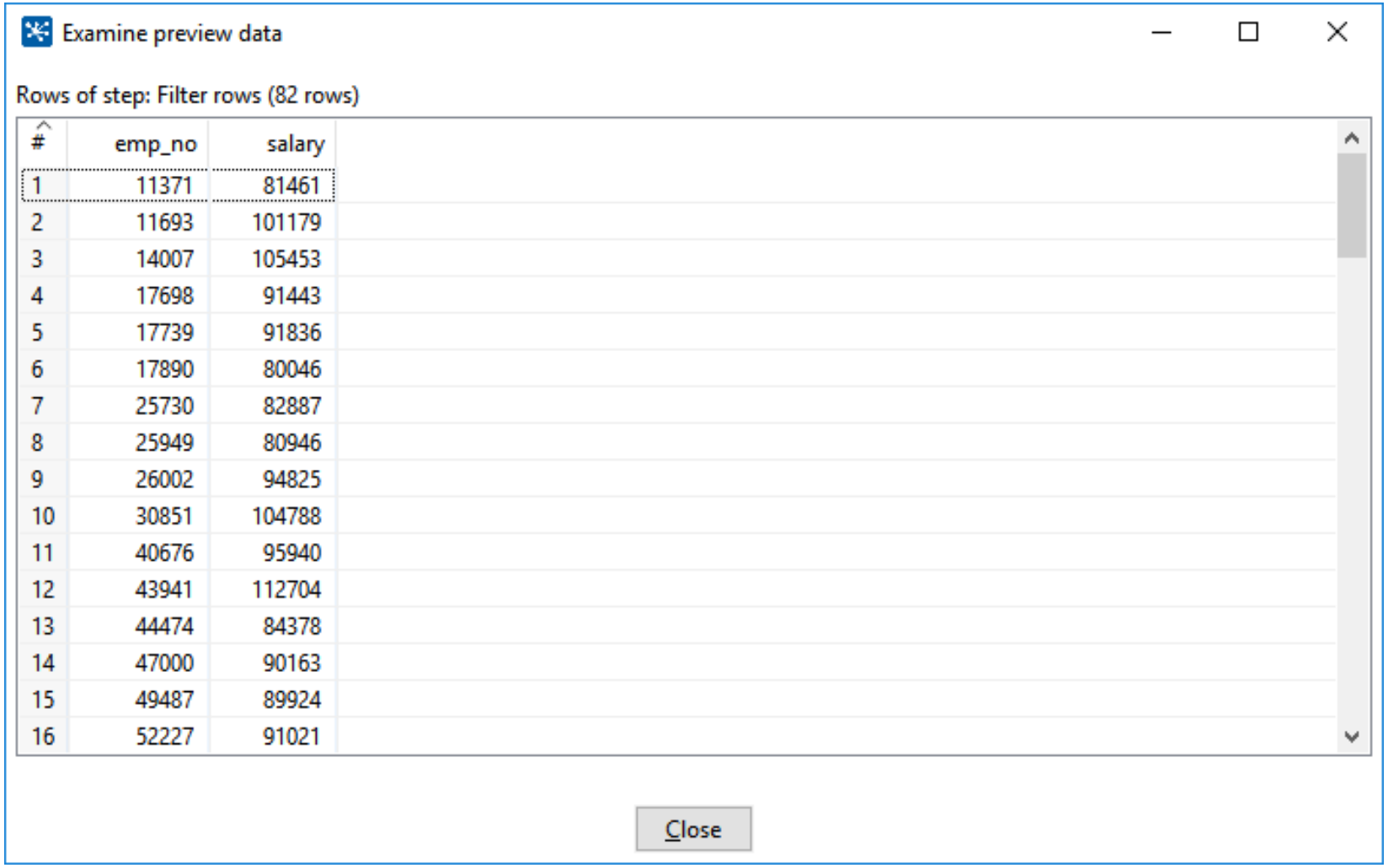
Filter rows


Flow
Abort
Annotate Stream
Append streams
Block this step until steps finish
Blocking Step
Detect empty stream
Dummy (do nothing)
ETL Metadata Injection
Filter rows
Identify last row in a stream
Java Filter
Job Executor
Prioritize streams
Table input
Filter
Execution Results
Execution History
Logging
Stepname Copynr
1 Table input 0
New Hop...
Edit...
Description...
Data Movement
Change Number of Copies to Start...
Copy CTRL-C
Duplicate
Delete DEL
Hide
Detach
Input Fields...
Output Fields...
Sniff Test During Execution
Check Selected Step(s)
Error Handling...
Preview...
Align / Distribute

Filter rows

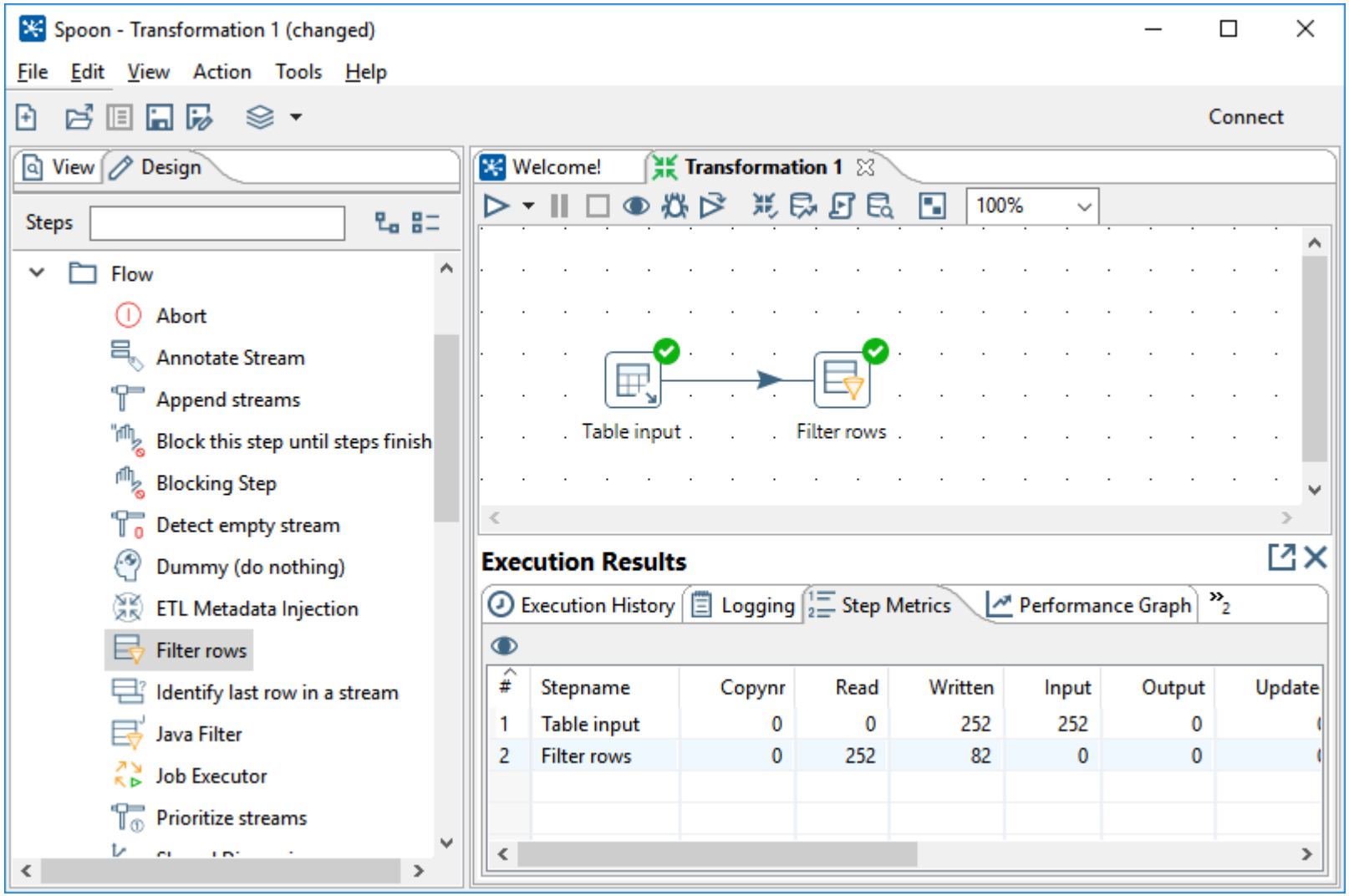

Transformation debug dialog
Table input
Filter rows
Number of rows to retrieve
1000
Retrieve first rows (preview)
Pause transformation on condition
Break-point / pause condition
<field>
=
<field>
<value>
Clear
Quick Launch
Configure
Cancel

Filter rows

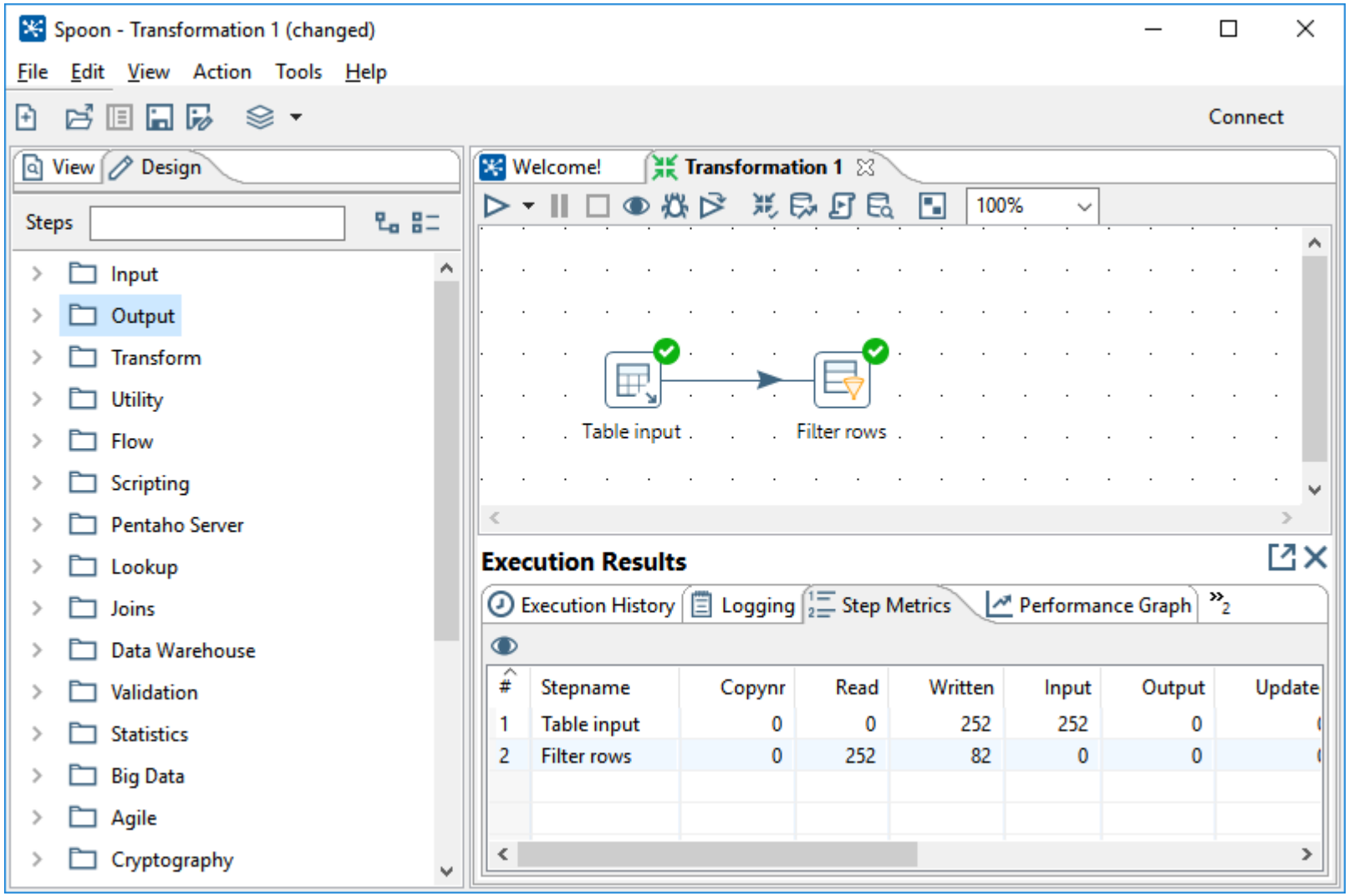


#	emp_no	salary
1	11371	81461
2	11693	101179
3	14007	105453
4	17698	91443
5	17739	91836
6	17890	80046
7	25730	82887
8	25949	80946
9	26002	94825
10	30851	104788
11	40676	95940
12	43941	112704
13	44474	84378
14	47000	90163
15	49487	89924
16	52227	91021

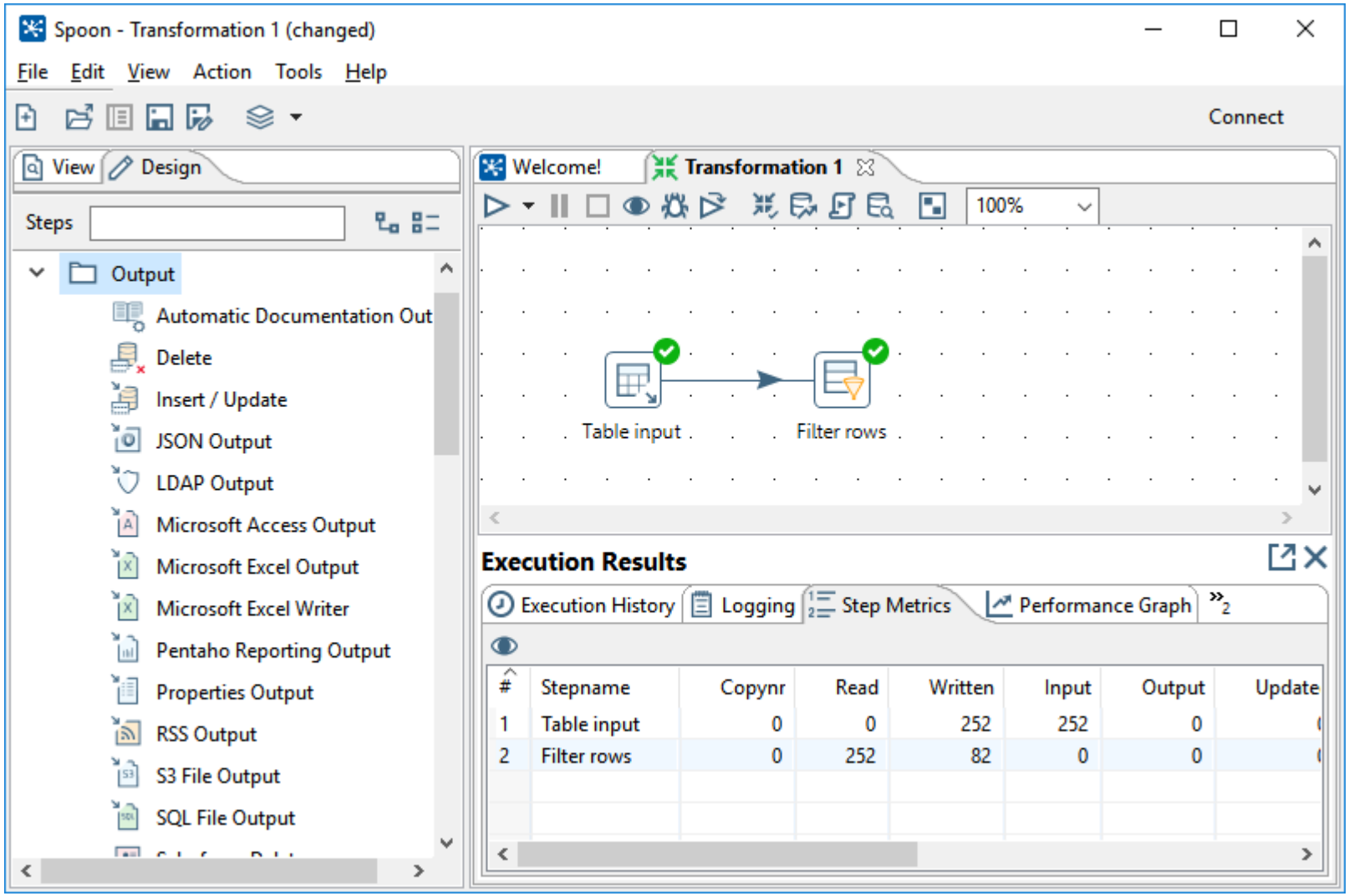
Filter rows


Spoon - Transformation 1 (changed)
File Edit View Action Tools Help
Connect
View Design
Steps
Flow
Abort
Annotate Stream
Append streams
Block this step until steps finish
Blocking Step
Detect empty stream
Dummy (do nothing)
ETL Metadata Injection
Filter rows
Identify last row in a stream
Java Filter
Job Executor
Prioritize streams
Welcome!
Transformation 1
100%
Table input
Filter rows
Execution Results
Execution History
Logging
Step Metrics
Performance Graph
Stepname Copynr Read Written Input Output Update
1 Table input 0 0 252 252 0
2 Filter rows 0 252 82 0 0

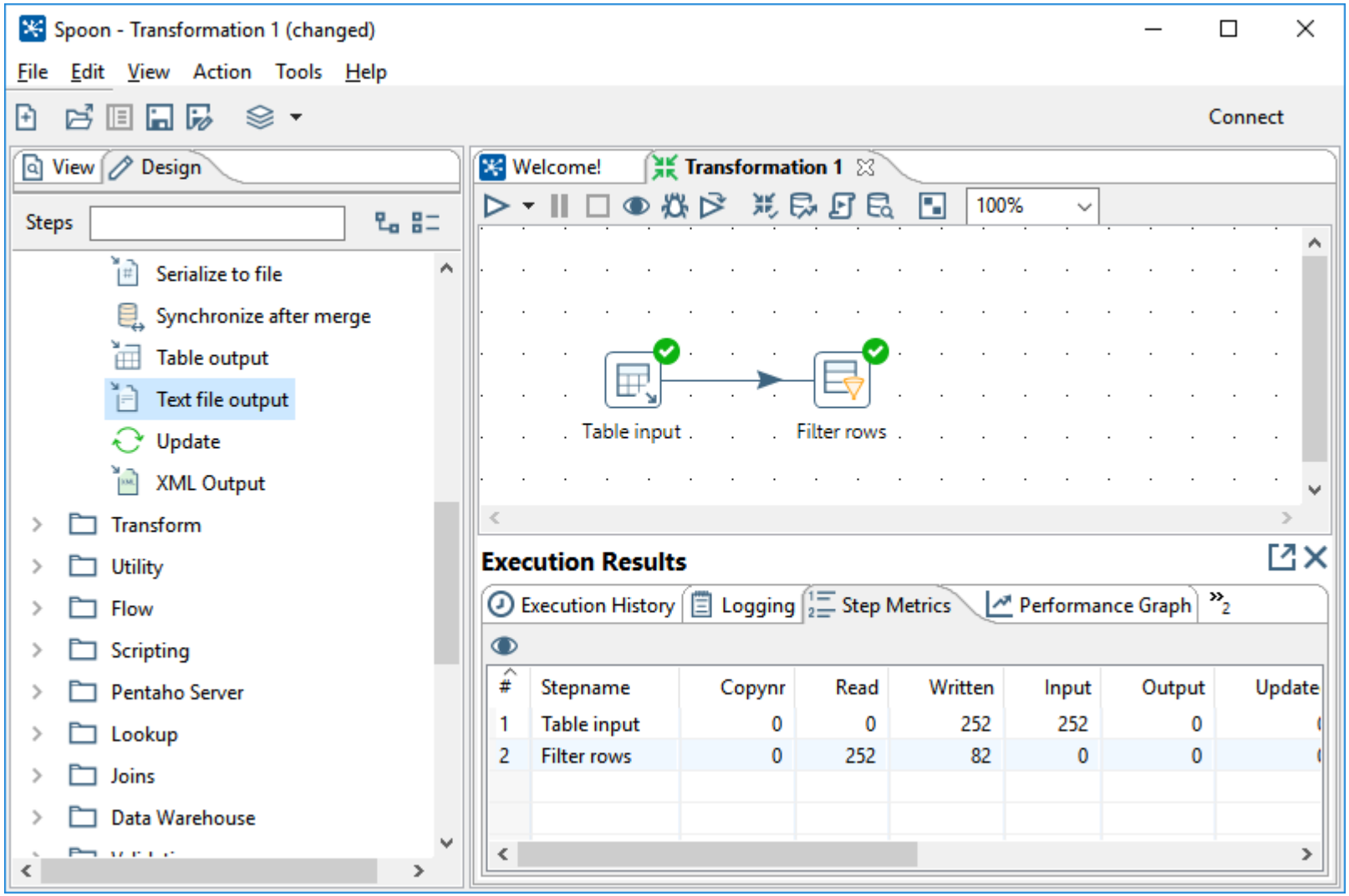
Text file output



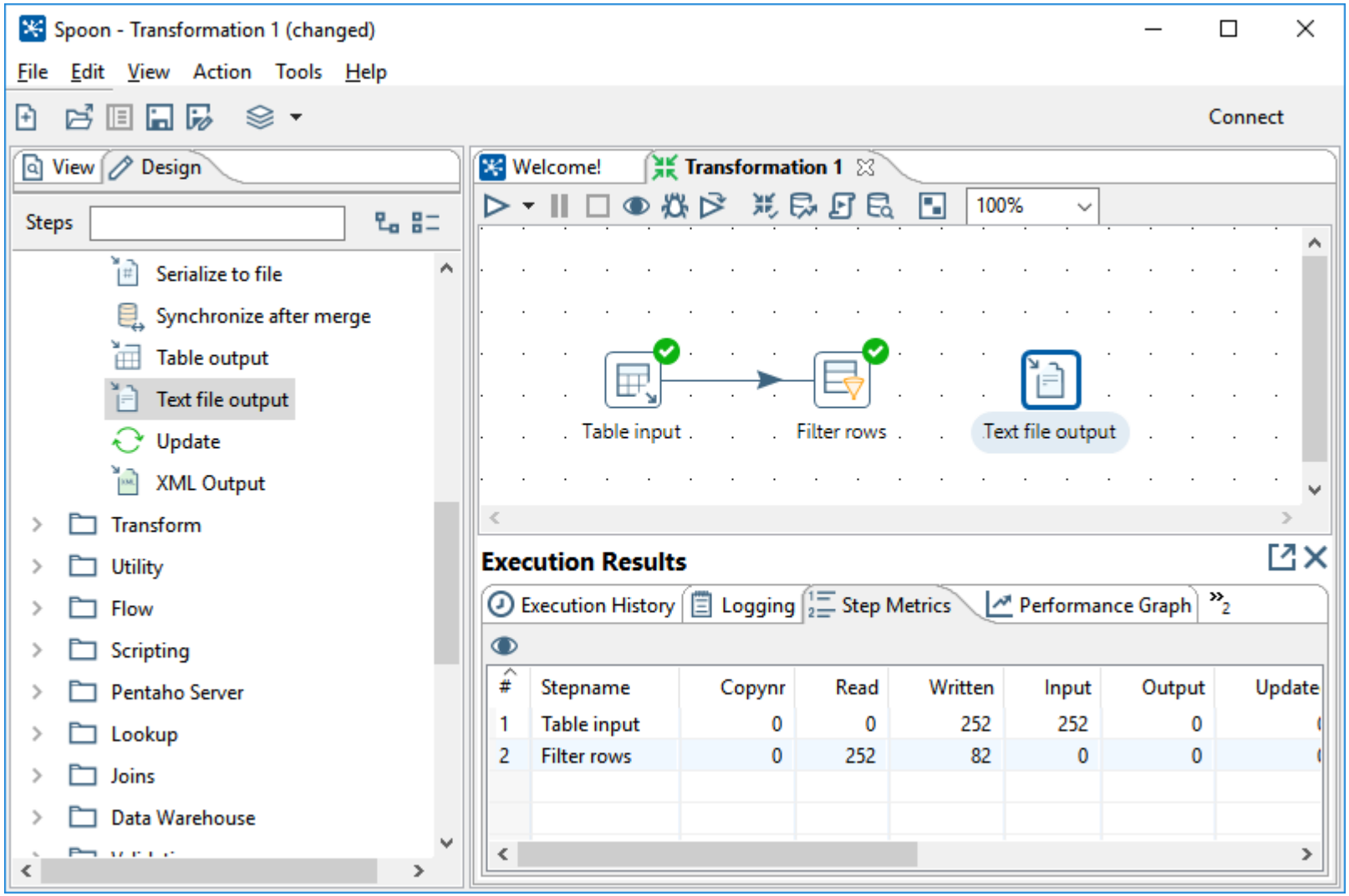
Text file output



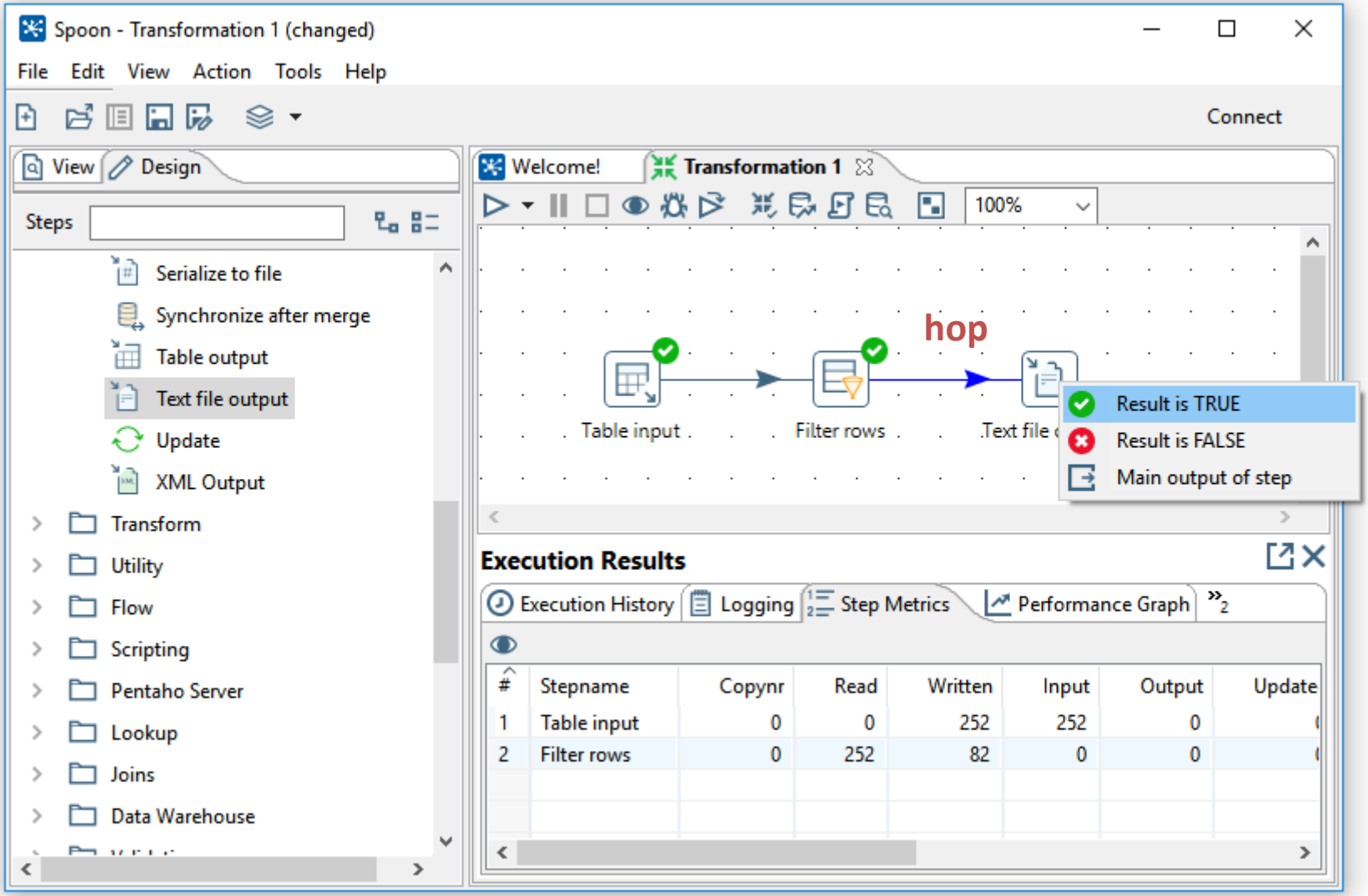
Text file output



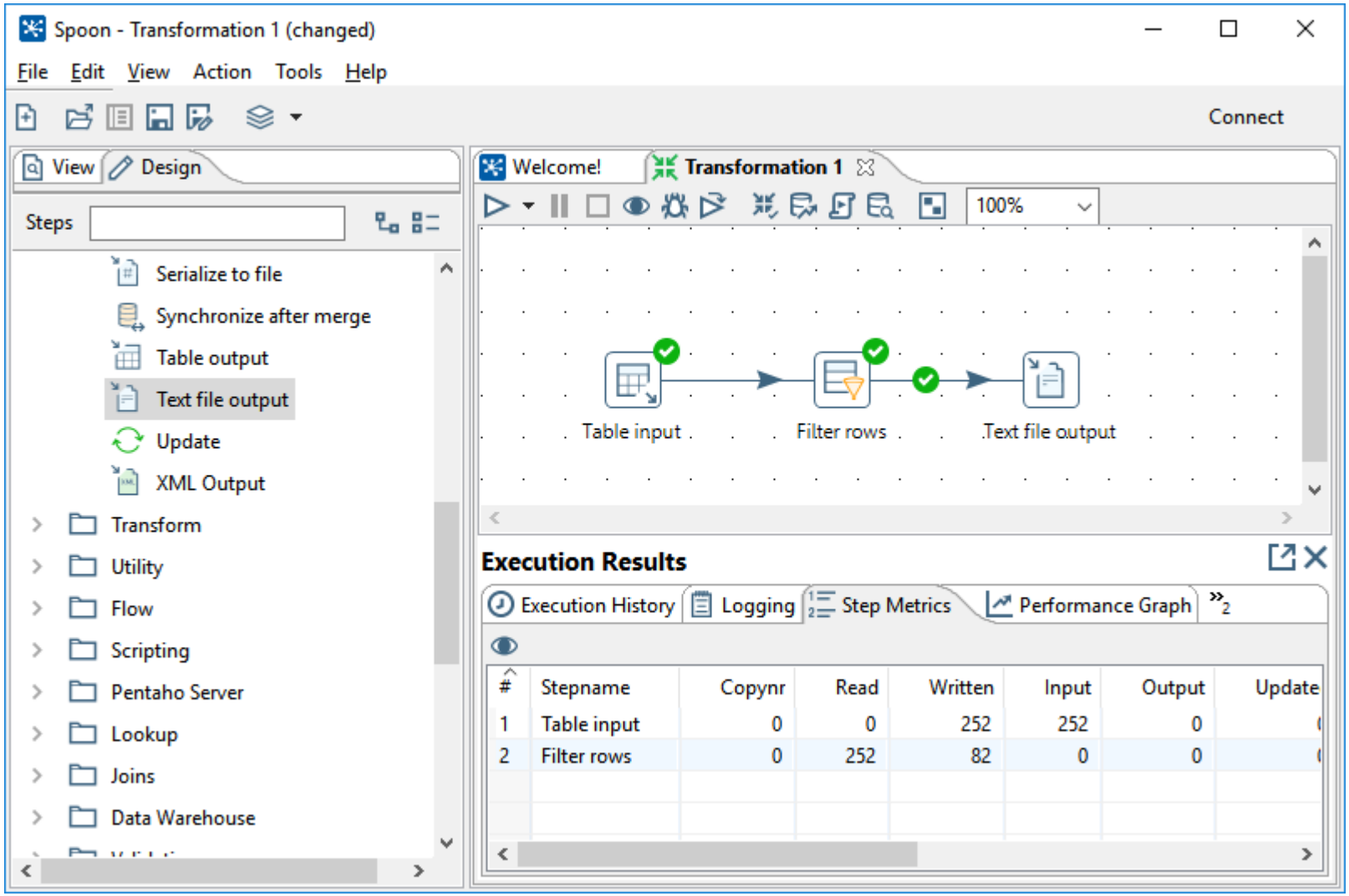
Text file output



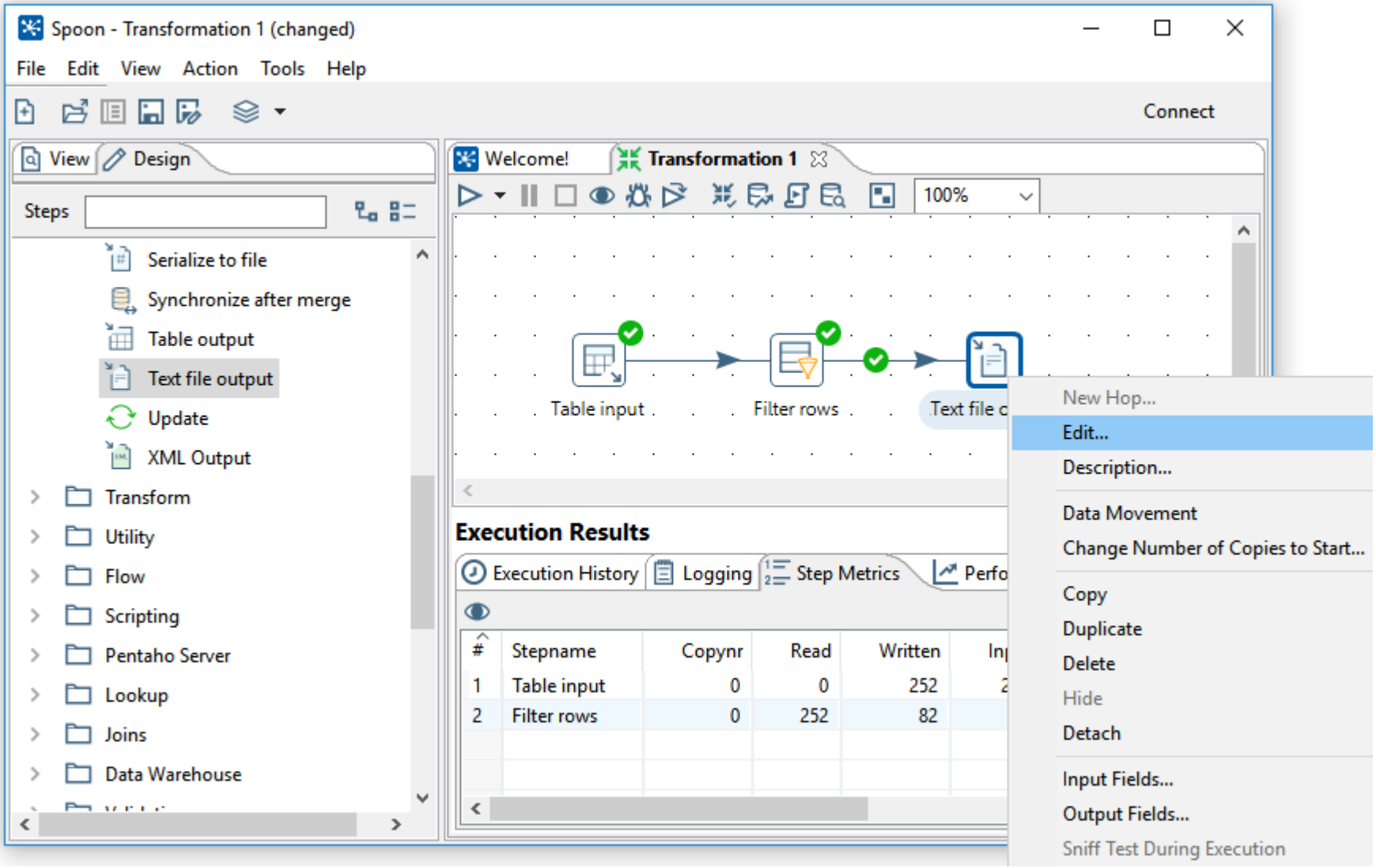
Text file output



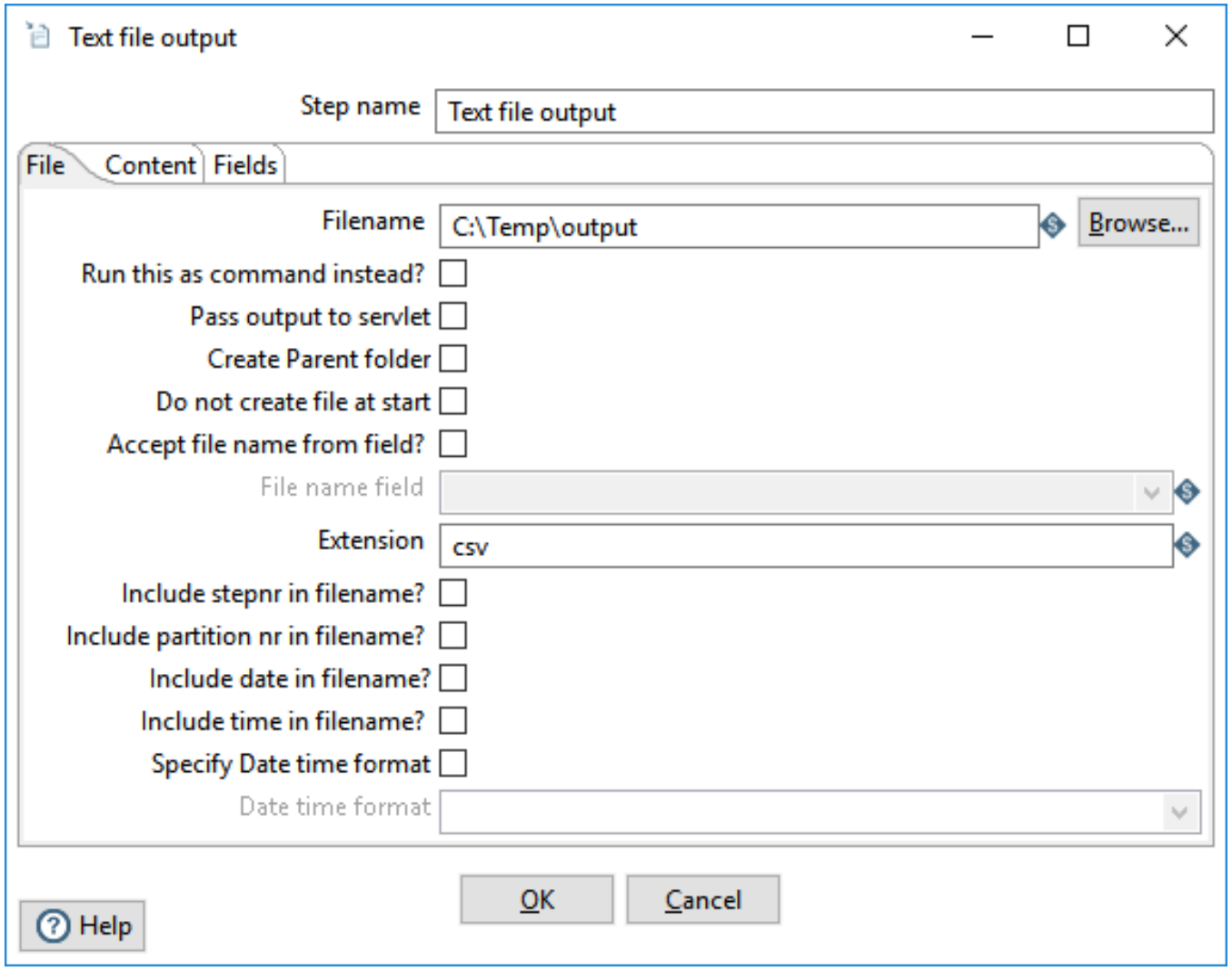
Text file output



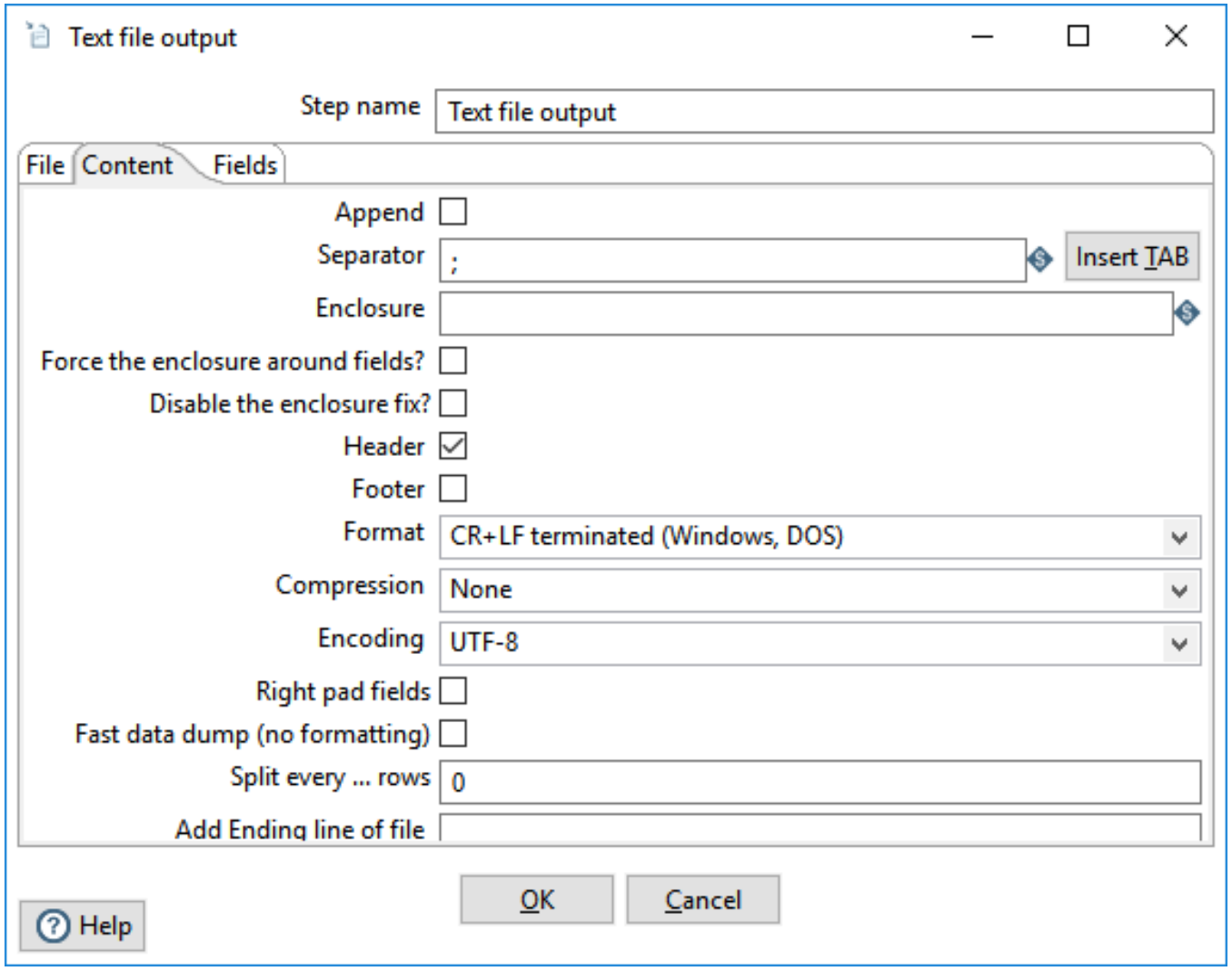
Text file output



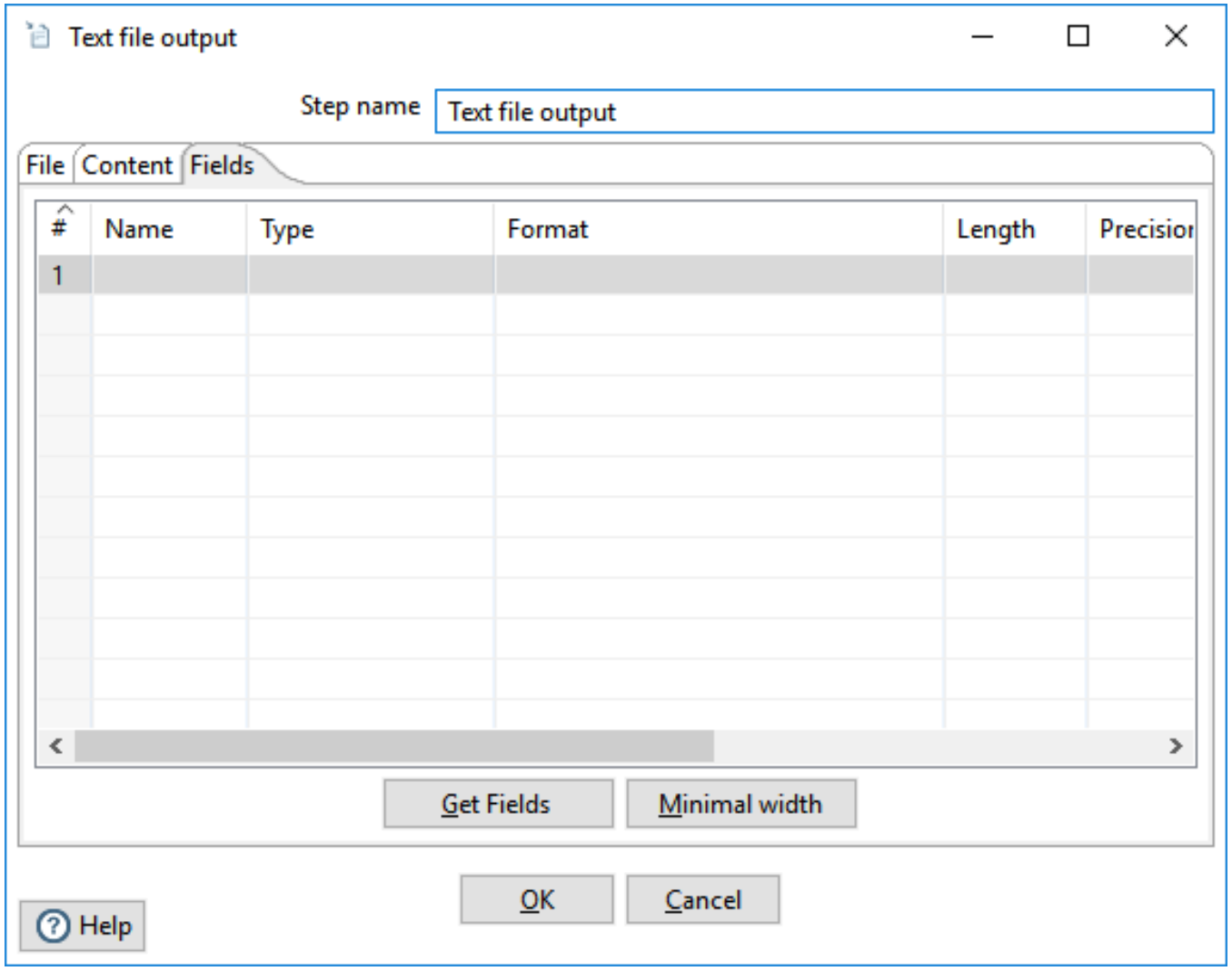
Text file output



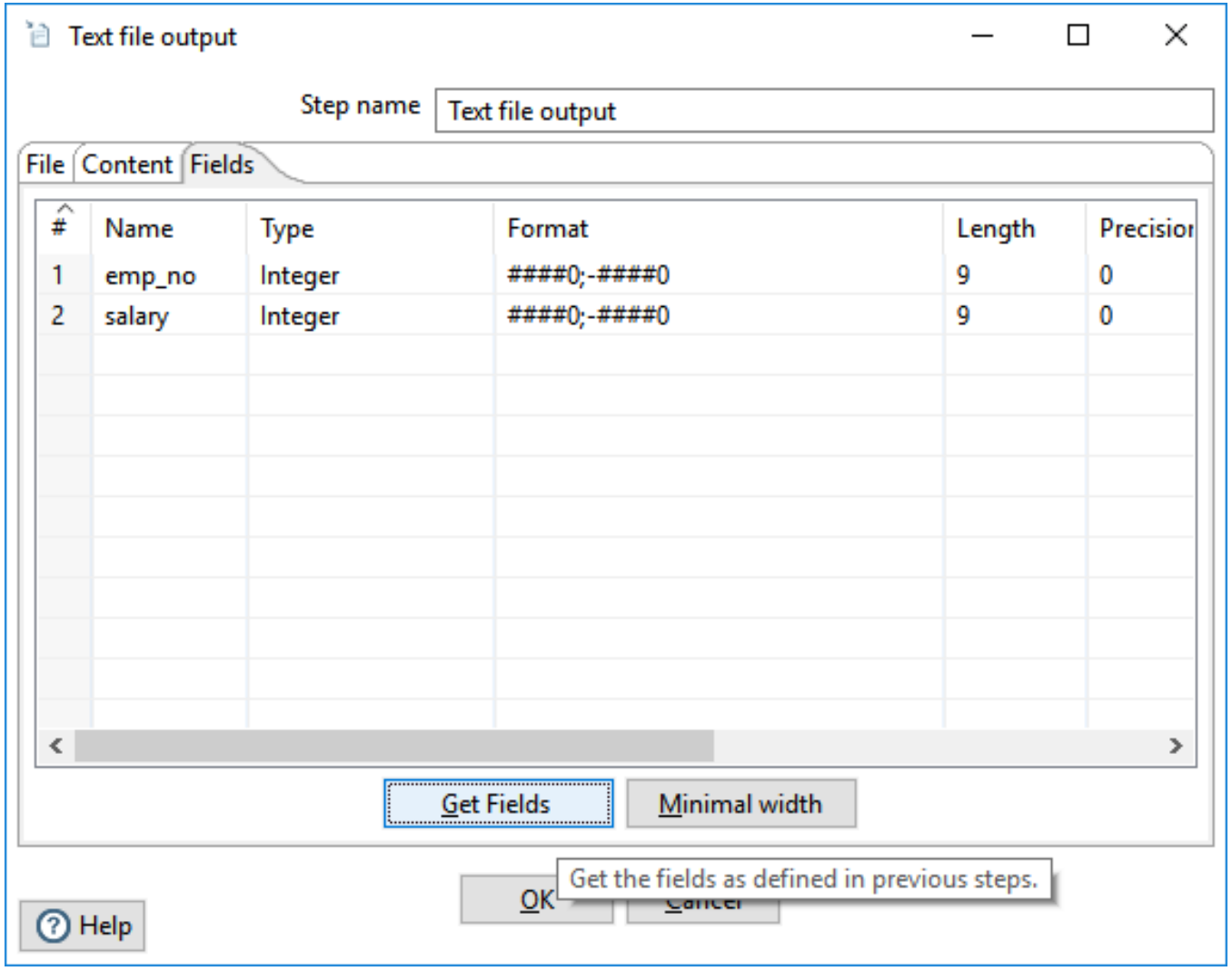
Text file output



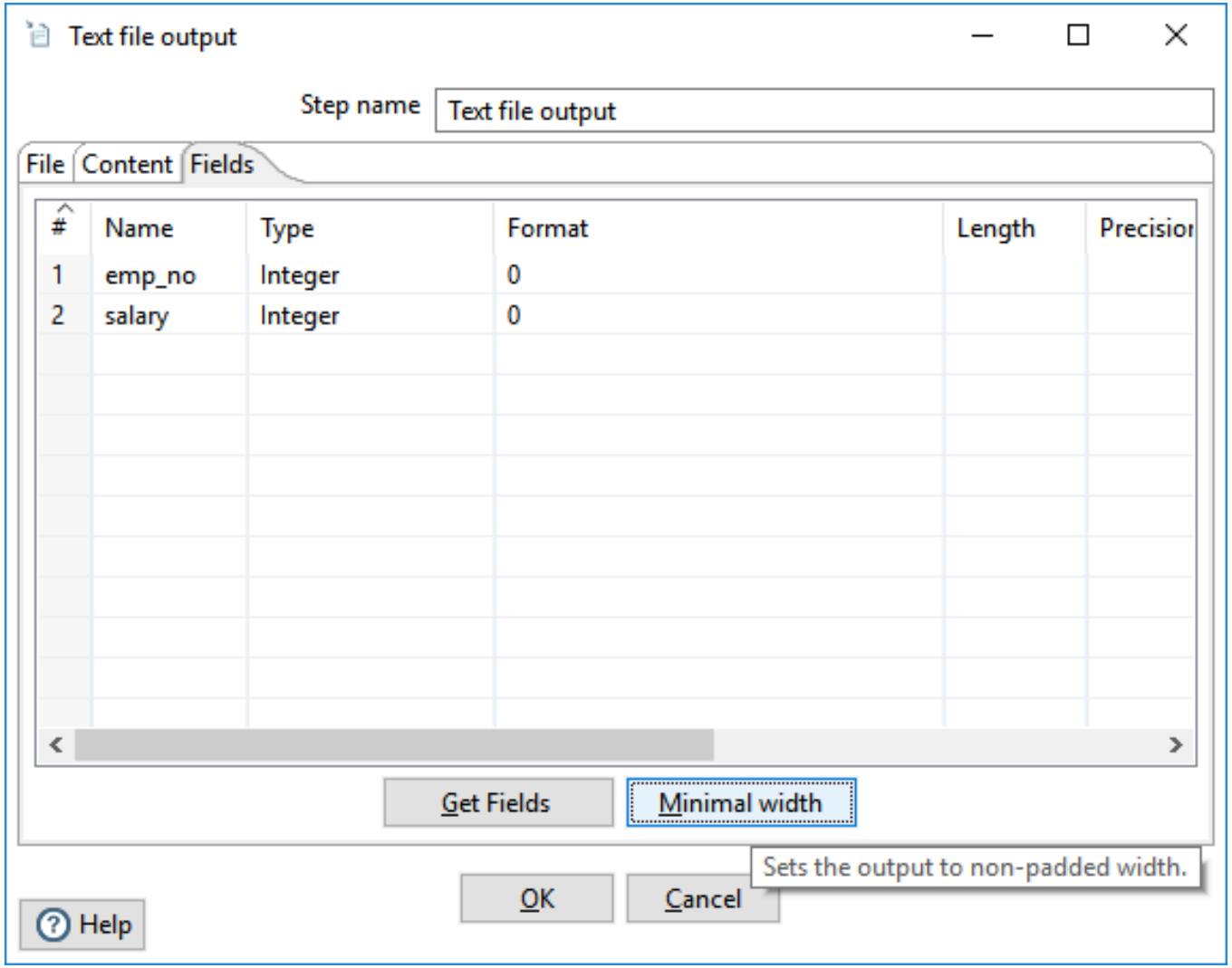
Text file output



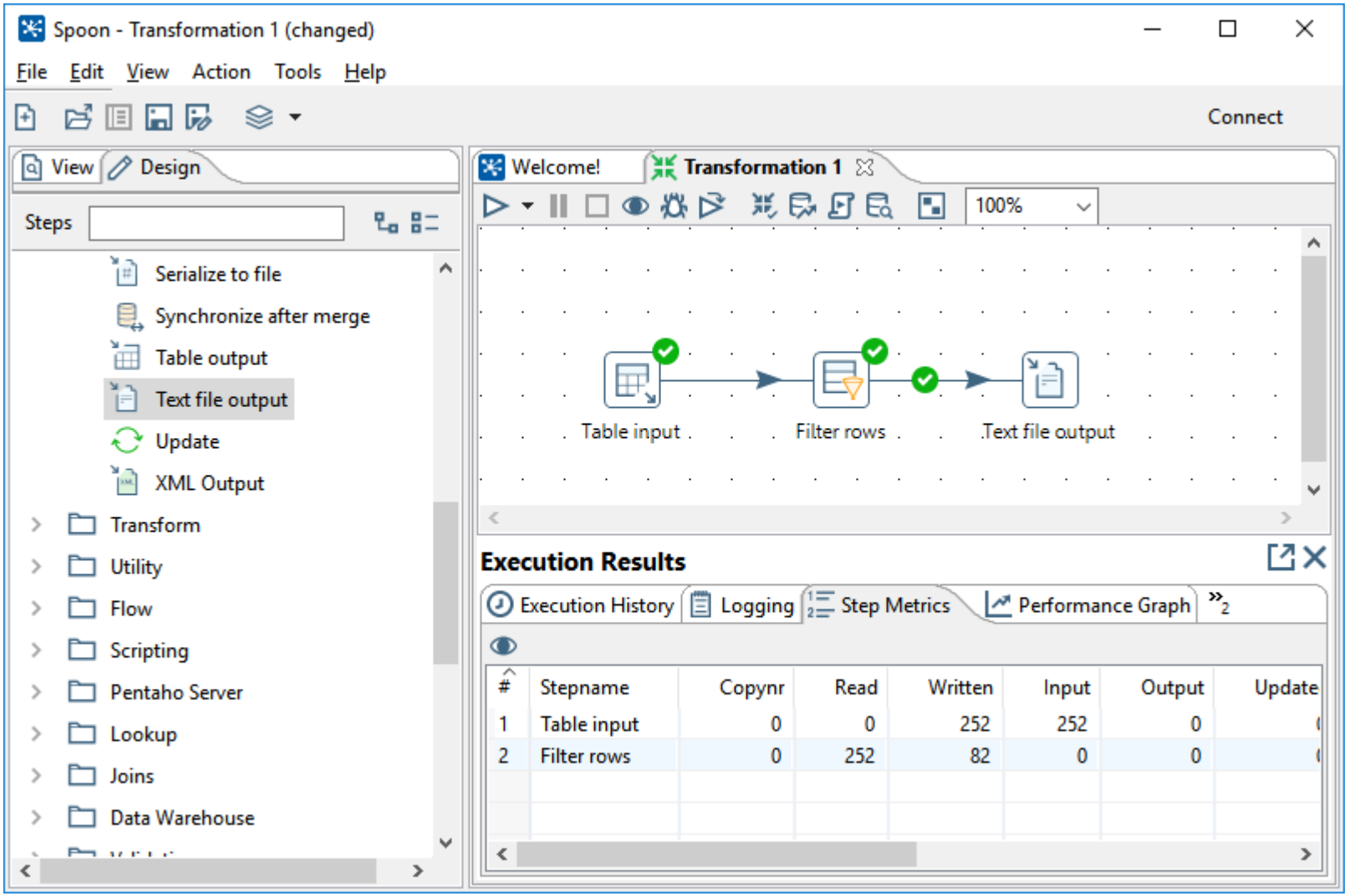
Text file output



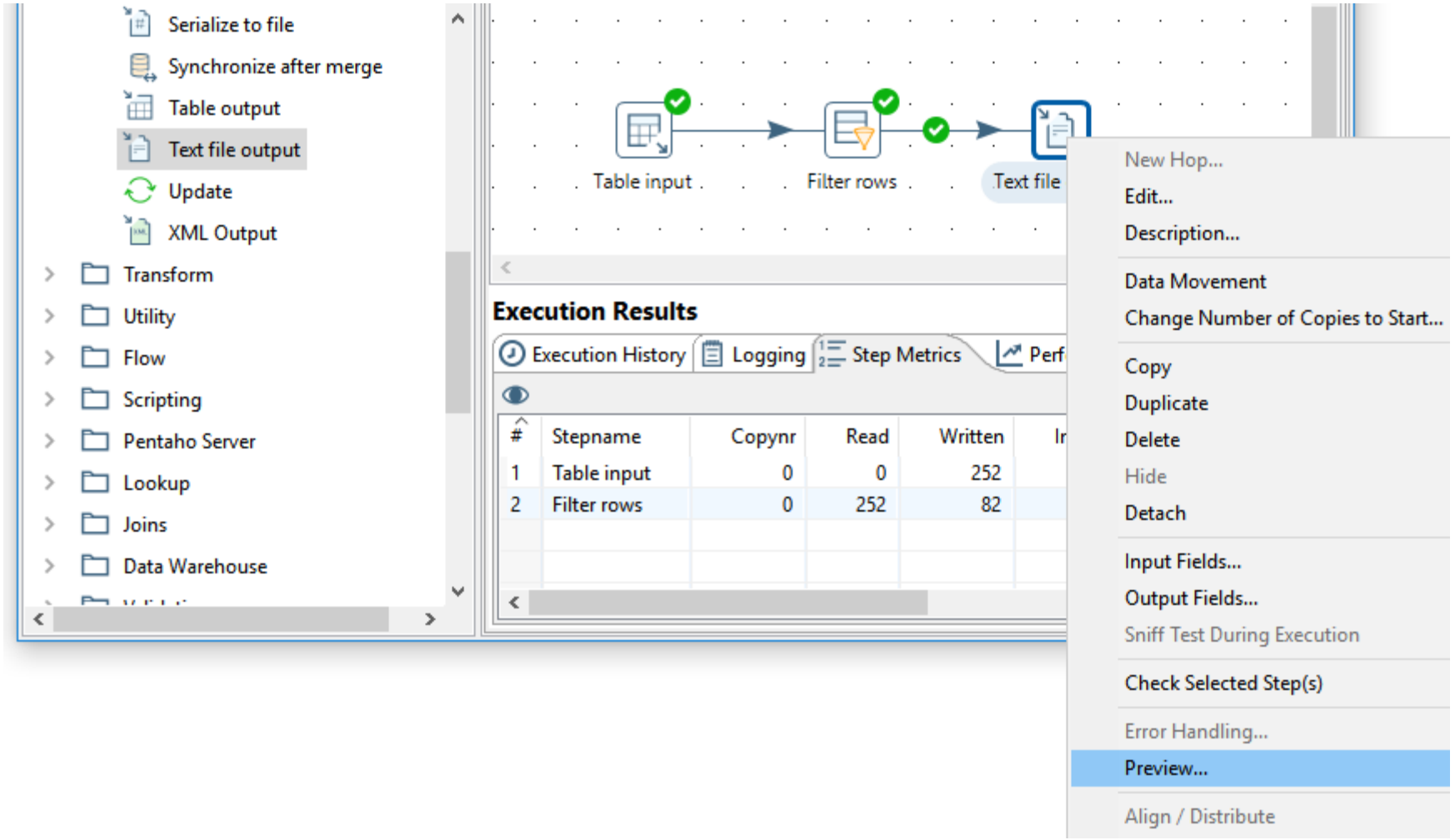
Text file output



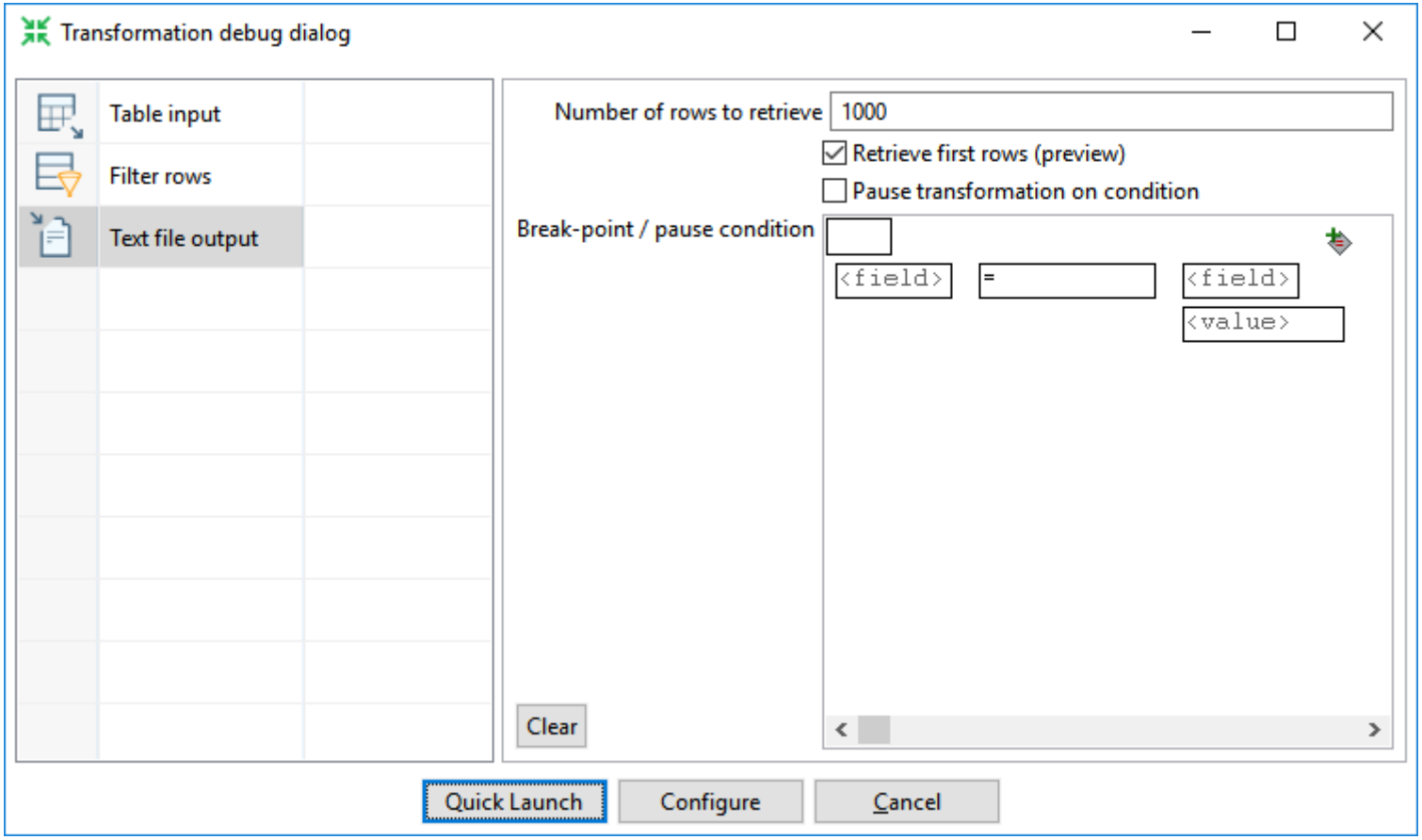
Text file output



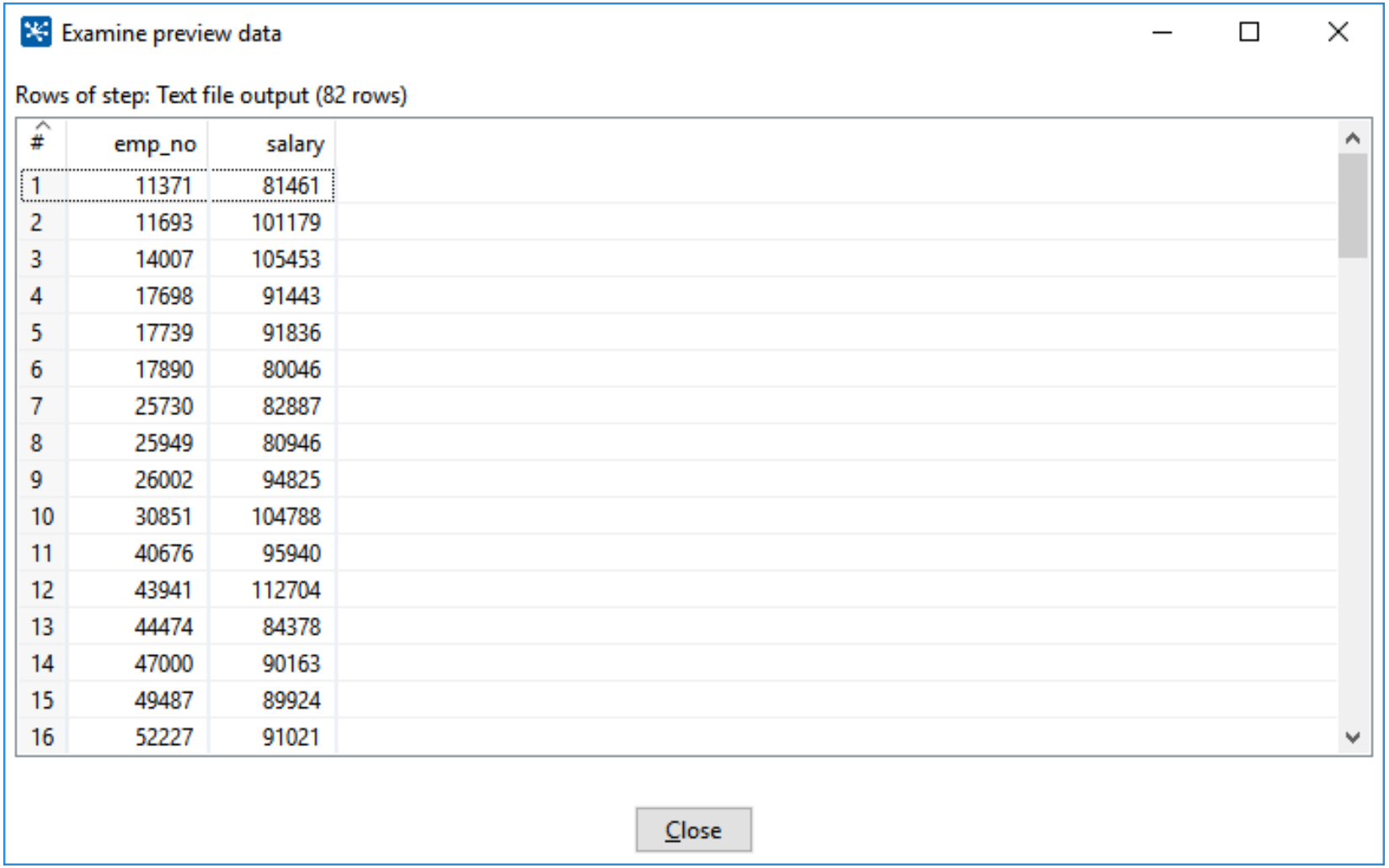
Text file output



Text file output

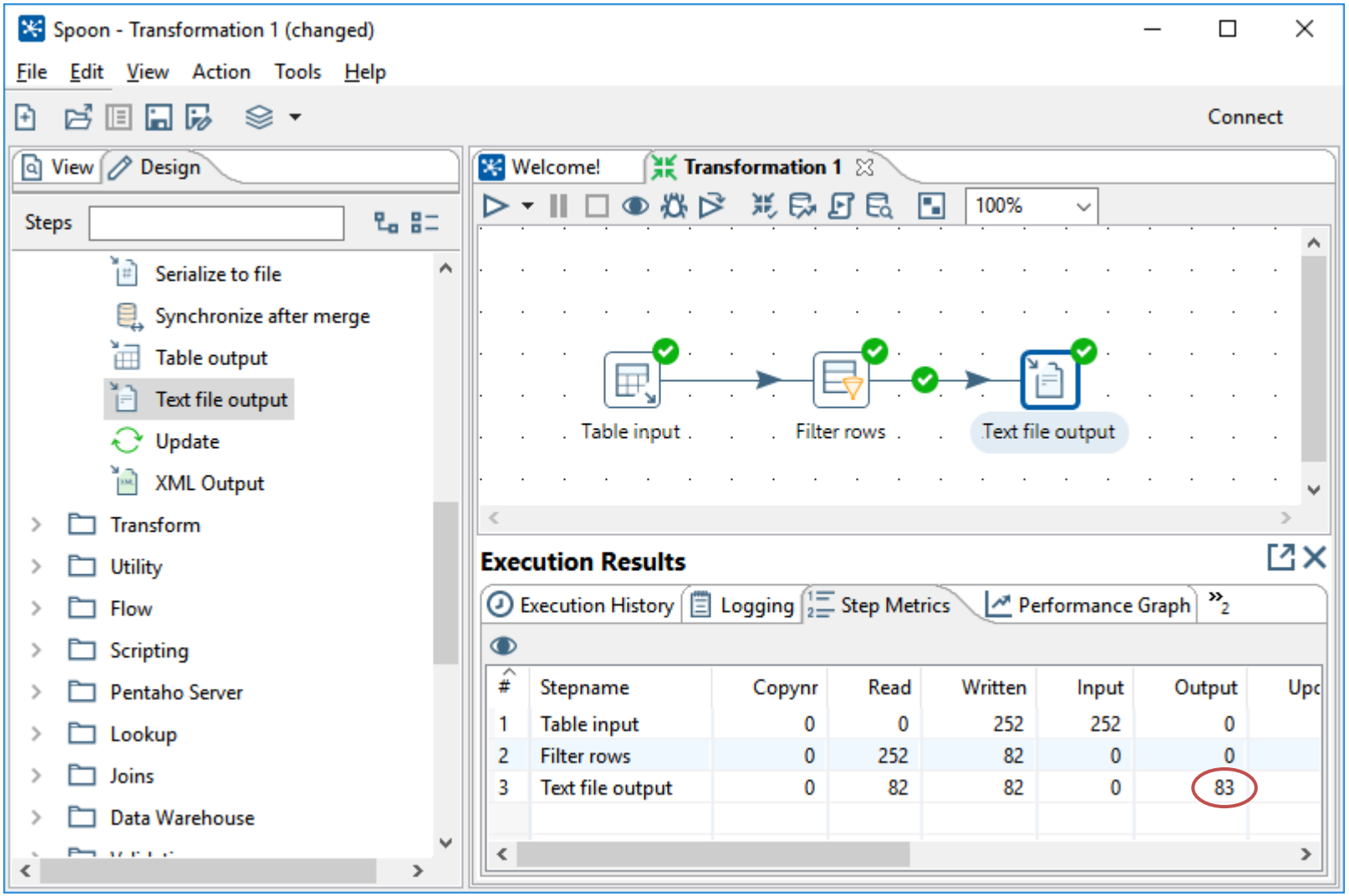


Text file output

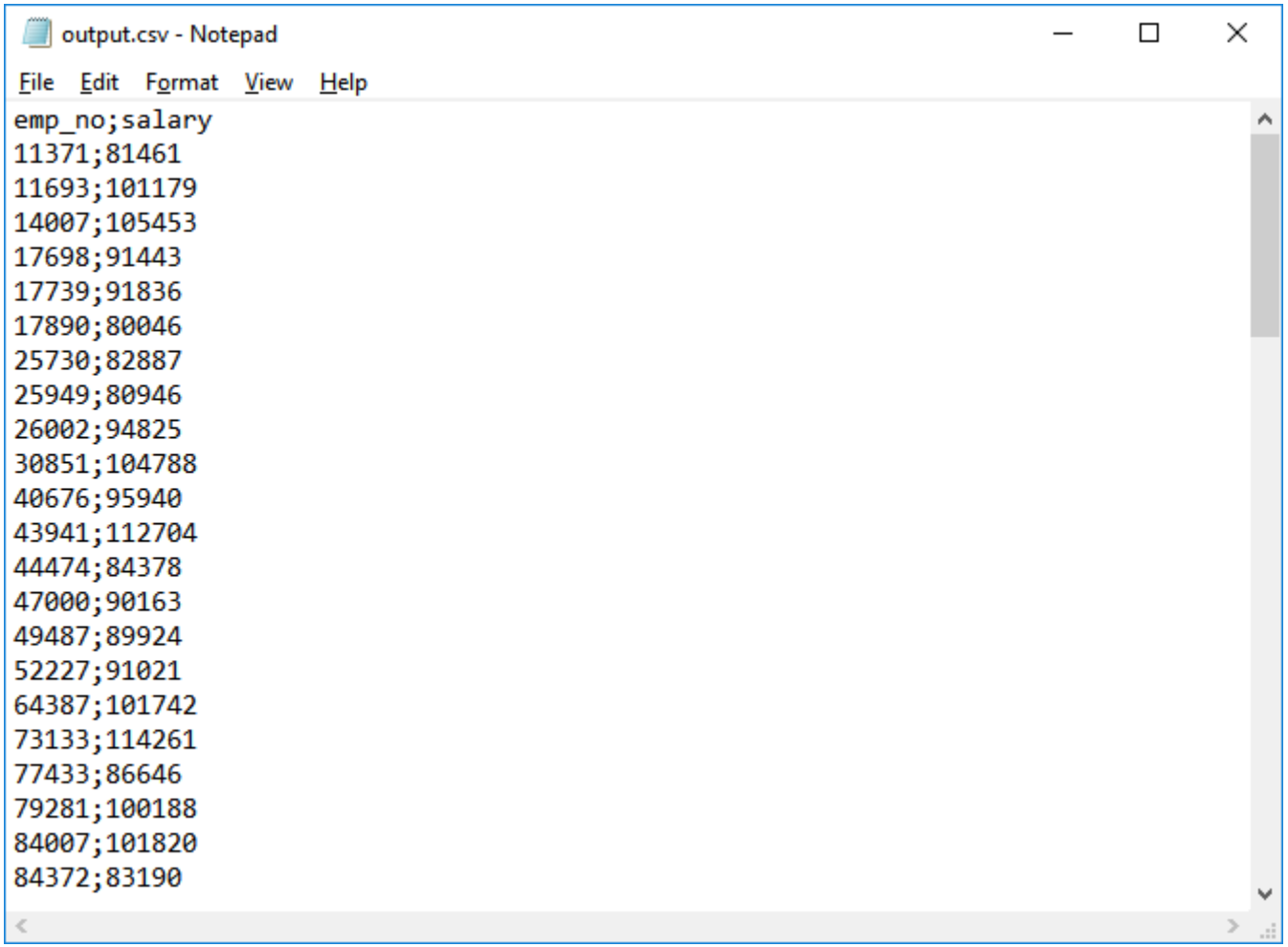


#	emp_no	salary
1	11371	81461
2	11693	101179
3	14007	105453
4	17698	91443
5	17739	91836
6	17890	80046
7	25730	82887
8	25949	80946
9	26002	94825
10	30851	104788
11	40676	95940
12	43941	112704
13	44474	84378
14	47000	90163
15	49487	89924
16	52227	91021

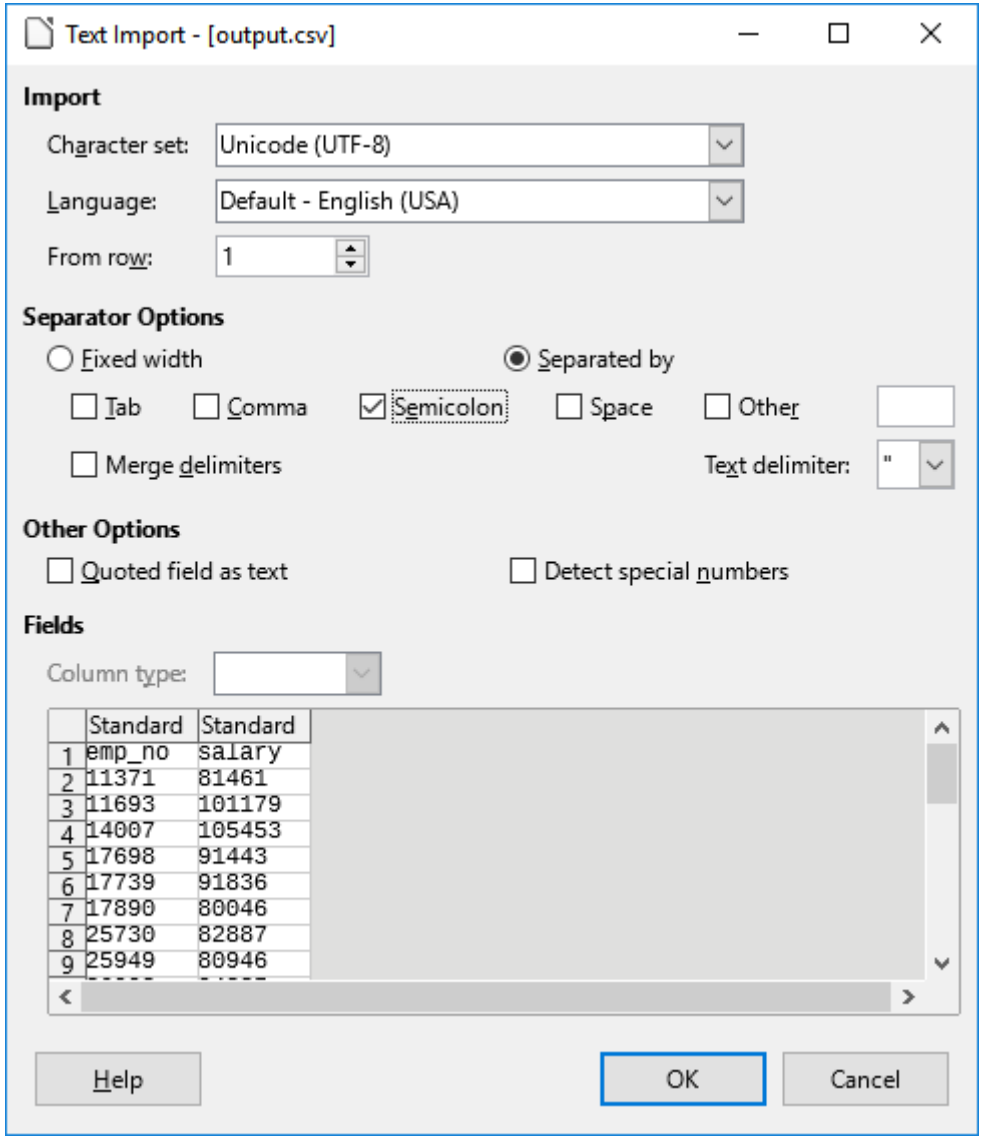
Text file output



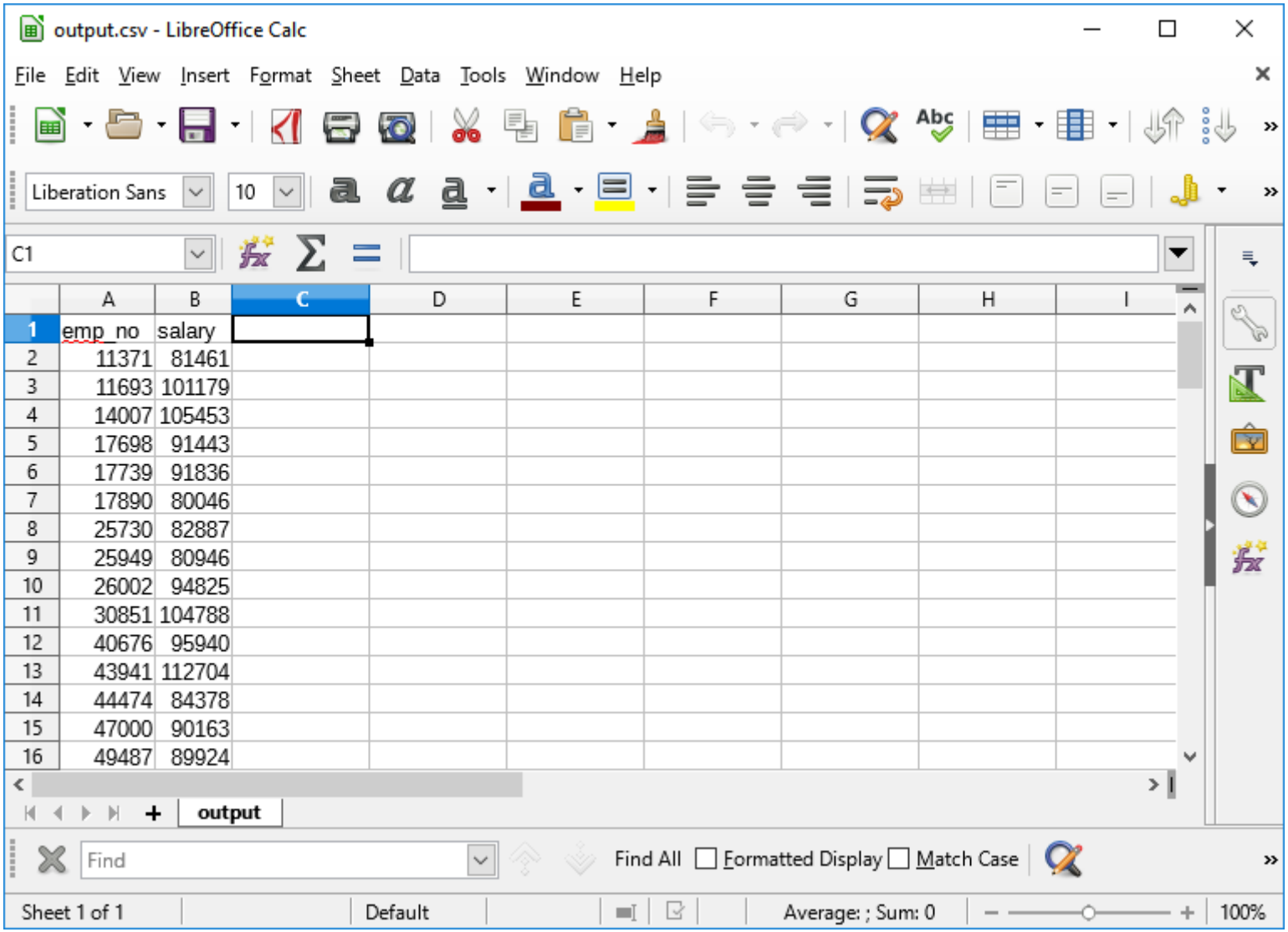
Output


output.csv - Notepad
File Edit Format View Help
emp_no;salary
11371;81461
11693;101179
14007;105453
17698;91443
17739;91836
17890;80046
25730;82887
25949;80946
26002;94825
30851;104788
40676;95940
43941;112704
44474;84378
47000;90163
49487;89924
52227;91021
64387;101742
73133;114261
77433;86646
79281;100188
84007;101820
84372;83190

Output

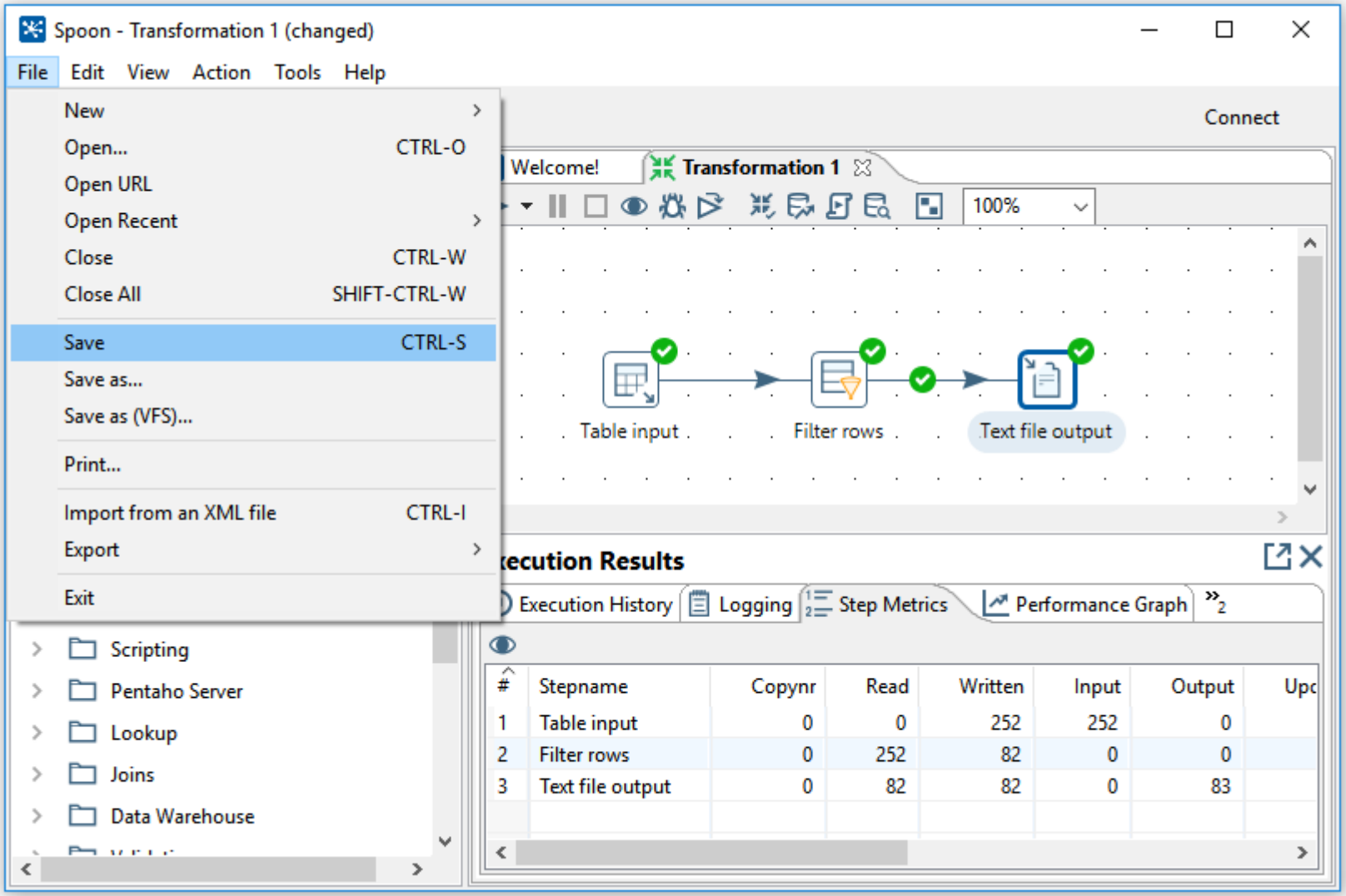

Text Import - [output.csv]
Import
Character set: Unicode (UTF-8)
Language: Default - English (USA)
From row: 1
Separator Options
Fixed width
Separated by
Tab
Comma
Semicolon
Space
Other
Merge delimiters
Text delimiter: "
Other Options
Quoted field as text
Detect special numbers
Fields
Column type:
Standard Standard
1 emp_no salary
2 11371 81461
3 11693 101179
4 14007 105453
5 17698 91443
6 17739 91836
7 17890 80046
8 25730 82887
9 25949 80946
Help
OK
Cancel

Output



	A	B	C	D	E	F	G	H	I
1	emp_no	salary							
2	11371	81461							
3	11693	101179							
4	14007	105453							
5	17698	91443							
6	17739	91836							
7	17890	80046							
8	25730	82887							
9	25949	80946							
10	26002	94825							
11	30851	104788							
12	40676	95940							
13	43941	112704							
14	44474	84378							
15	47000	90163							
16	49487	89924							

Save transformation



Run transformation