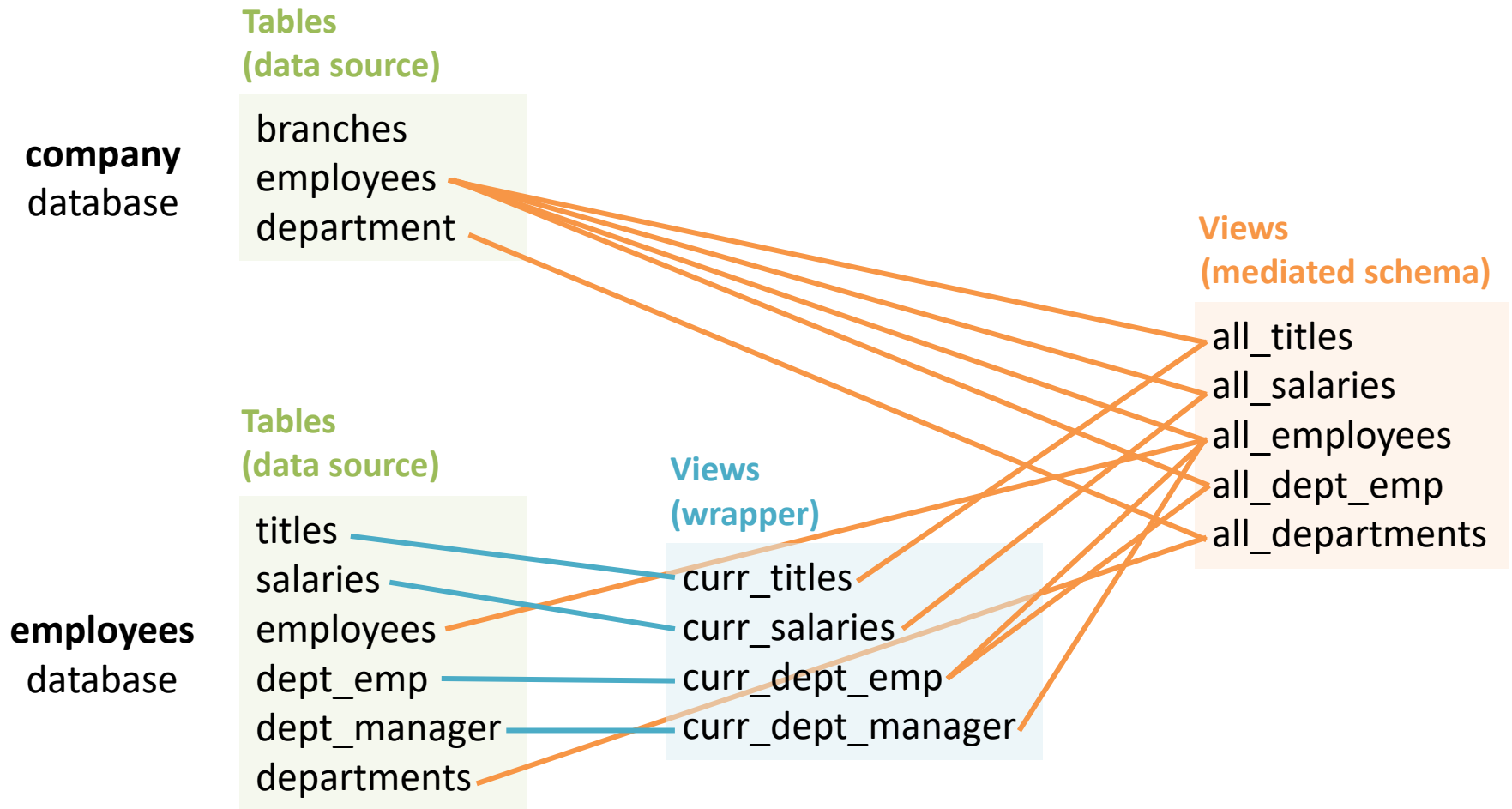


# Data Analysis and Integration

## Introduction to ETL tools

ETL:  
Extract  
Transform  
e Load

# Data integration using views

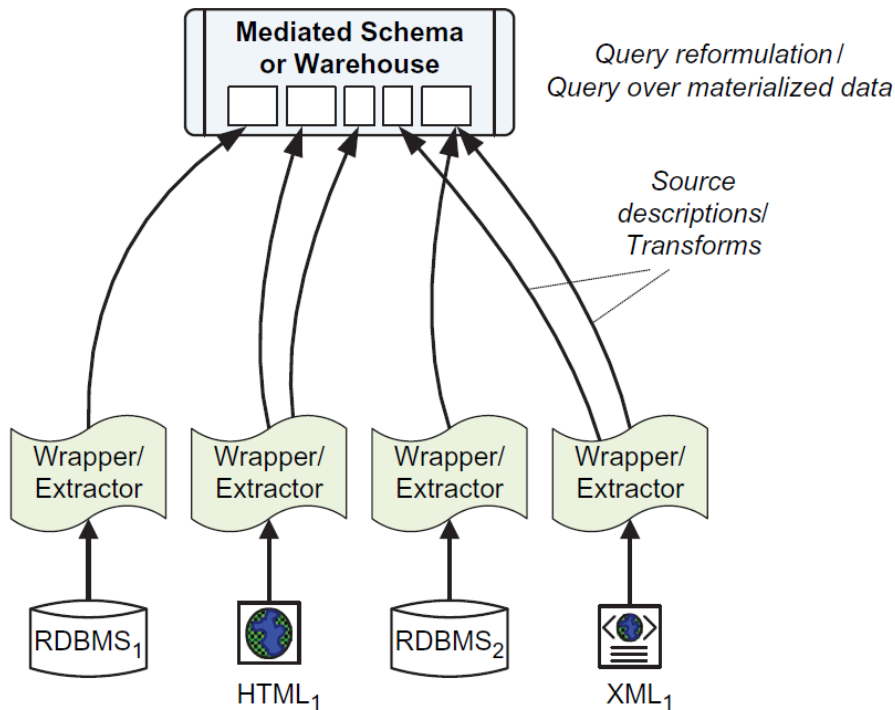


# Data integration using views

- 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

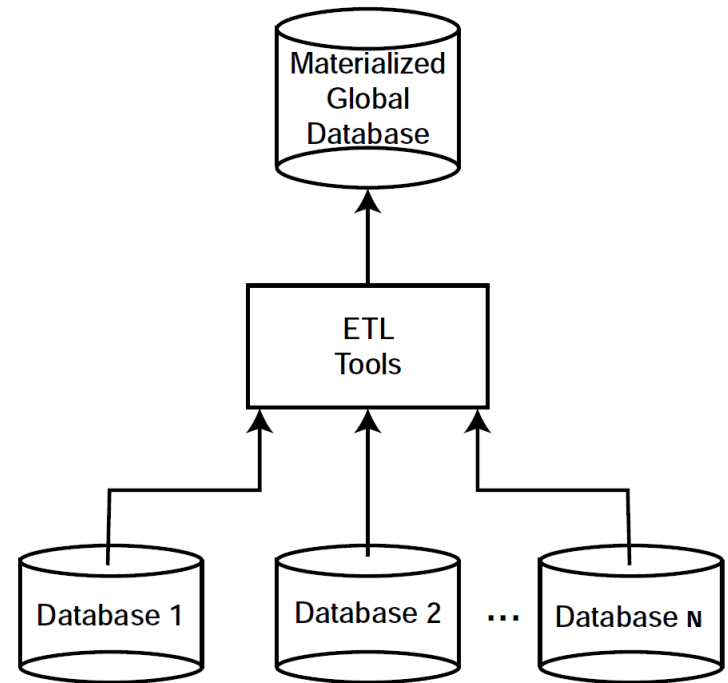
# Data integration vs. data warehousing

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

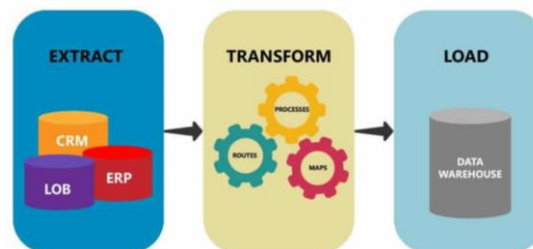
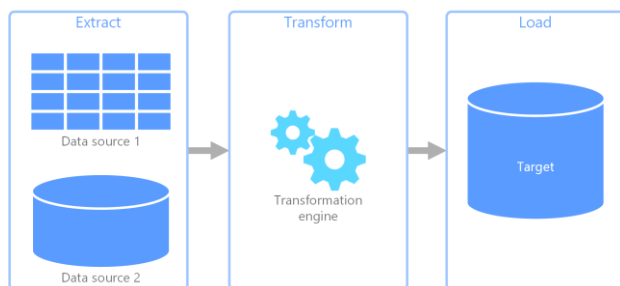
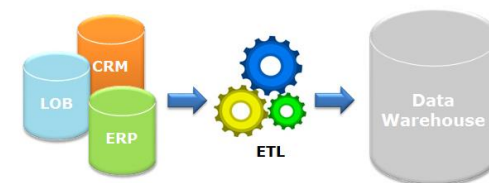
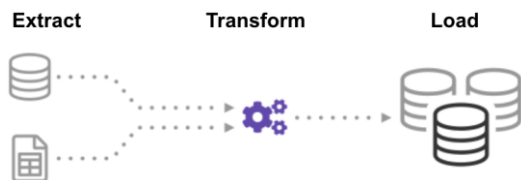
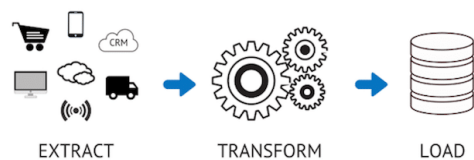
# Data warehousing

- 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

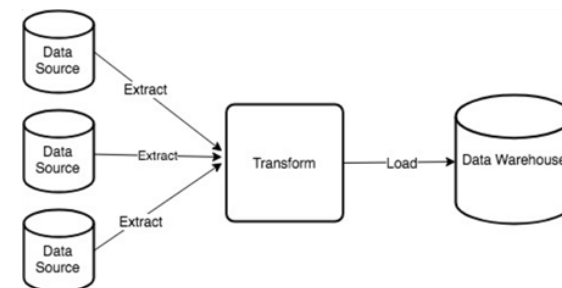
# ETL process

(Pentaho is an ETL tool)

- Extract, transform, load

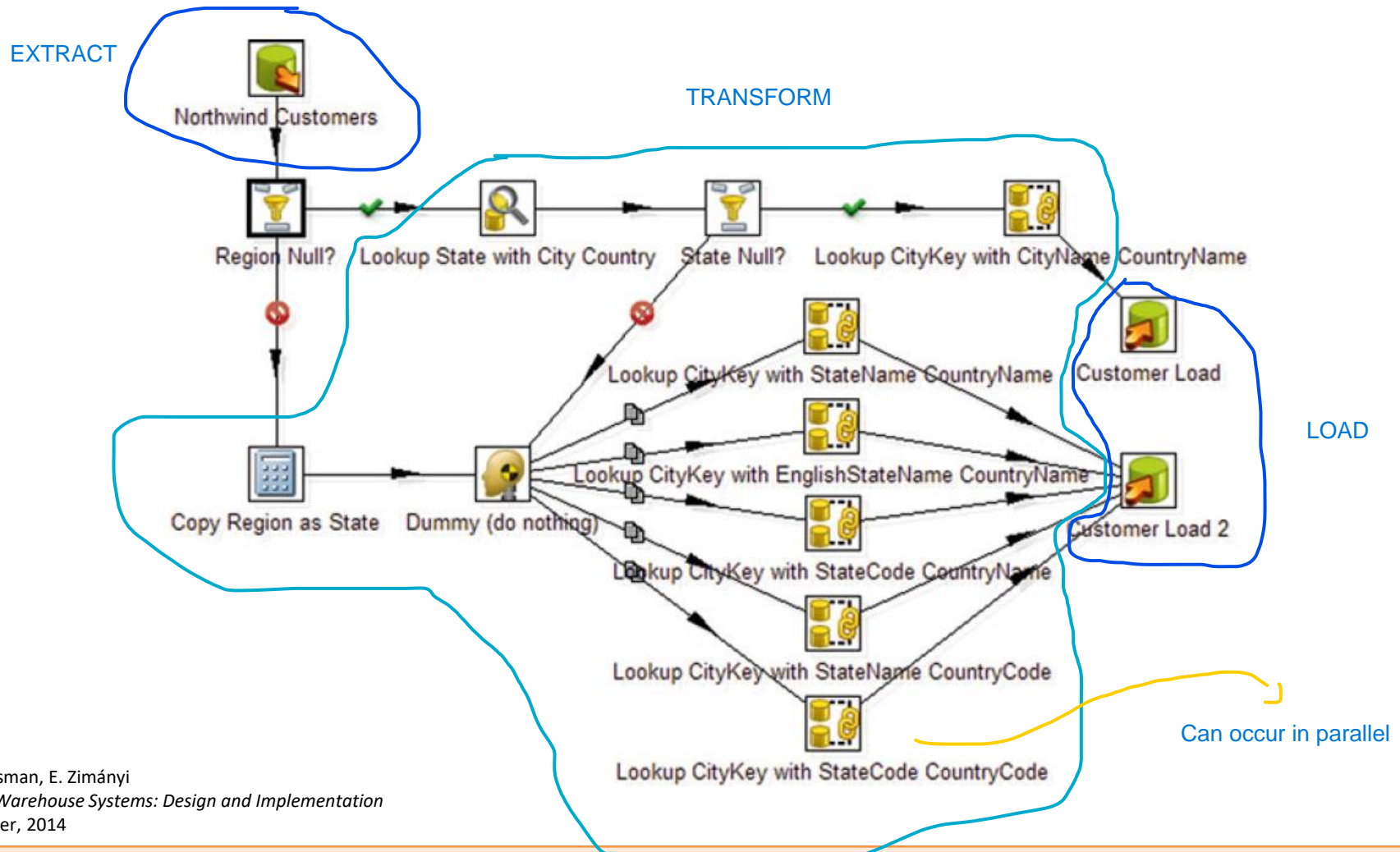


ETL - Extract, Transform, Load



# ETL process

- Extracting, transforming and loading customers



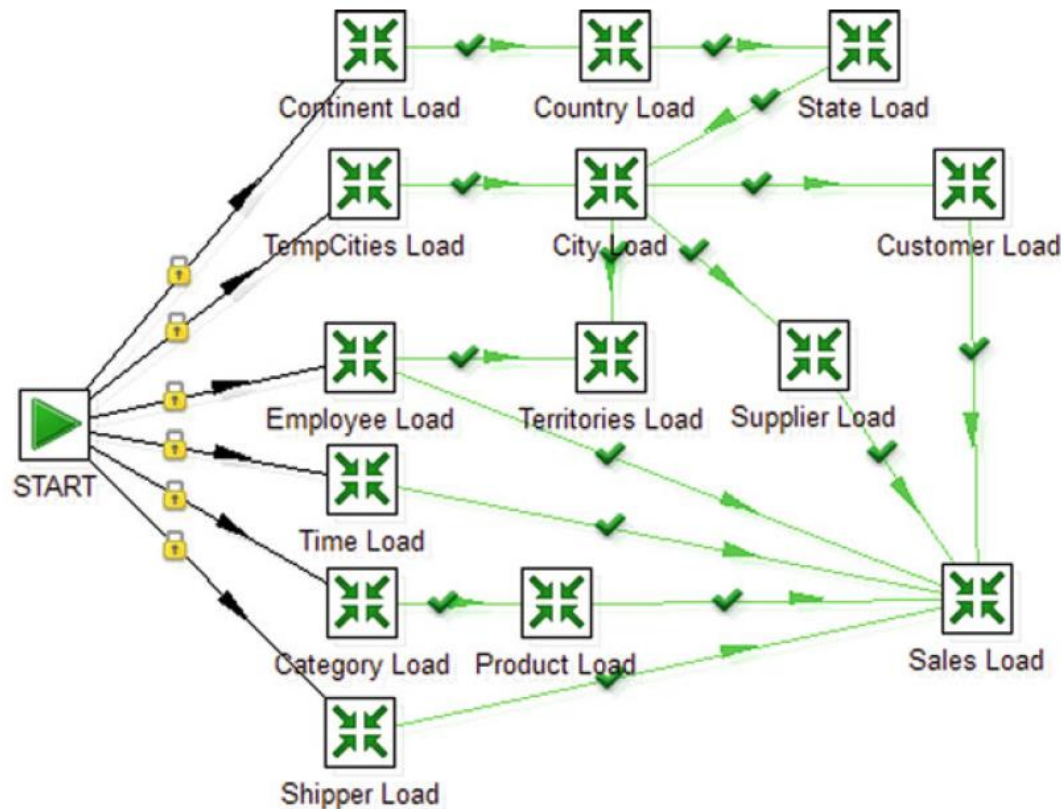
# ETL process

- An ETL process comprises many such transformations

We might need to transform terabytes of data!

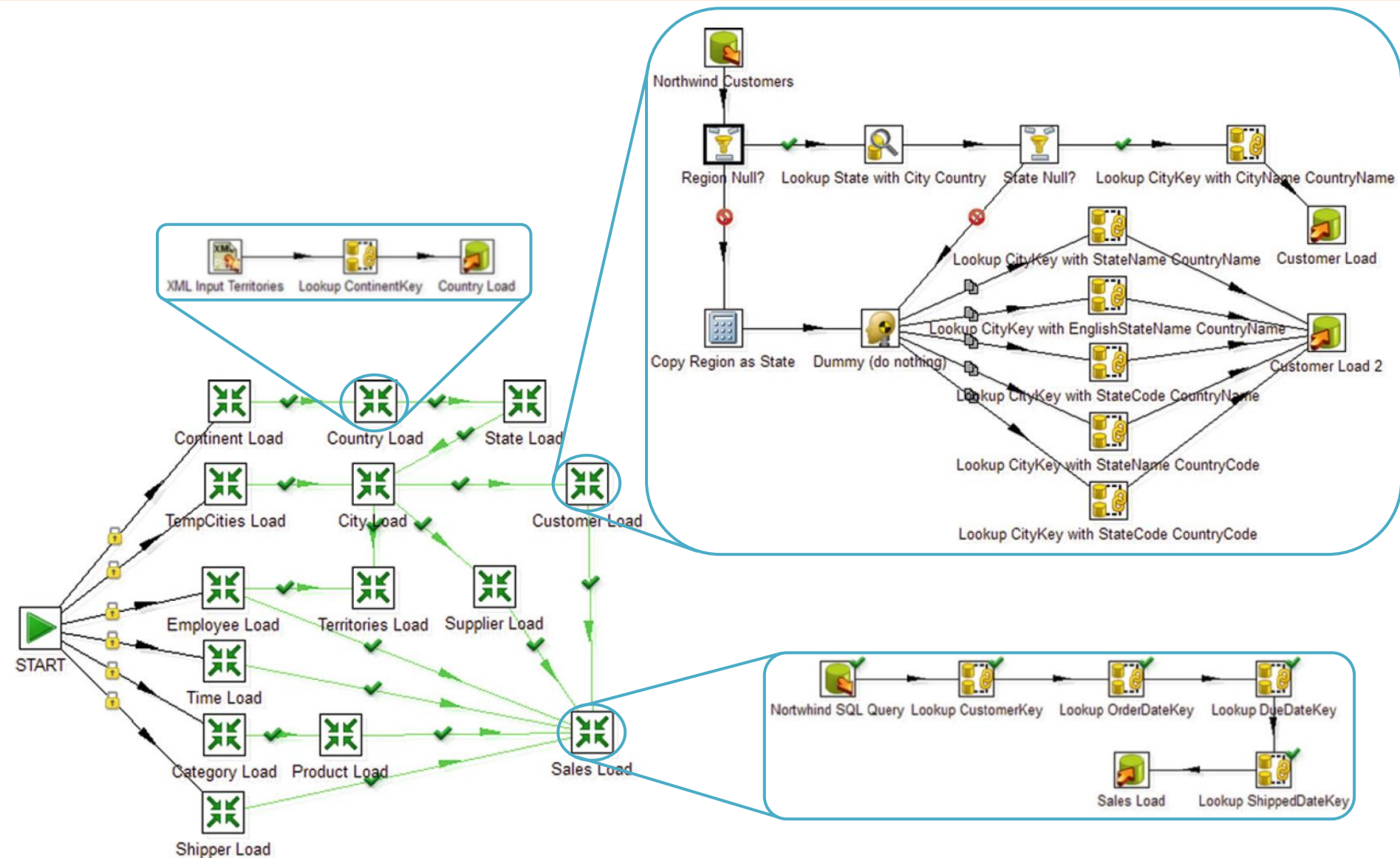
Not many computers can handle that  
So! We pipeline/stream this system

All steps may occur at the same time





# ETL process

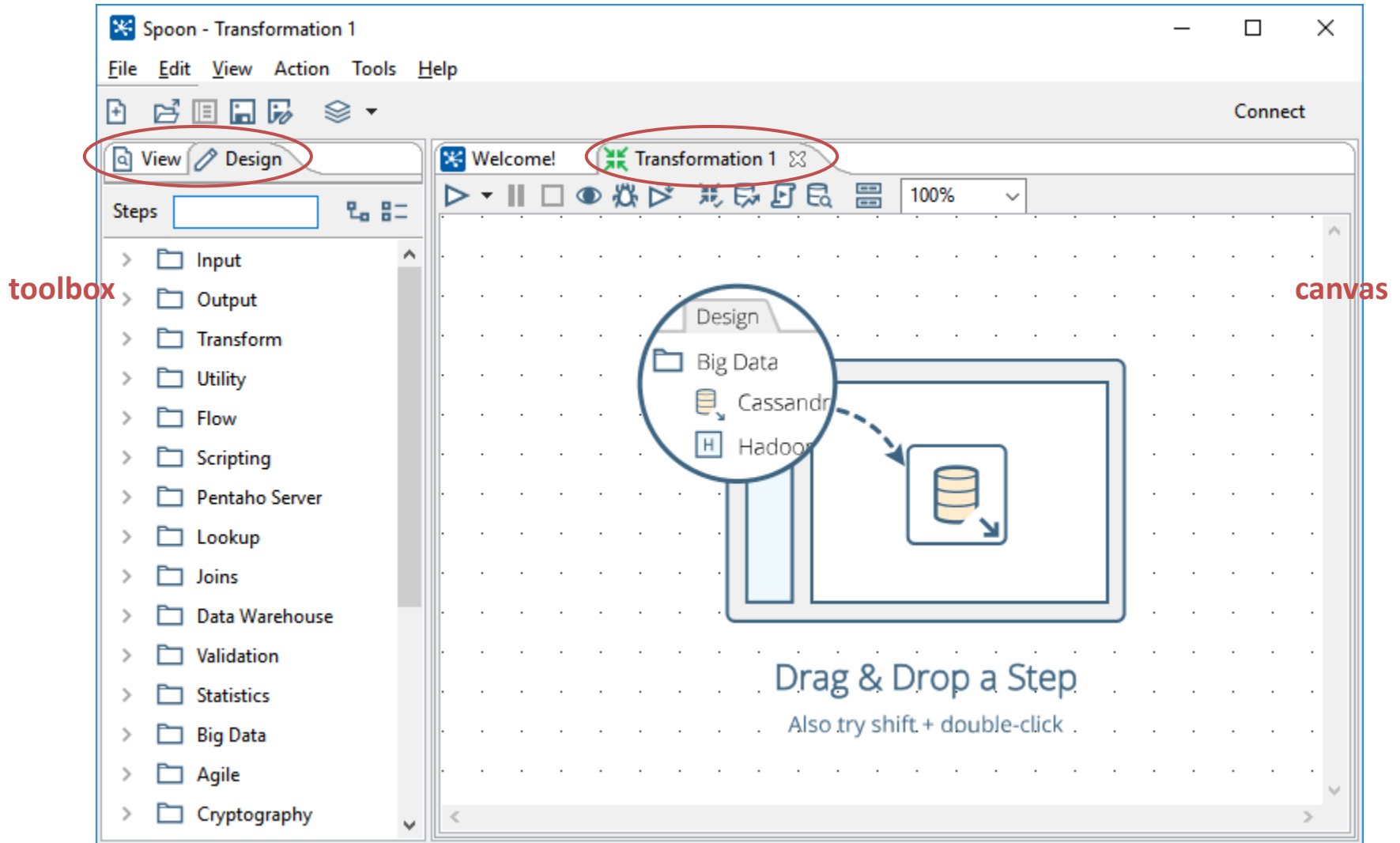


# ETL tools

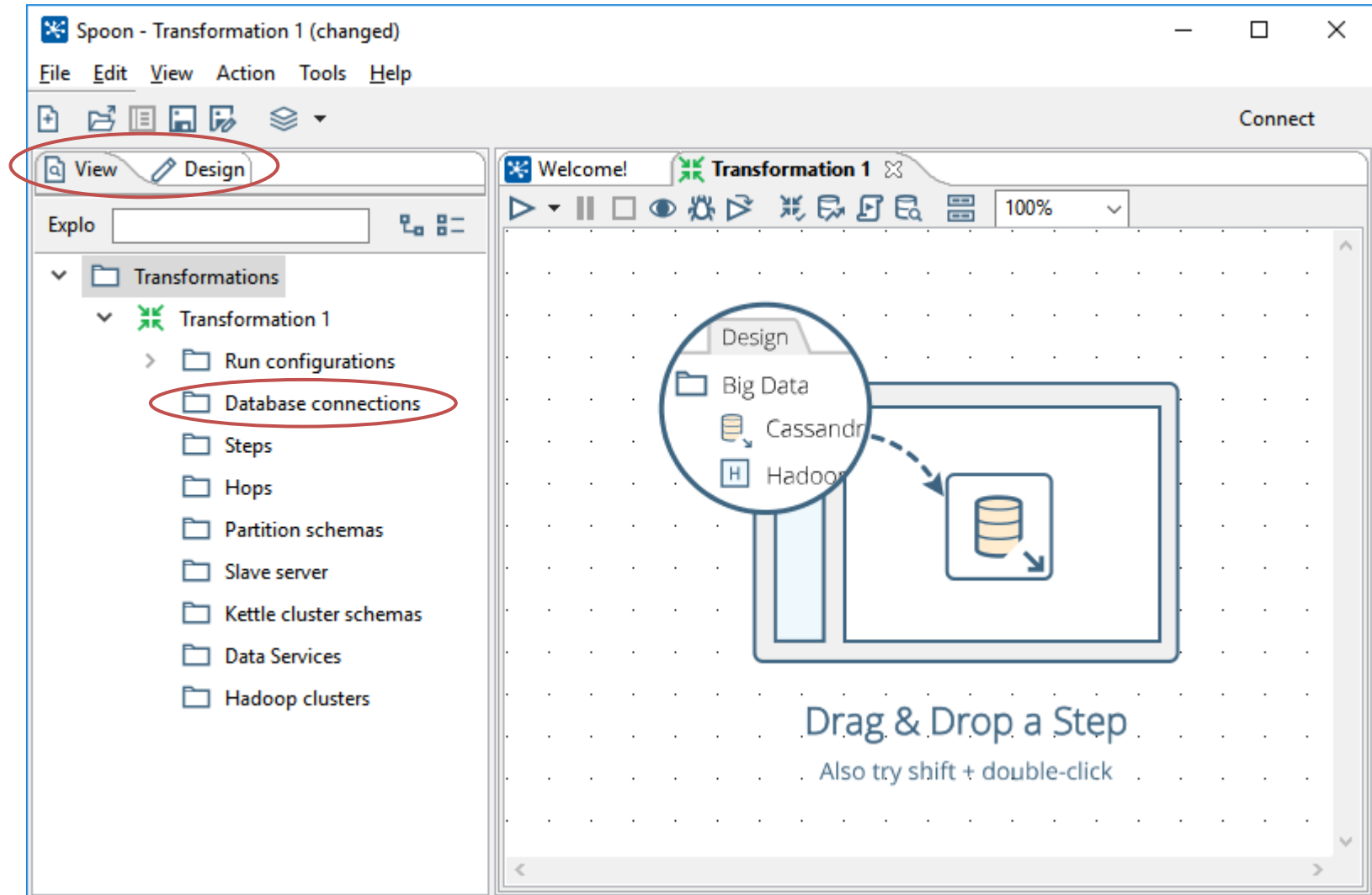
- ETL tools
  - what do they do?
  - how do they work?
  - how can we use them?
- The ETL tool that we will be using:
  - Pentaho Data Integration
  - also known as PDI, Kettle, or Spoon
  - competing products (e.g. SQL Server Integration Services)

# Workspace

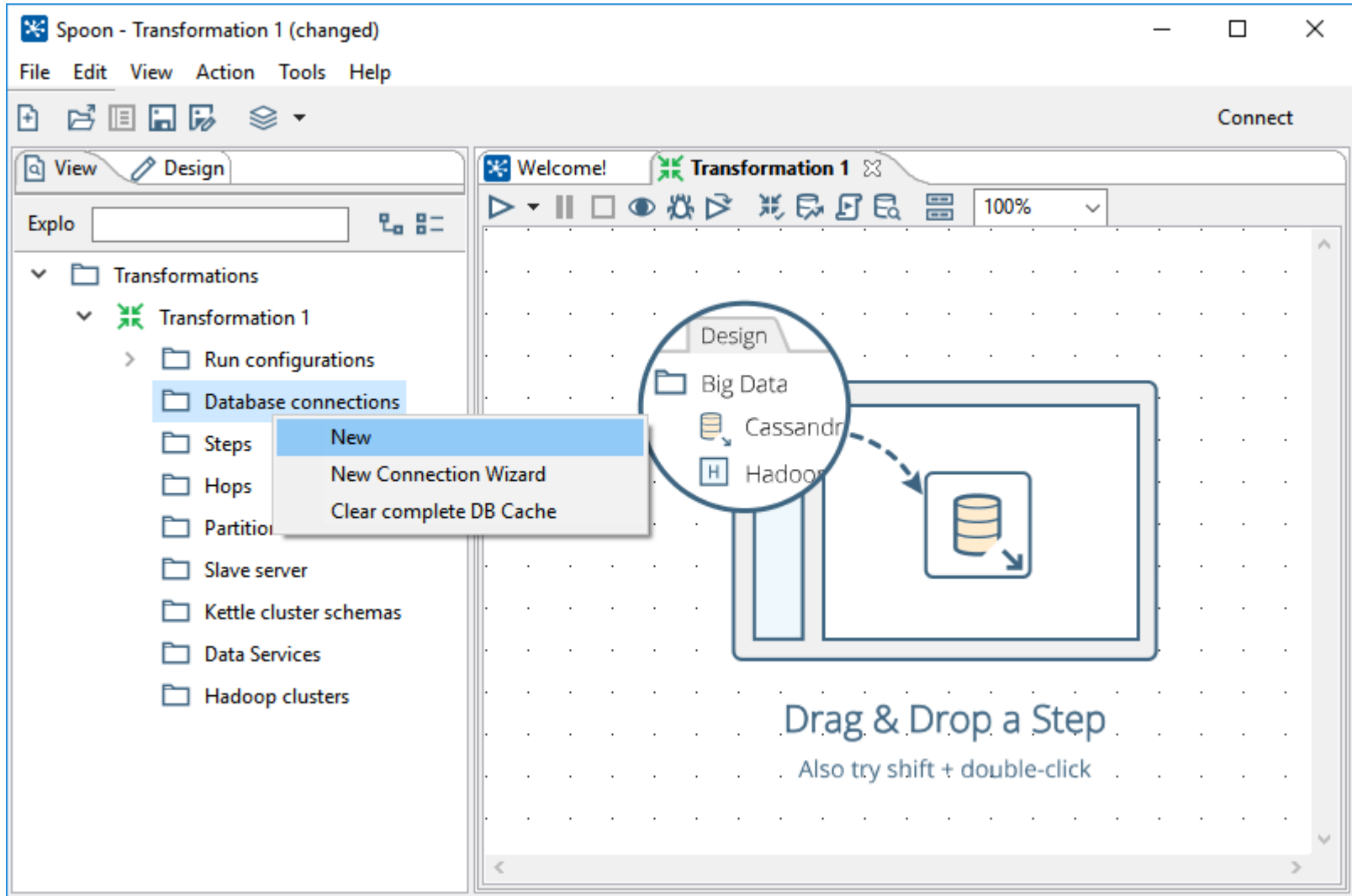
Every step in the tool must be configured



# Database connections



# New database connection



# New database connection

Database Connection

General  
Advanced  
Options  
Pooling  
Clustering

Connection Name:  
employees

Connection Type:  
MS SQL Server (Native)  
MariaDB  
MaxDB (SAP DB)  
MonetDB  
MySQL  
Native Mondrian  
Neoview  
Netezza  
OpenERP Server  
Oracle  
Oracle RDB

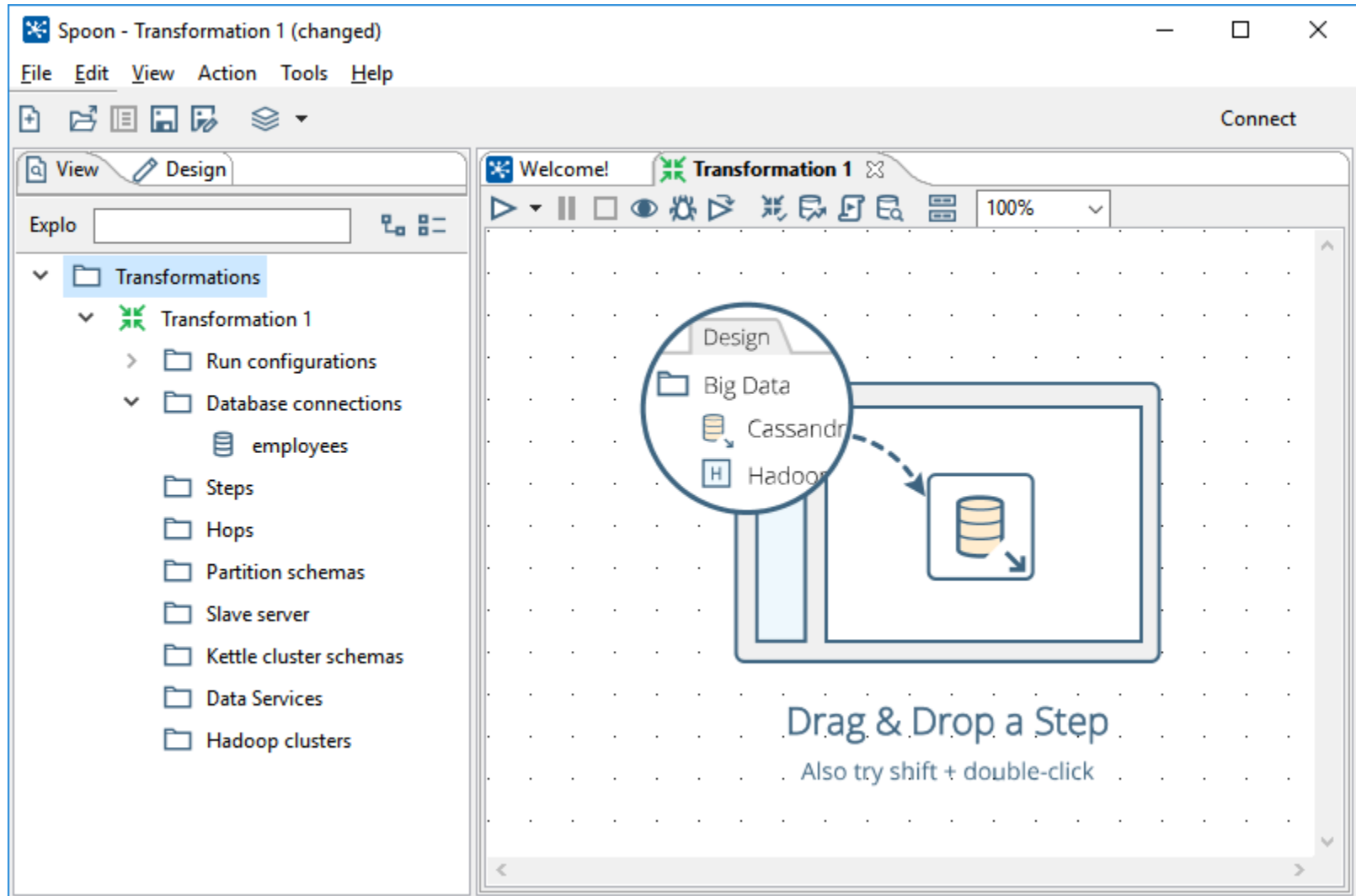
Access:  
Native (JDBC)  
ODBC  
JNDI

Settings  
Host Name:  
localhost  
Database Name:  
employees  
Port Number:  
3306  
User Name:  
aid  
Password:  
...  
☒ Use Result Streaming Cursor

Test Feature List Explore

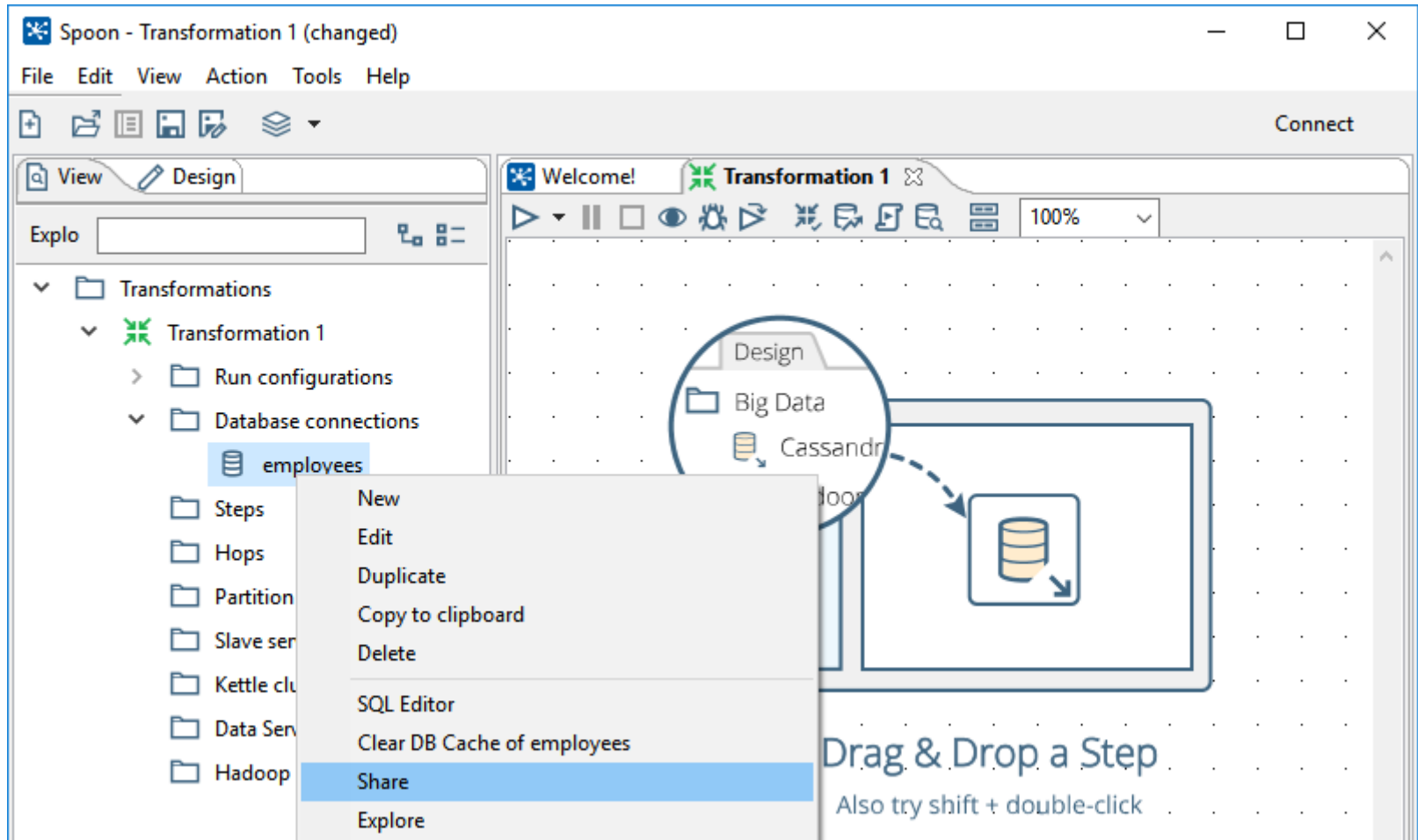
OK Cancel

# New database connection



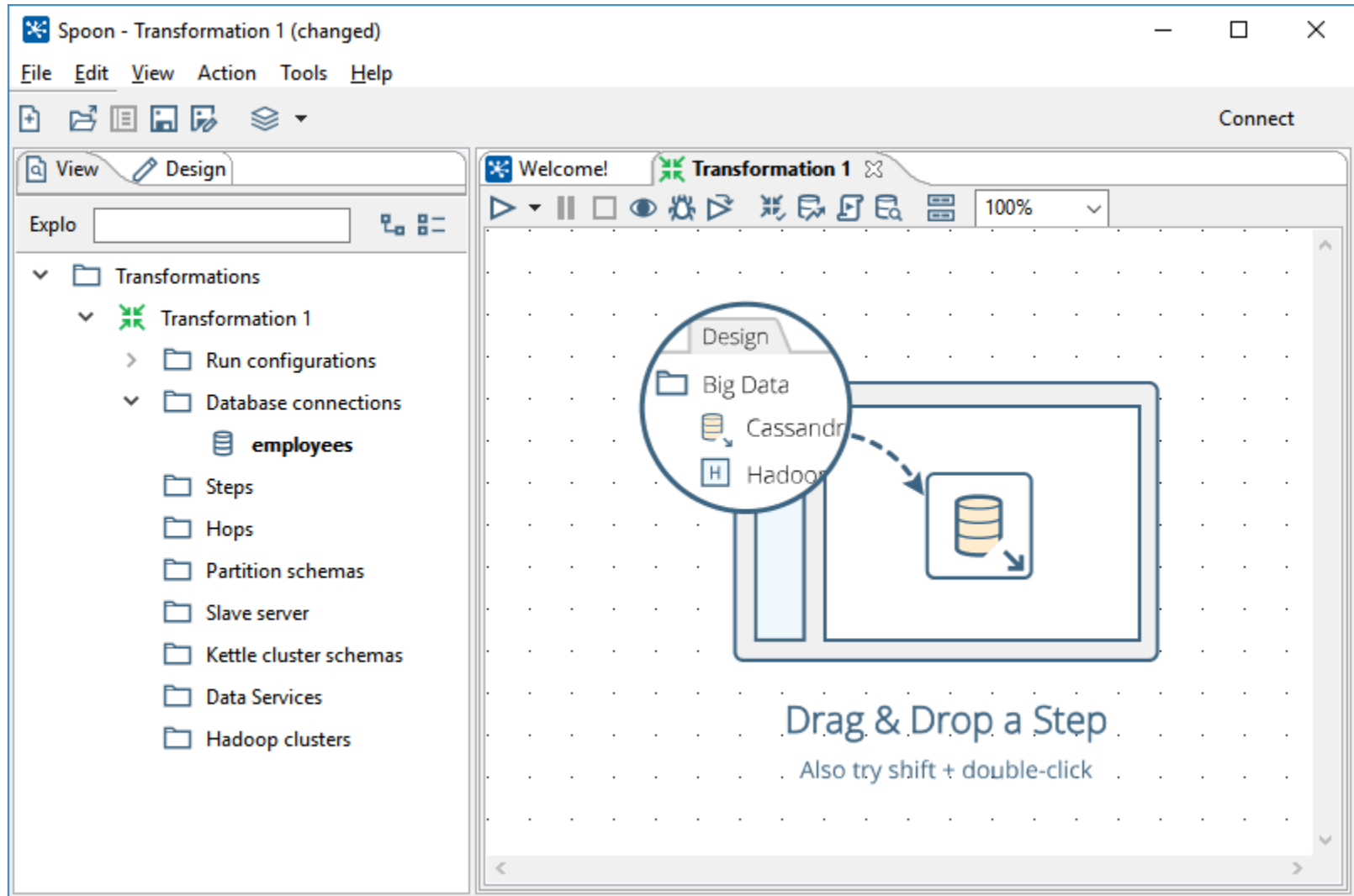
# New database connection

- Using the same connection in multiple transformations

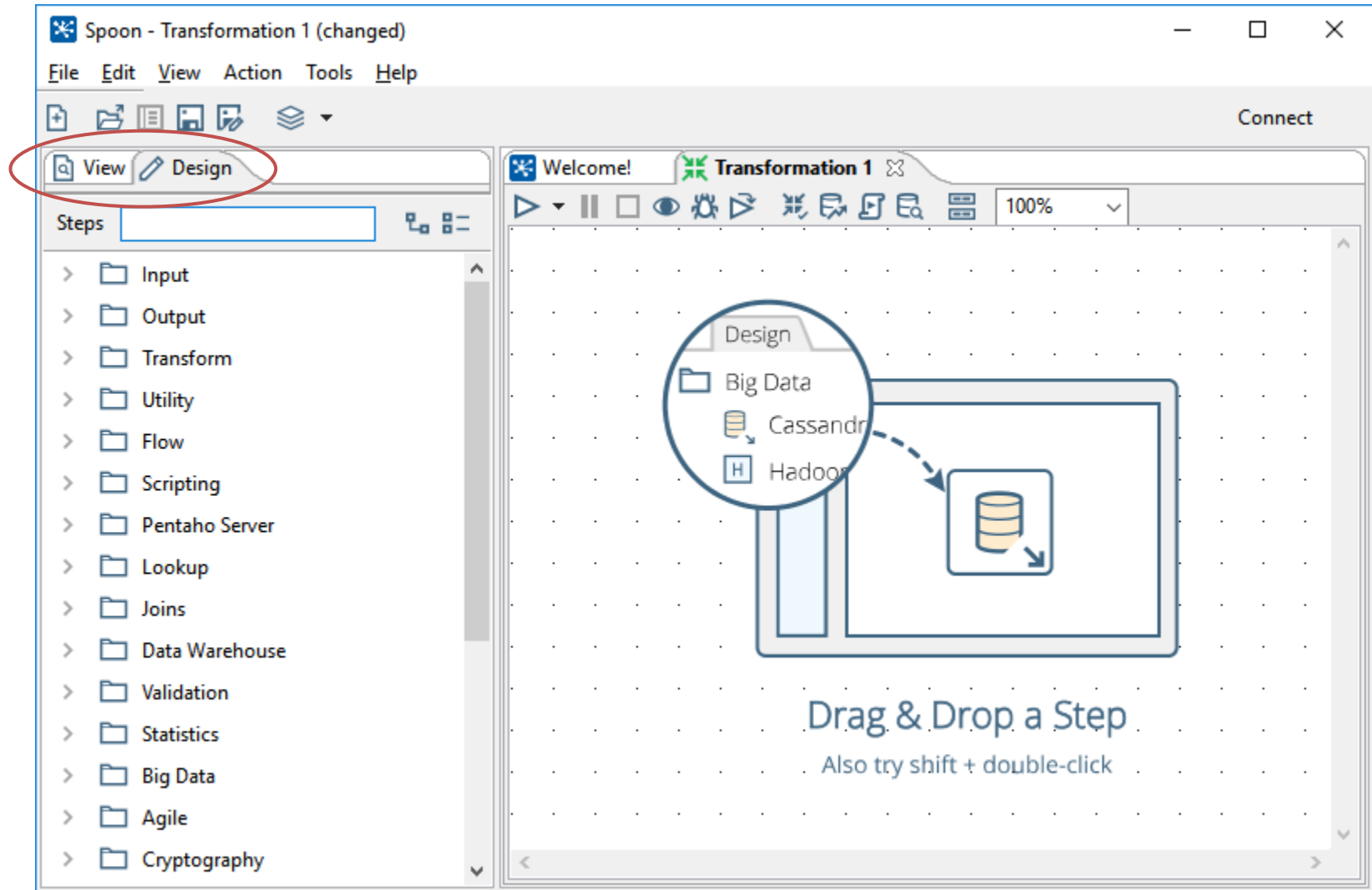




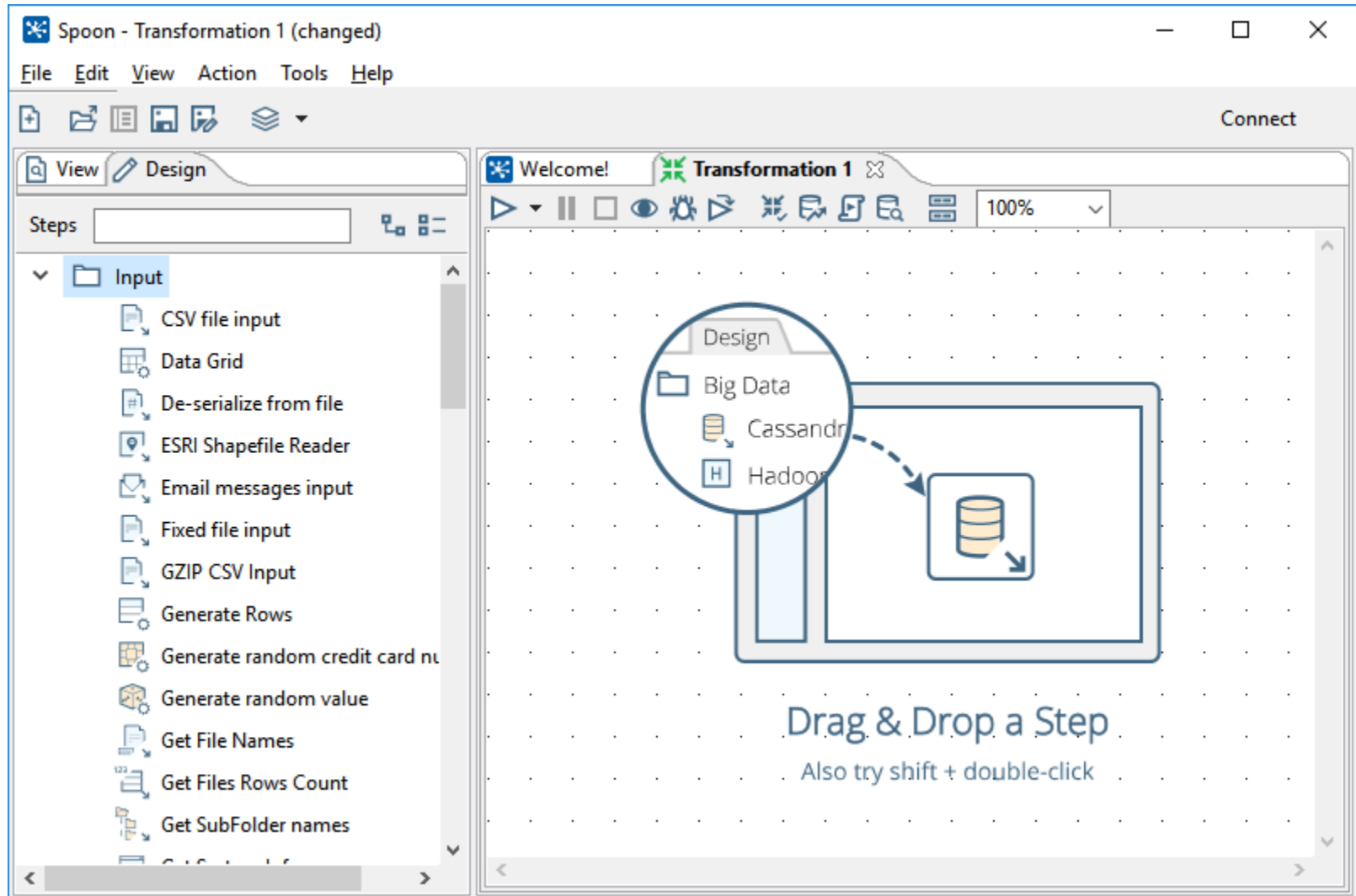
# New database connection



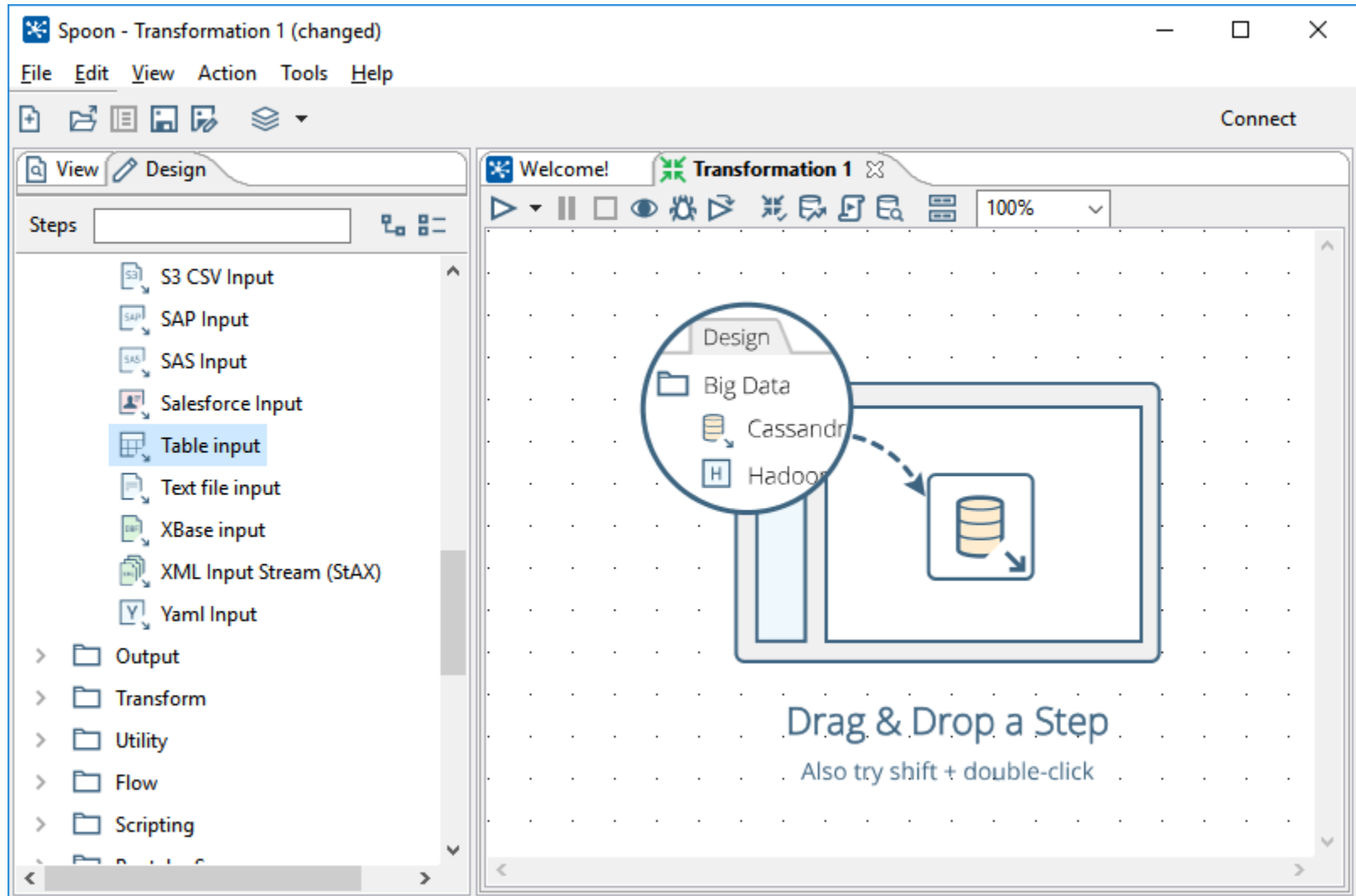
# Table input



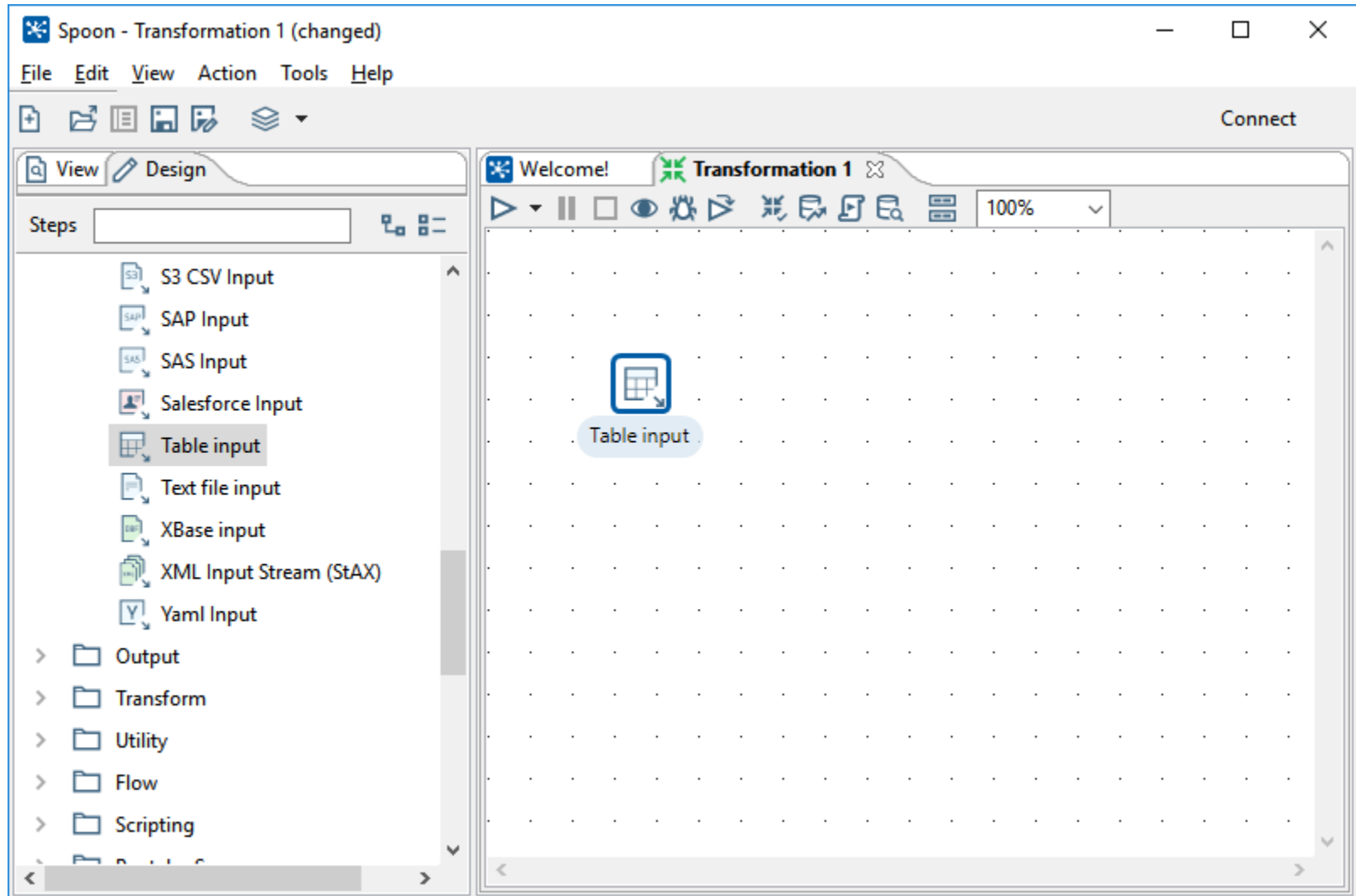
# Table input



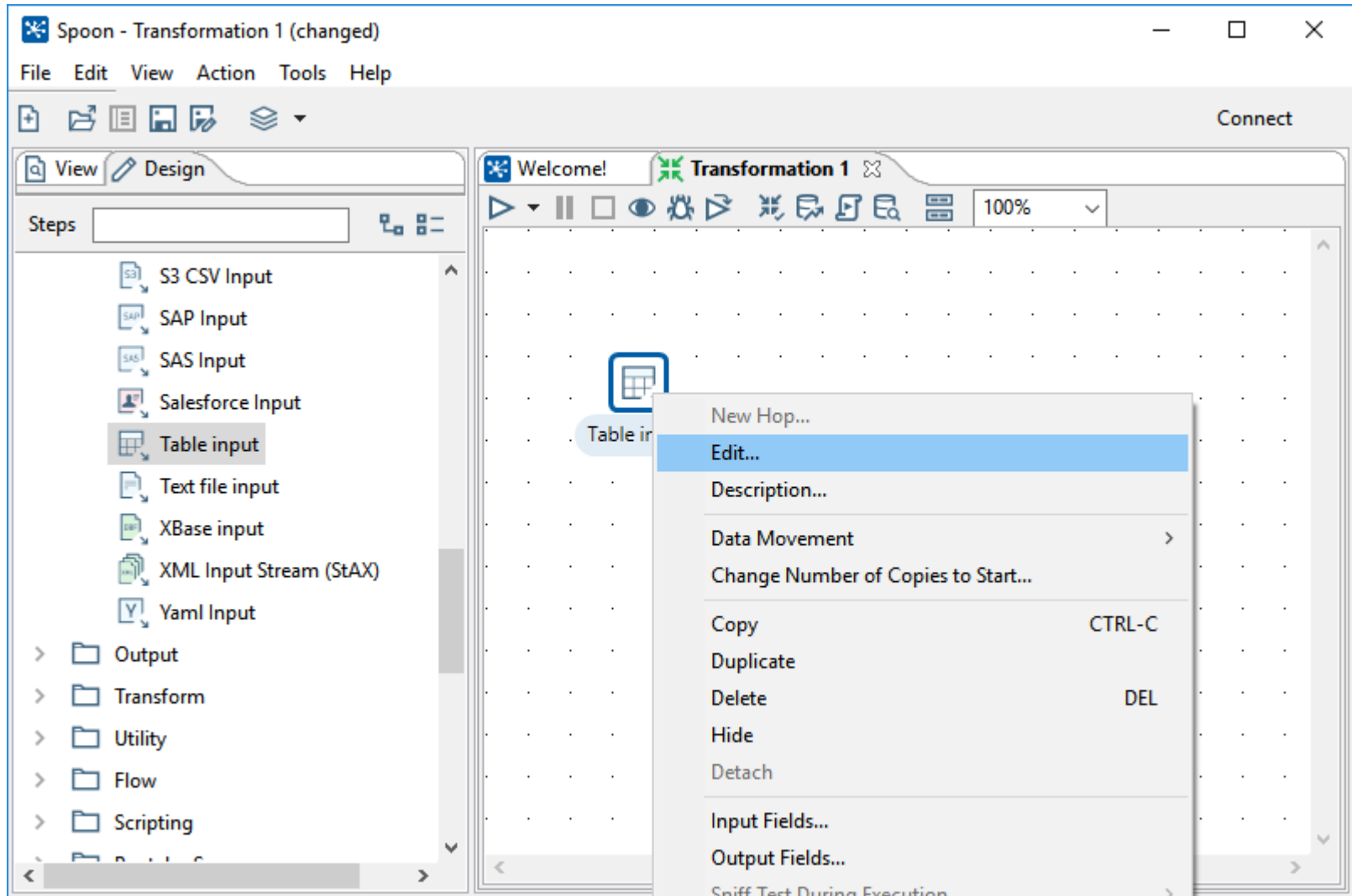
# Table input



# Table input



# Table input



# 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? ☐

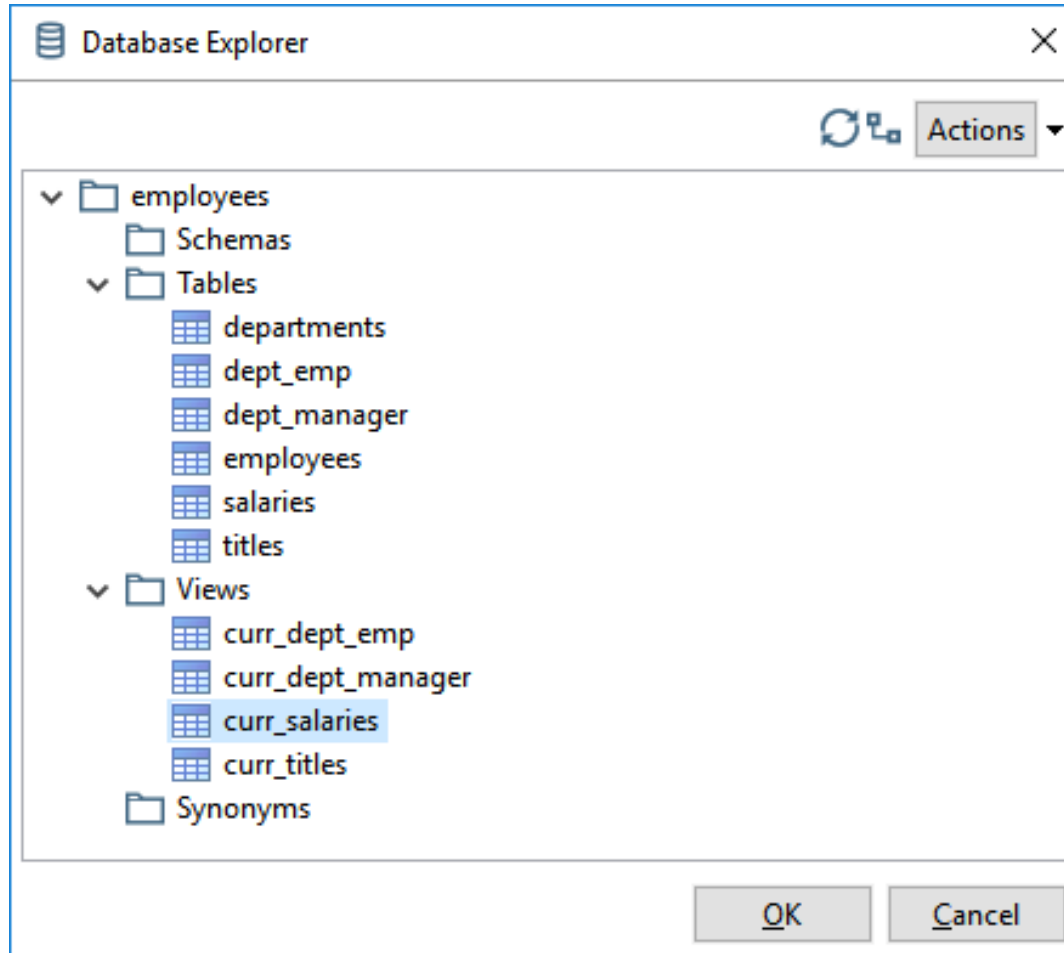
Insert data from step

Execute for each row? ☐

Limit size: 0

Help OK Preview Cancel

# Table input





# 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? ☐

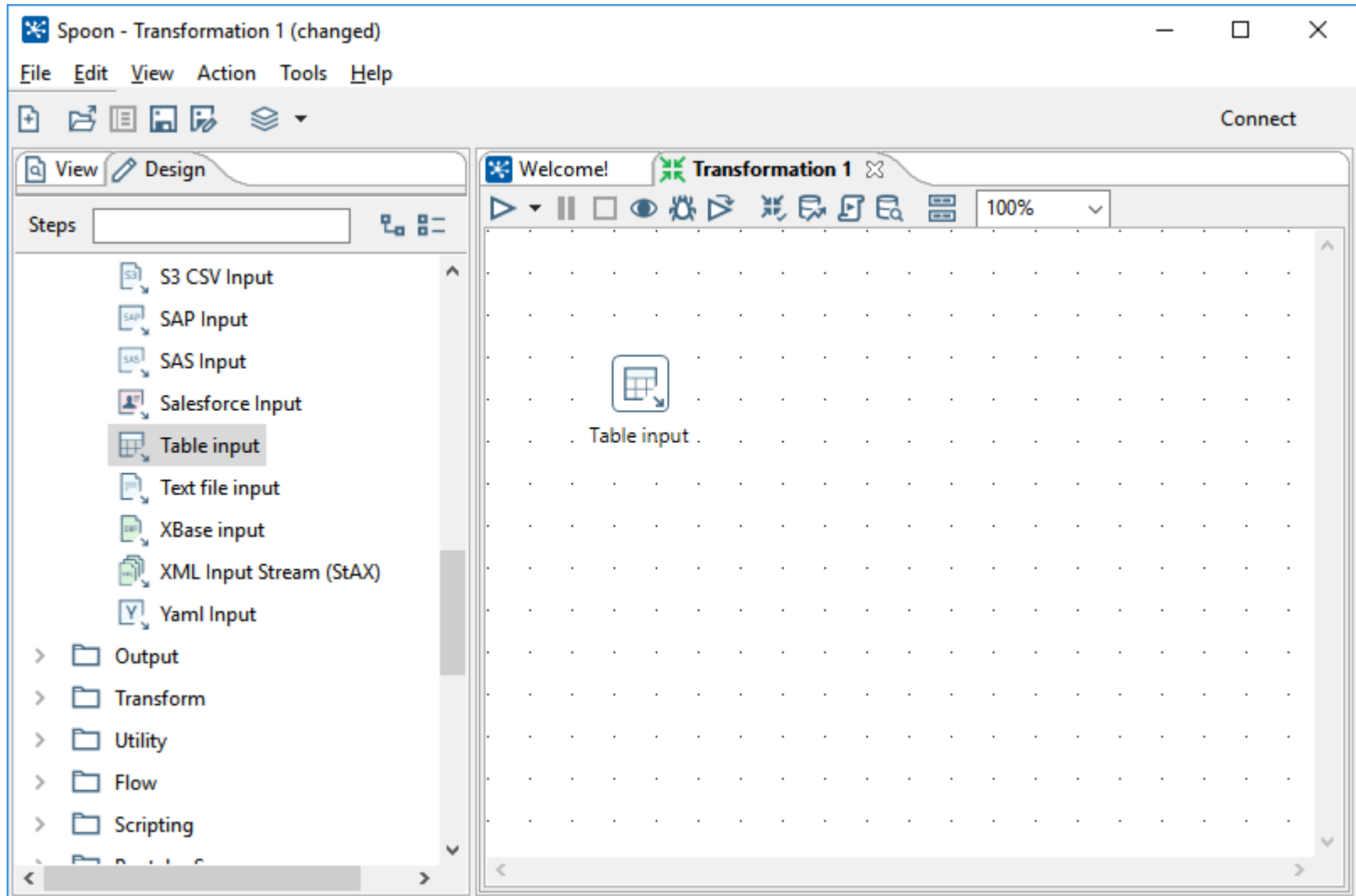
Insert data from step

Execute for each row? ☐

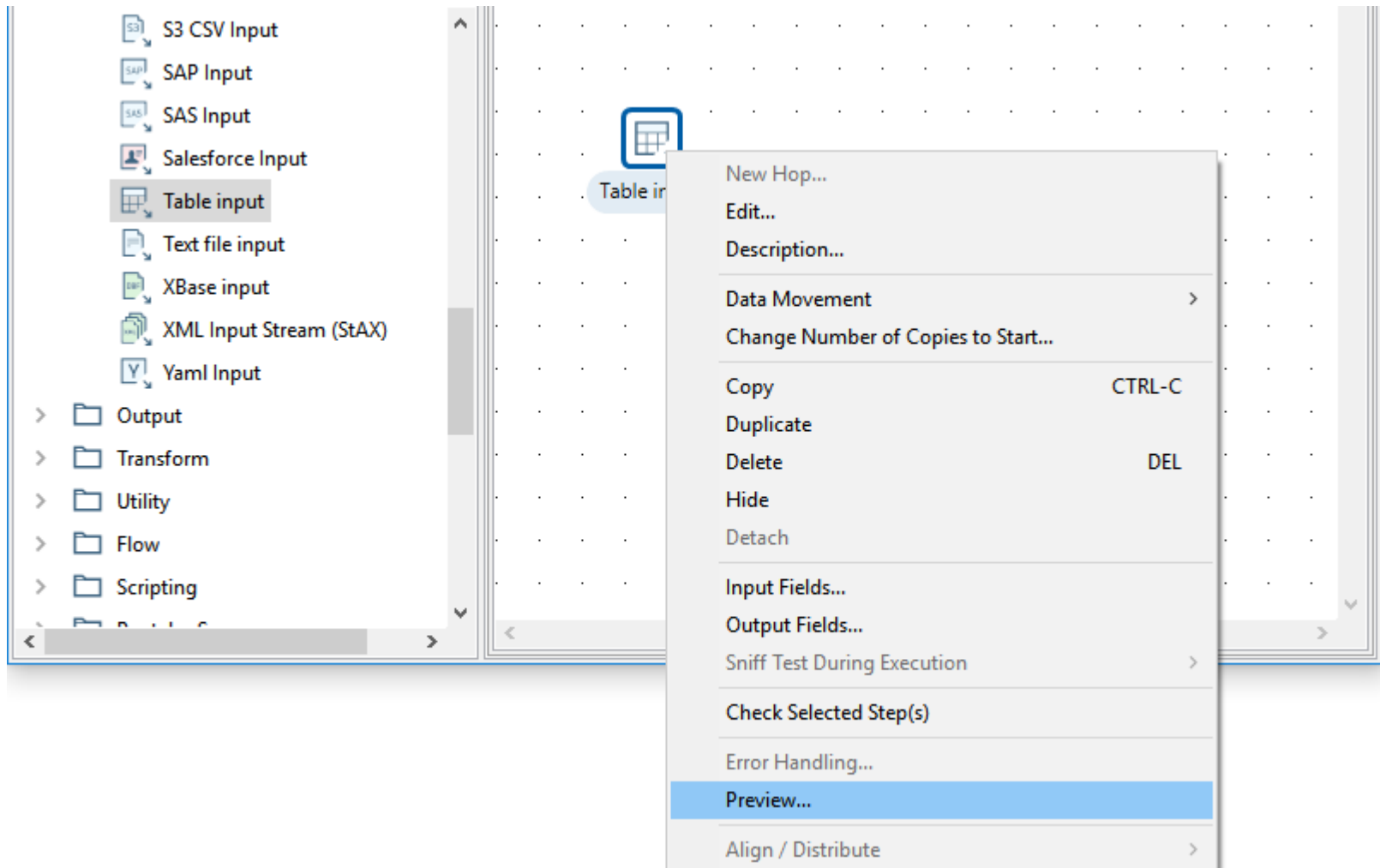
Limit size: 0

Help OK Preview Cancel

# Table input



# Table input



# 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

Examine preview data

Rows of step: Table input (252 rows)

#	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

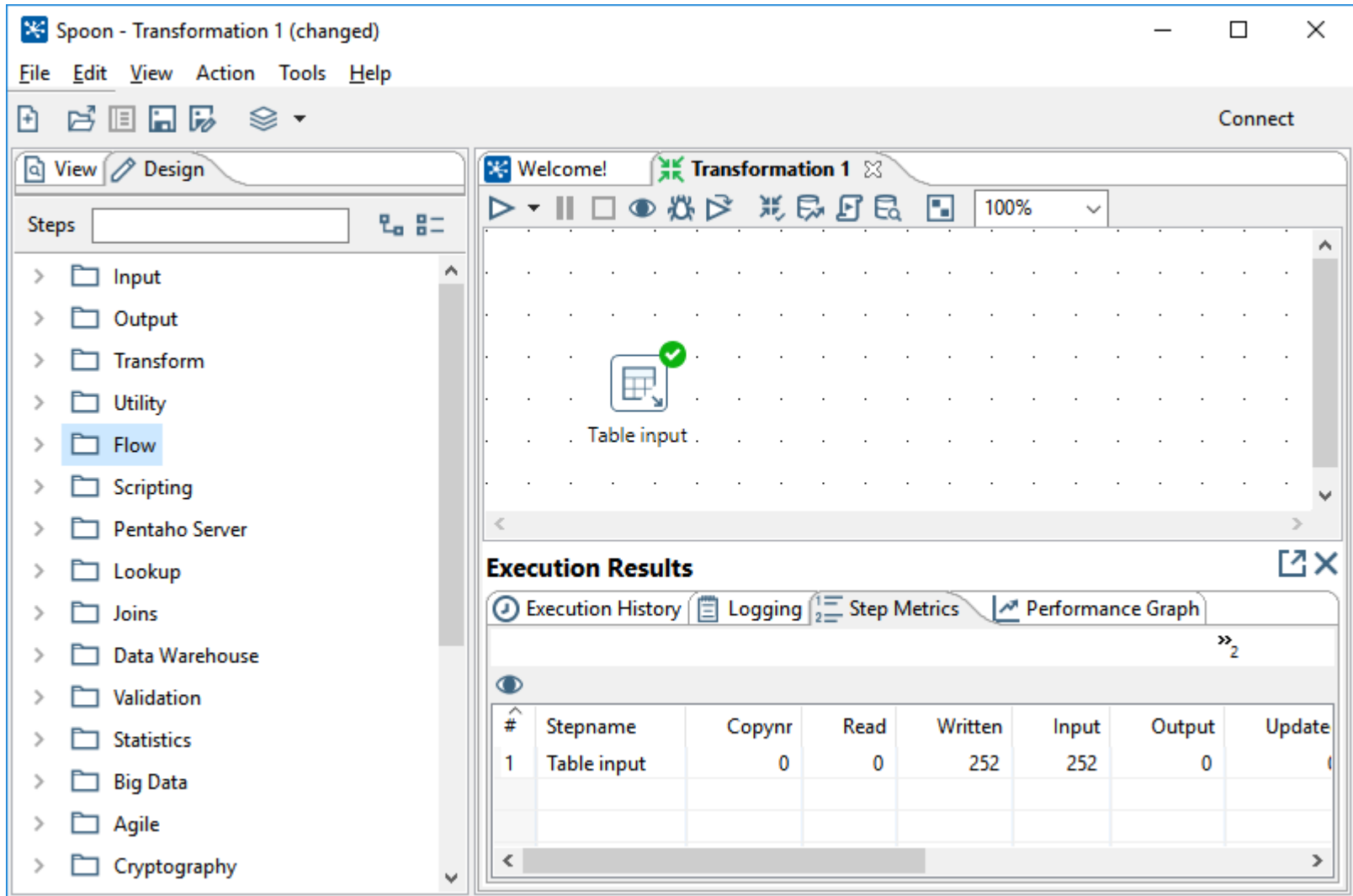
Close

# Table input

The screenshot shows the Spoon - Transformation 1 (changed) window. The left pane displays the 'Steps' list with 'Table input' selected. The right pane shows the 'Design' view with a 'Table input' step icon. Below the design view is the 'Execution Results' section, which includes a table with the following data:

#	Stepname	Copynr	Read	Written	Input	Output	Update
1	Table input	0	0	252	252	0	

# Filter rows



The screenshot shows the Spoon - Transformation 1 (changed) window. The left sidebar contains a tree view of steps, with 'Flow' selected. The main canvas displays the 'Table input' step in the Design view. The 'Execution Results' tab is active, showing a table with the following data:

#	Stepname	Copynr	Read	Written	Input	Output	Update
1	Table input	0	0	252	252	0	

# Filter rows

The screenshot shows the Spoon - Transformation 1 (changed) window. The 'Steps' list on the left includes various transformation steps, with 'Filter rows' highlighted. The main canvas displays a 'Table input' step. The 'Execution Results' section at the bottom shows the 'Execution History' tab with a table of results.

#	Stepname	Copynr	Read	Written	Input	Output	Update
1	Table input	0	0	252	252	0	



# Filter rows

The screenshot shows the SAP Data Services Spoon interface. The main workspace displays a transformation named 'Transformation 1' with two steps: 'Table input' and 'Filter rows'. The 'Filter rows' step is highlighted with a blue selection box. The left sidebar shows the 'Steps' list, with 'Filter rows' selected. The bottom right pane shows the 'Execution Results' table.

**Execution Results**

#	Stepname	Copynr	Read	Written	Input	Output	Update
1	Table input	0	0	252	252	0	

# Filter rows

The screenshot displays the SAP Data Services Spoon interface for a transformation named 'Transformation 1 (changed)'. The left sidebar shows a list of steps under the 'Flow' category, with 'Filter rows' selected. The main canvas shows a workflow starting with a 'Table input' step, followed by a 'hop' (indicated by a red arrow), and then a 'Filter rows' step. The 'Filter rows' step is highlighted with a green checkmark. Below the canvas, the 'Execution Results' section is visible, showing a table with execution metrics.

**Execution Results**

#	Stepname	Copynr	Read	Written	Input	Output	Update
1	Table input	0	0	252	252	0	

# Filter rows

The screenshot displays the SAP Spoon interface for 'Transformation 1 (changed)'. The left sidebar shows the 'Steps' palette with 'Filter rows' selected. The main canvas shows a 'Table input' step connected to a 'Filter rows' step. A context menu is open over the 'Filter rows' step, listing various actions. The 'Execution Results' panel at the bottom shows the execution history.

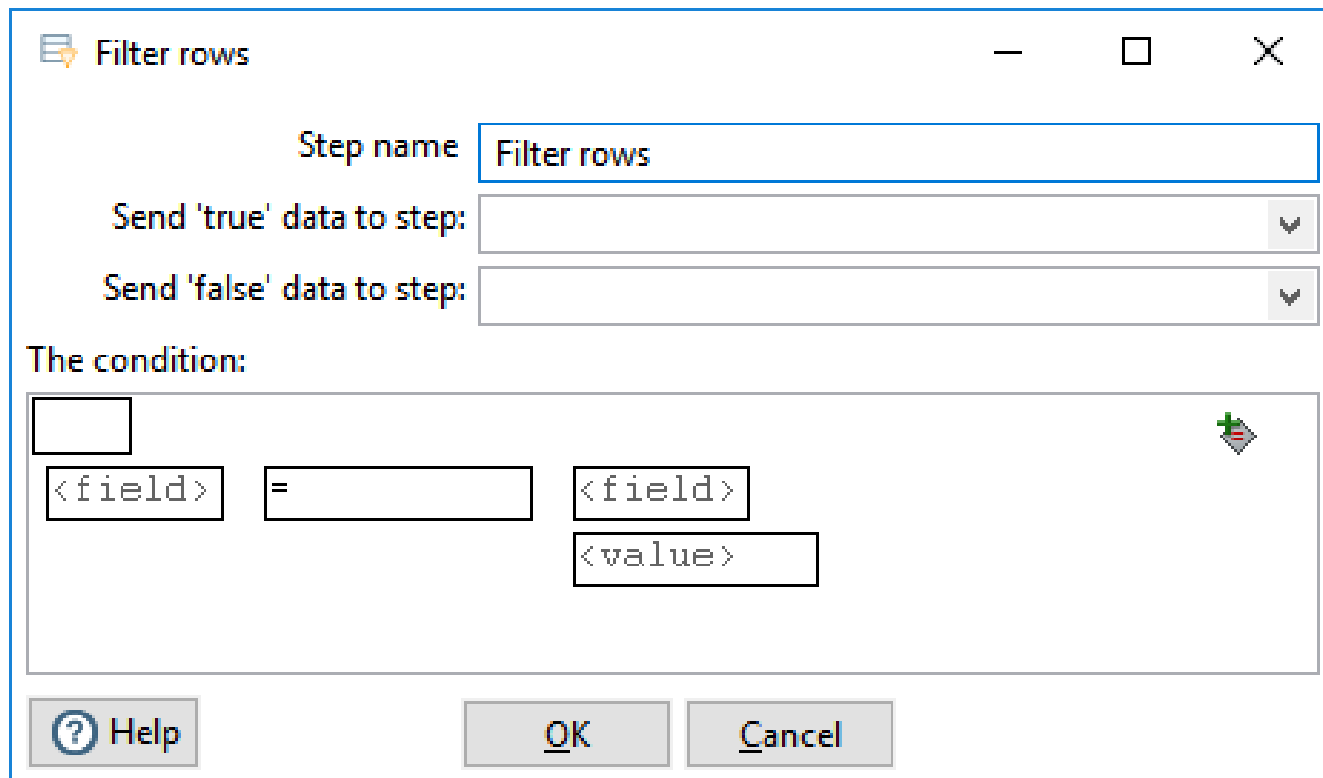
**Steps Palette:**

- 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

**Execution Results:**

#	Stepname	Copynr	R
1	Table input	0	

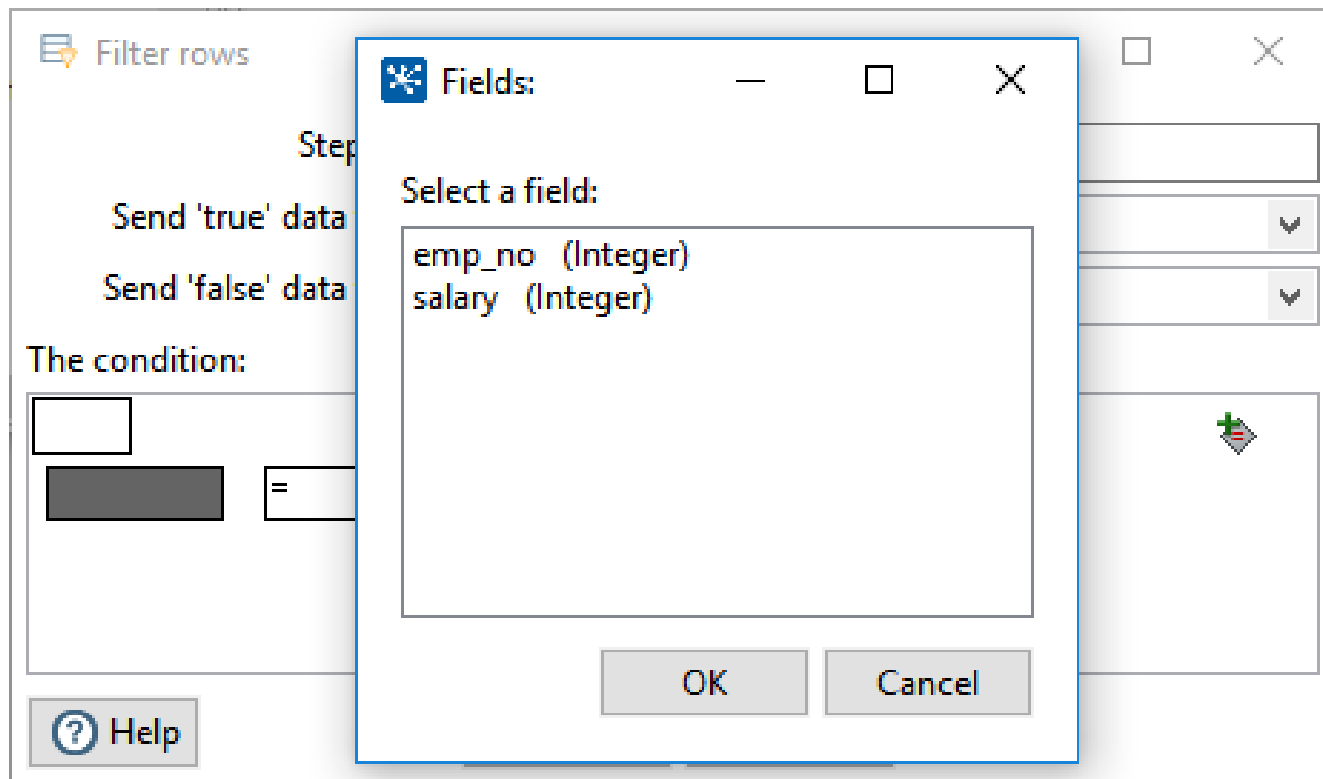
# Filter rows



The screenshot shows a 'Filter rows' dialog box with the following elements:

- Title bar:** 'Filter rows' with standard window controls (minimize, maximize, close).
- Step name:** A text field containing 'Filter rows'.
- Send 'true' data to step:** A dropdown menu.
- Send 'false' data to step:** A dropdown menu.
- The condition:** A large text area containing a logical expression: `<field> = <field>` and `<value>`. A small icon with a green plus sign and a red minus sign is visible in the top right corner of this area.
- Buttons:** 'Help' (with a question mark icon), 'OK', and 'Cancel'.

# Filter rows



# Filter rows

**Filter rows**

Step name:

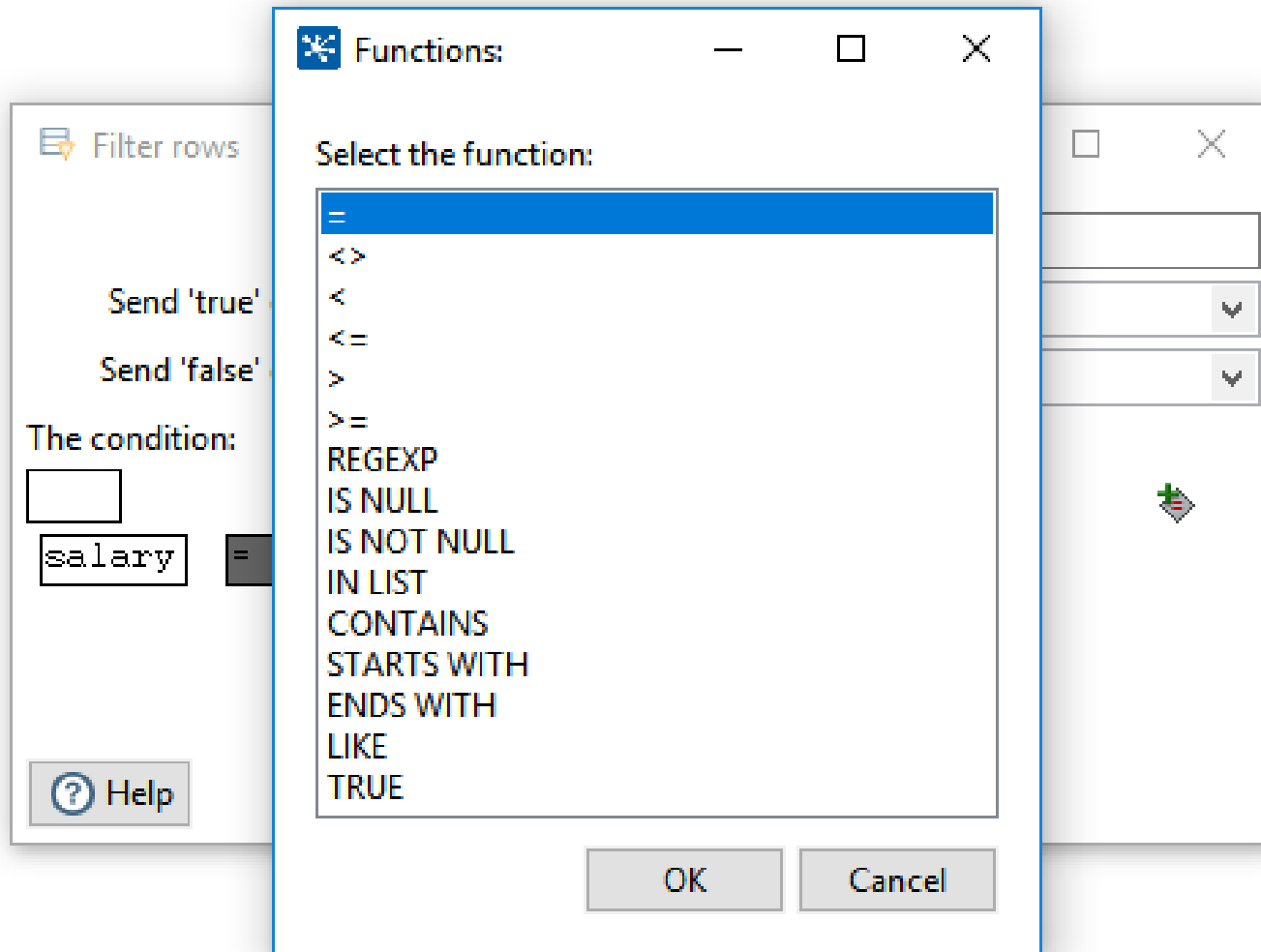
Send 'true' data to step:

Send 'false' data to step:

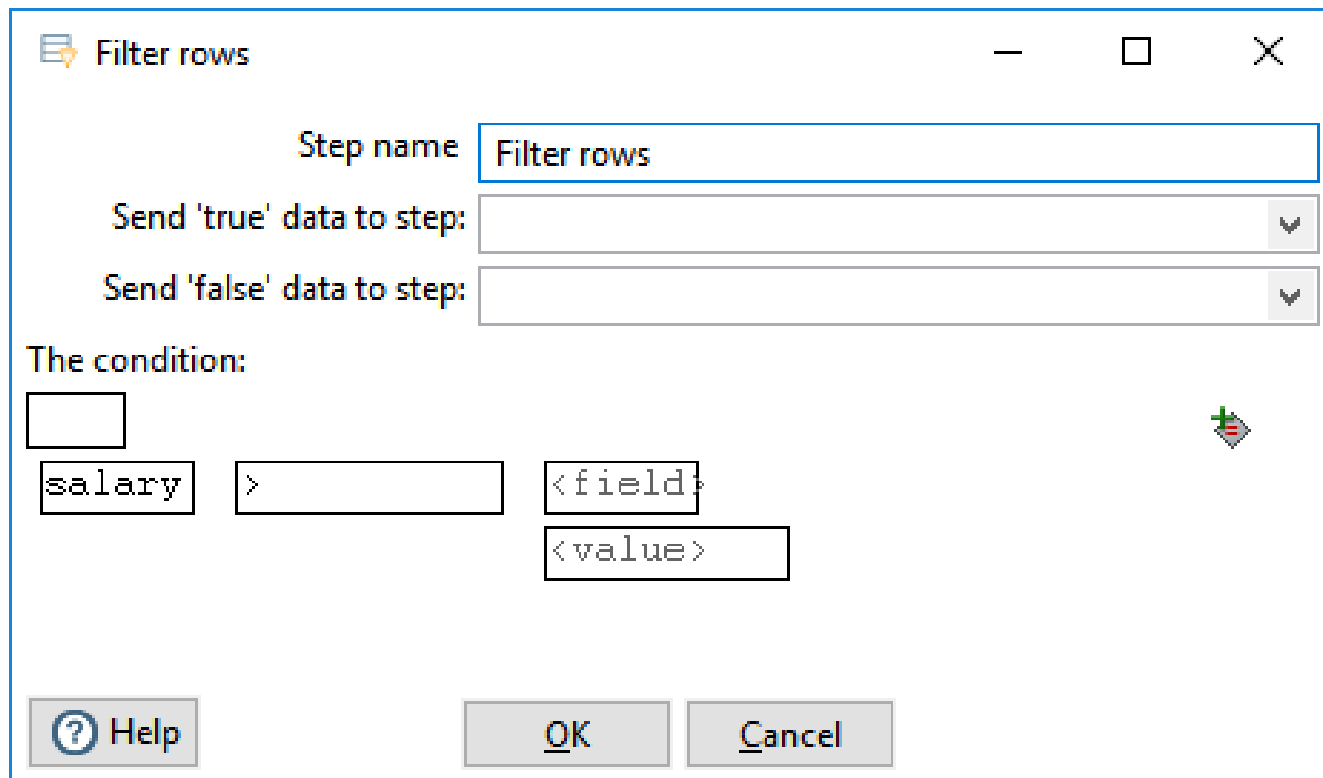
The condition:

=

# Filter rows



# Filter rows



The screenshot shows a 'Filter rows' dialog box with the following fields and controls:

- Step name:** A text field containing 'Filter rows'.
- Send 'true' data to step:** A dropdown menu.
- Send 'false' data to step:** A dropdown menu.
- The condition:** A section containing a checkbox, a text field with 'salary', a comparison operator '>', a field with '<field>', and a field with '<value>'. A small icon with a green plus sign and a red minus sign is located to the right of the condition fields.
- Buttons:** 'Help' (with a question mark icon), 'OK', and 'Cancel'.



# Filter rows

The image shows a 'Filter rows' dialog box with the following fields and options:

- Step name:** Filter rows
- Send 'true' data to step:** (empty field)
- Send 'false' data to step:** (empty field)
- The condition:**
  - ☐ salary >

Buttons at the bottom: ? Help, OK, Cancel.

The 'Enter value' sub-dialog is open, showing:

- Type:** Integer
- Value:** 80000
- Conversion format:** ####0;-####0
- Length:** -1
- Precision:** 0

Buttons at the bottom: OK, Test, Cancel.

# Filter rows

Filter rows

Step name: Filter rows

Send 'true' data to step:

Send 'false' data to step:

The condition:

(Integer)

Help OK Cancel

# Filter rows

The screenshot shows the SAP Data Services Spoon interface. The main workspace displays a transformation named 'Transformation 1' with two steps: 'Table input' and 'Filter rows'. The 'Filter rows' step is highlighted with a green checkmark. The left sidebar shows the 'Steps' list, with 'Filter rows' selected. The bottom right panel shows the 'Execution Results' table.

**Execution Results**

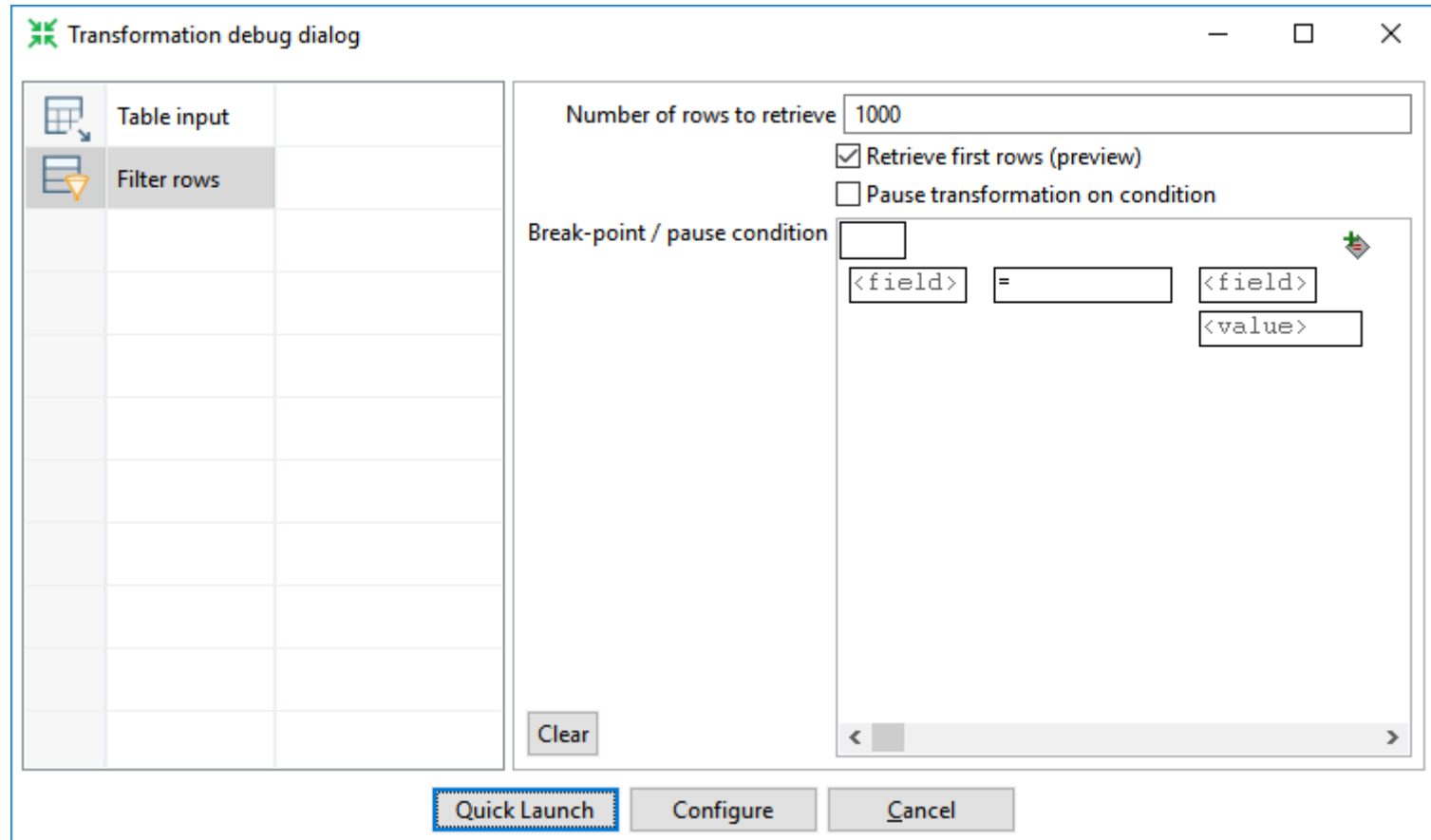
#	Stepname	Copynr	Read	Written	Input	Output	Update
1	Table input	0	0	252	252	0	

# Filter rows

The screenshot displays a data integration tool interface. On the left, a 'Flow' panel lists various steps, with 'Filter rows' highlighted. The main workspace shows a flow diagram with a 'Table input' step connected to a 'Filter' step. A context menu is open over the 'Filter' step, listing actions such as 'New Hop...', 'Edit...', 'Description...', 'Data Movement', 'Change Number of Copies to Start...', 'Copy', 'Duplicate', 'Delete', 'Hide', 'Detach', 'Input Fields...', 'Output Fields...', 'Sniff Test During Execution', 'Check Selected Step(s)', 'Error Handling...', 'Preview...', and 'Align / Distribute'. The 'Preview...' option is currently selected. Below the flow diagram, the 'Execution Results' section is visible, showing a table with the following data:

#	Stepname	Copynr
1	Table input	0

# Filter rows



# Filter rows

Examine preview data

Rows of step: Filter rows (82 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

Close

# Filter rows

The screenshot displays the Spoon - Transformation 1 (changed) window. The left sidebar shows the 'Steps' list with 'Filter rows' selected. The main canvas shows a workflow with 'Table input' and 'Filter rows' steps. The 'Execution Results' tab is active, showing the following table:

#	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

The screenshot shows the Spoon - Transformation 1 (changed) window. The left sidebar contains a tree view of the transformation steps, with 'Output' selected. The main canvas displays a data flow diagram with two steps: 'Table input' and 'Filter rows', connected by an arrow. Both steps have green checkmarks above them, indicating successful execution. Below the canvas, the 'Execution Results' tab is active, showing a table with execution metrics.

**Execution Results**

#	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

The screenshot shows the Spoon - Transformation 1 (changed) window. The left sidebar displays a list of output connectors under the 'Output' folder, including Automatic Documentation Out, Delete, Insert / Update, JSON Output, LDAP Output, Microsoft Access Output, Microsoft Excel Output, Microsoft Excel Writer, Pentaho Reporting Output, Properties Output, RSS Output, S3 File Output, and SQL File Output. The main canvas shows a data flow diagram with two steps: 'Table input' and 'Filter rows', connected by an arrow. Both steps have a green checkmark icon above them. The bottom panel displays the 'Execution Results' section, which includes tabs for Execution History, Logging, Step Metrics, and Performance Graph. The Execution History tab is active, showing a table with the following data:

#	Stepname	Copynr	Read	Written	Input	Output	Update
1	Table input	0	0	252	252	0	0
2	Filter rows	0	252	82	0	0	0

# Text file output

The screenshot shows the Spoon - Transformation 1 (changed) window. The left sidebar contains a tree view of steps, with 'Text file output' highlighted. The main canvas displays a data flow diagram with two steps: 'Table input' and 'Filter rows', connected by an arrow. The 'Execution Results' panel at the bottom shows a table with execution metrics for these steps.

**Execution Results**

#	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

The screenshot shows the Spoon - Transformation 1 (changed) window. The left sidebar contains a tree view of steps, with 'Text file output' selected. The main canvas displays a data flow diagram with three steps: 'Table input', 'Filter rows', and 'Text file output'. The 'Execution Results' panel at the bottom shows the execution history for these steps.

**Execution Results**

#	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

The screenshot displays the Spoon - Transformation 1 (changed) window. The left sidebar shows the 'Steps' list with 'Text file output' selected. The main canvas shows a data flow: 'Table input' (green checkmark) → 'Filter rows' (green checkmark) → 'Text file output' (green checkmark). A red 'hop' label is above the arrow between 'Filter rows' and 'Text file output'. A tooltip for the 'Filter rows' step shows a green checkmark for 'Result is TRUE' and a red X for 'Result is FALSE'. The 'Execution Results' table at the bottom shows the following data:

#	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

The screenshot shows the Spoon - Transformation 1 (changed) window. The left sidebar contains a tree view of the transformation steps, with 'Text file output' selected. The main canvas displays a data flow diagram with three steps: 'Table input', 'Filter rows', and 'Text file output'. The 'Execution Results' panel at the bottom shows the execution history for these steps.

**Execution Results**

#	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

The screenshot displays the Apache Kettle (Spoon) interface for a transformation named 'Transformation 1'. The left sidebar shows a tree view of steps, with 'Text file output' selected. The main canvas shows a flow from 'Table input' to 'Filter rows' to 'Text file output'. The 'Execution Results' tab is active, showing a table with the following data:

#	Stepname	Copynr	Read	Written	Inp
1	Table input	0	0	252	2
2	Filter rows	0	252	82	

The right-click context menu for the 'Text file output' step includes the following options:

- New Hop...
- Edit...
- Description...
- Data Movement
- Change Number of Copies to Start...
- Copy
- Duplicate
- Delete
- Hide
- Detach
- Input Fields...
- Output Fields...
- Sniff Test During Execution

# Text file output

The image shows a configuration window titled "Text file output". At the top, there is a "Step name" field containing the text "Text file output". Below this, there are three tabs: "File", "Content", and "Fields", with "File" being the active tab. The "File" tab contains several configuration options:

- Filename:** A text field with the value "C:\Temp\output" and a "Browse..." button to its right.
- Run this as command instead?** ☐
- Pass output to servlet** ☐
- Create Parent folder** ☐
- Do not create file at start** ☐
- Accept file name from field?** ☐
- File name field:** A dropdown menu currently showing an empty field.
- Extension:** A text field with the value "csv".
- Include stepnr in filename?** ☐
- Include partition nr in filename?** ☐
- Include date in filename?** ☐
- Include time in filename?** ☐
- Specify Date time format** ☐
- Date time format:** A dropdown menu currently showing an empty field.

At the bottom of the window, there are three buttons: a "Help" button (with a question mark icon), an "OK" button, and a "Cancel" button.

# Text file output

Text file output

Step name: Text file output

File Content Fields

Append ☐

Separator: ; Insert TAB

Enclosure:

Force the enclosure around fields? ☐

Disable the enclosure fix? ☐

Header ☒

Footer ☐

Format: CR+LF terminated (Windows, DOS)

Compression: None

Encoding: UTF-8

Right pad fields ☐

Fast data dump (no formatting) ☐

Split every ... rows: 0

Add Ending line of file:

Help OK Cancel



# Text file output

The screenshot shows a software window titled "Text file output". At the top right are standard window controls (minimize, maximize, close). Below the title bar, there's a label "Step name" followed by a text input field containing "Text file output". The main area has three tabs: "File", "Content", and "Fields". The "Fields" tab is selected, displaying a table with five columns: "#", "Name", "Type", "Format", "Length", and "Precision". The first row of the table contains the number "1" under the "#" column, while the other columns are empty. Below the table is a horizontal scrollbar. At the bottom of the dialog, there are four buttons: "Get Fields", "Minimal width", "OK", and "Cancel". A small "Help" button with a question mark icon is located at the bottom left corner.

# Text file output

Text file output

Step name: Text file output

File Content Fields

#	Name	Type	Format	Length	Precision
1	emp_no	Integer	#####0;-#####0	9	0
2	salary	Integer	#####0;-#####0	9	0

< >

Get Fields Minimal width

Get the fields as defined in previous steps.

Help OK Cancel

# Text file output

[illegible]

# Text file output

The screenshot shows the Spoon - Transformation 1 (changed) window. The left sidebar contains a tree view of the transformation steps, with 'Text file output' selected. The main canvas displays a flow diagram with three steps: 'Table input', 'Filter rows', and 'Text file output'. The 'Execution Results' panel at the bottom shows the execution history and metrics for these steps.

**Execution Results**

#	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

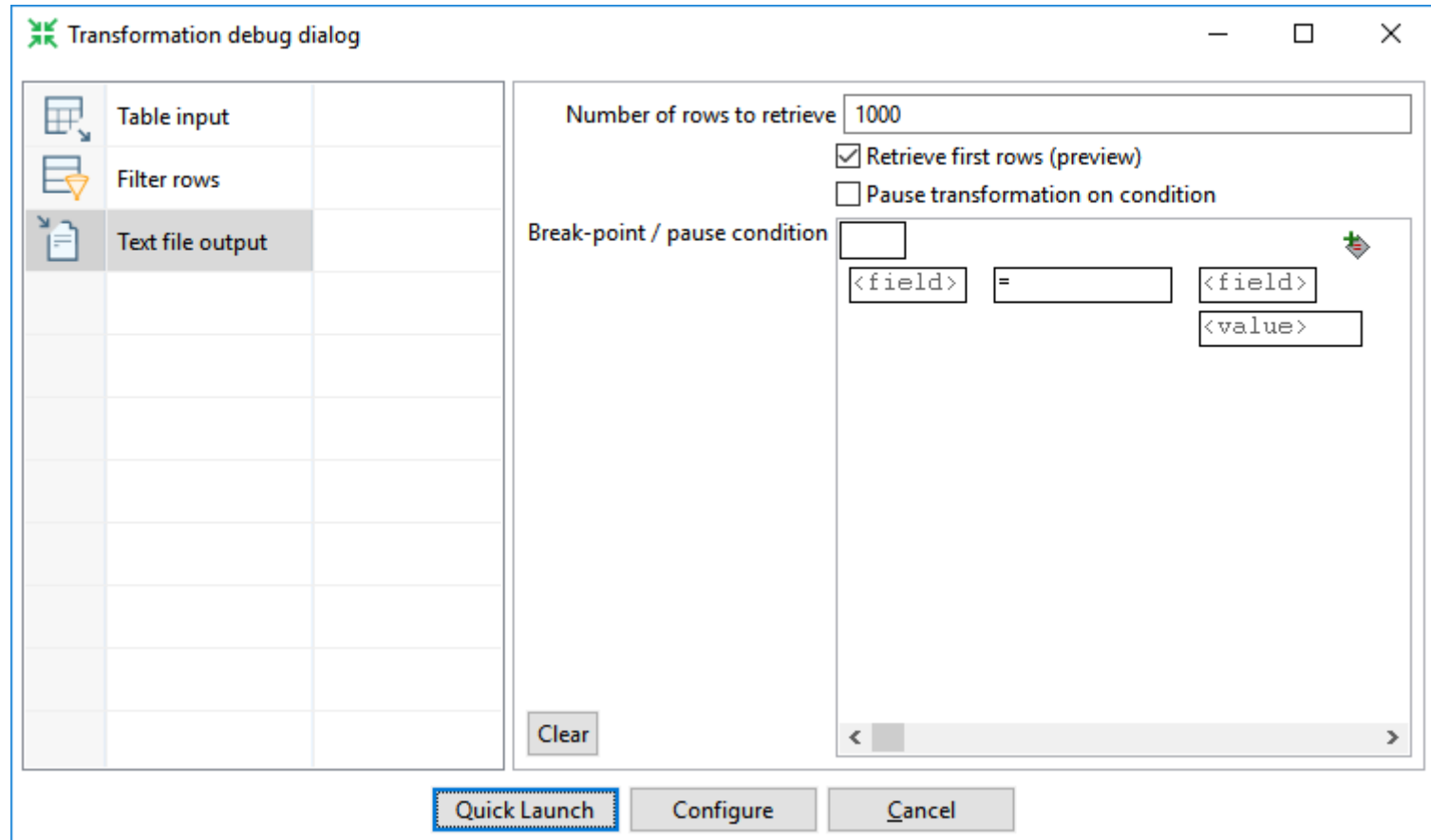
The screenshot displays the Pentaho Data Integration (Kettle) interface. On the left, a tree view shows the project structure, with 'Text file output' selected under the 'Transform' folder. The main workspace shows a workflow with three steps: 'Table input', 'Filter rows', and 'Text file output'. All steps have green checkmarks indicating successful execution. Below the workflow, the 'Execution Results' tab is active, showing a table with execution metrics.

#	Stepname	Copynr	Read	Written	Input
1	Table input	0	0	252	
2	Filter rows	0	252	82	

A context menu is open over the 'Text file output' step, listing various actions. The 'Preview...' option is highlighted in blue.

- New Hop...
- Edit...
- Description...
- Data Movement
- Change Number of Copies to Start...
- Copy
- Duplicate
- Delete
- Hide
- Detach
- Input Fields...
- Output Fields...
- Sniff Test During Execution
- Check Selected Step(s)
- Error Handling...
- Preview...**
- Align / Distribute

# Text file output



# Text file output

Examine preview data

Rows of step: Text file output (82 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

Close

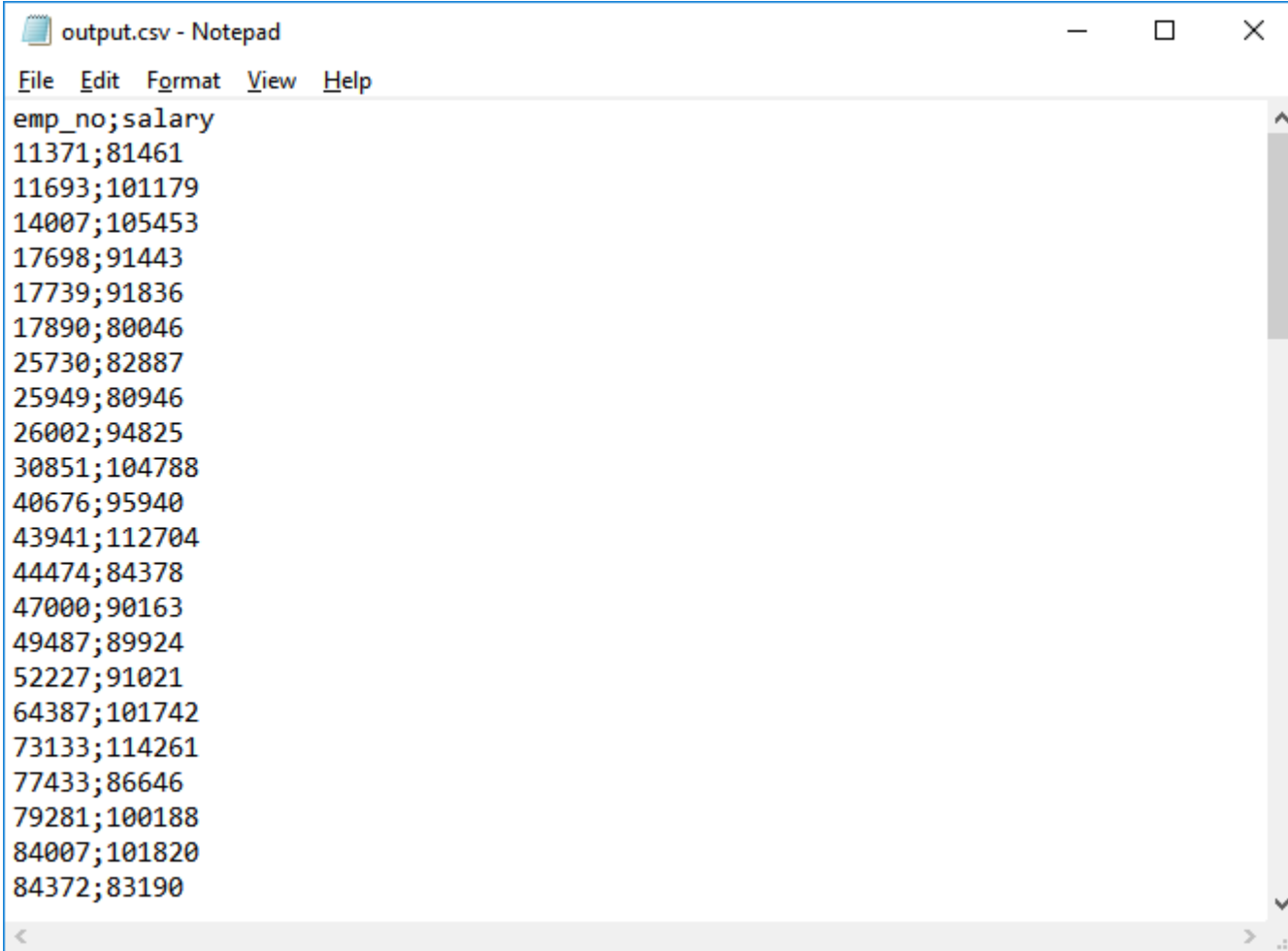
# Text file output

The screenshot displays the Spoon - Transformation 1 (changed) window. The left sidebar shows a tree view of components, with 'Text file output' selected under the 'Transform' folder. The main canvas shows a data flow diagram with three steps: 'Table input', 'Filter rows', and 'Text file output'. The 'Execution Results' tab is active, showing a table with the following data:

#	Stepname	Copynr	Read	Written	Input	Output	Upd
1	Table input	0	0	252	252	0	
2	Filter rows	0	252	82	0	0	
3	Text file output	0	82	82	0	83	



# Output



A screenshot of a Notepad window titled "output.csv - Notepad". The window displays a list of employee data in a CSV format, with each line representing an employee's ID and salary separated by a semicolon. The data is as follows:

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

output.csv - LibreOffice Calc

File Edit View Insert Format Sheet Data Tools Window Help

Liberation Sans 10

C1

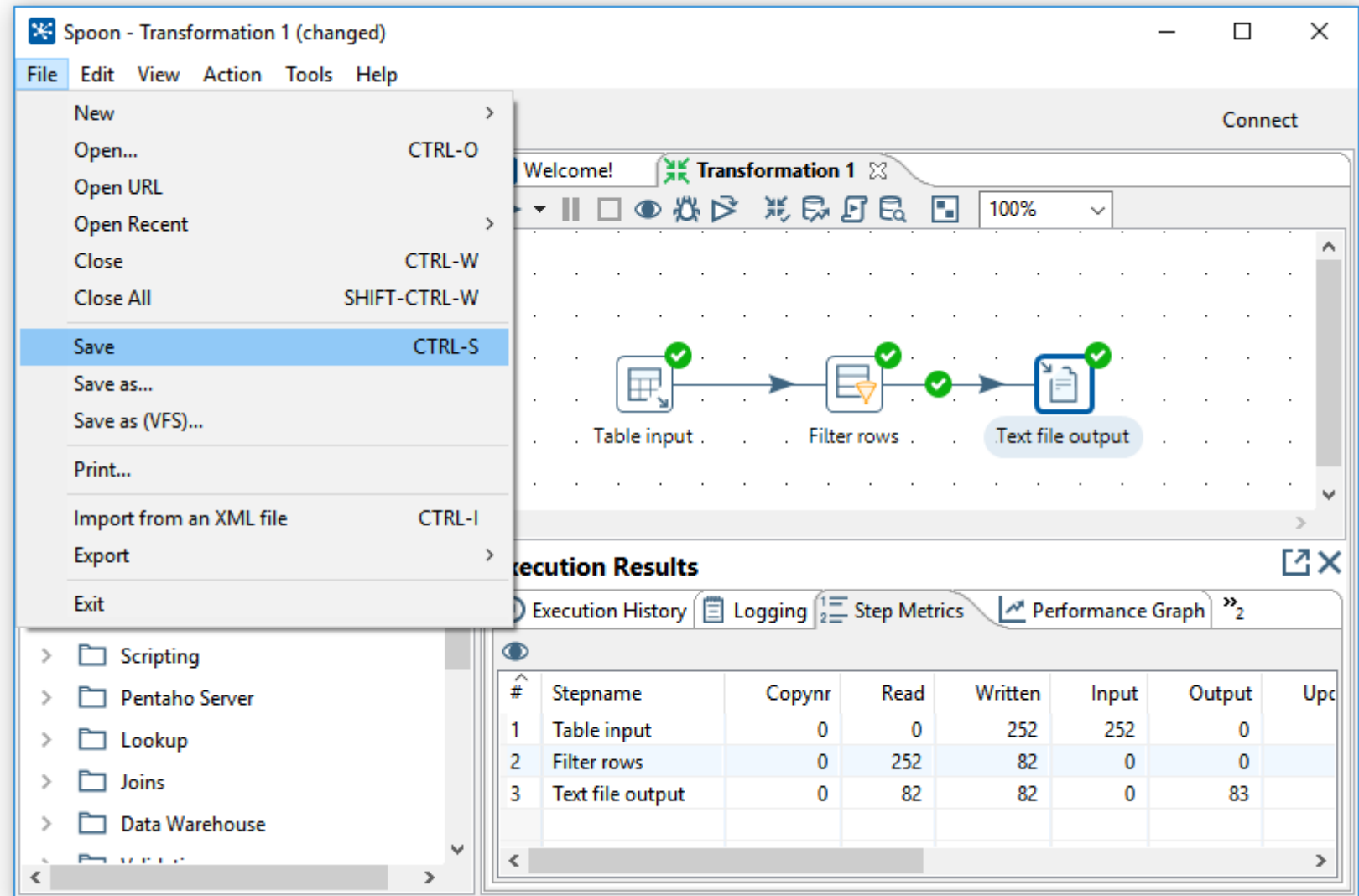
	A	B	C	D	E	F	G	H	I
1	emp_no	salary							
2	11371	81461							
3	11693	101179							
4	14007	105453							
5	17698	91443							
6	17739	91836							
7	17890	80046							
8	25730	82887							
9	25949	80946							
10	26002	94825							
11	30851	104788							
12	40676	95940							
13	43941	112704							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

output

Find Find All ☐ Formatted Display ☐ Match Case

Sheet 1 of 1 Default Average: ; Sum: 0 100%

# Save transformation



# Run transformation

The screenshot shows the Apache Kettle (Spoon) interface for running a transformation. The 'Run' button is circled in red. The transformation flow is as follows:

```
graph LR; A[Table input] --> B[Filter rows]; B --> C[Text file output];
```

**Execution Results**

#	Stepname	Copynr	Read	Written	Input	Output	Upc
1	Table input	0	0	252	252	0	
2	Filter rows	0	252	82	0	0	
3	Text file output	0	82	82	0	83	