Practical No. 04

Title: - Implementation of ETL transformation with Pentaho
Aim: - ETL Transformation with Pentaho.

Lab Objectives: -

Students will understand following concepts:

I. Copy data from Source (Table/Excel/ Oracle) and store it to Target (Table/Excel/ Oracle)

II. Adding sequence, Adding Calculator, Concatenation of two fields, Splitting of two fields

III. String Operations, Sorting data, Implement the merge join

transformation on tables. Description: -

Pentaho Data Integration(PDI)

It is a business Intelligence system

(BI) Also known as KETTLE.

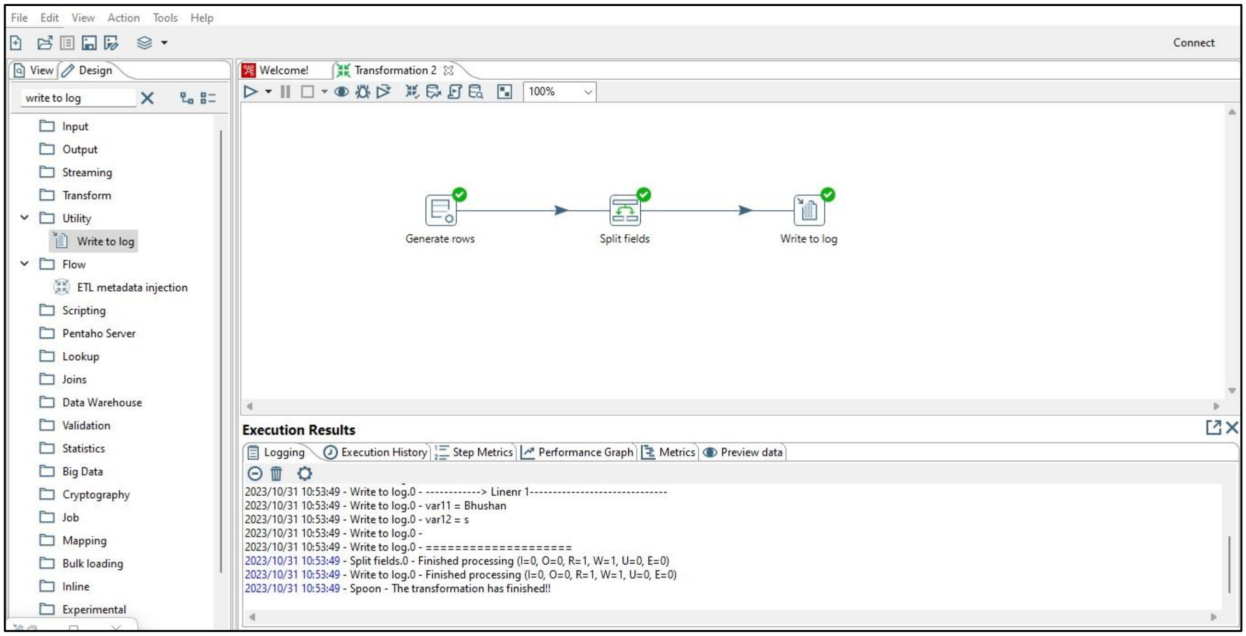
Pentaho, a subsidiary of Hitachi Vantara, is free and open- source platform for data integration and analytics. The software comes in a free community edition and a subscription-based enterprise edition.

Pentaho Data Integration (PDI) is one of the most powerful tool for building ETL processes. Founded in 2004 and Stable released on 9.1.0.0-324 / September 7, 2020.

Available for Windows, Linux, MAC OSX.

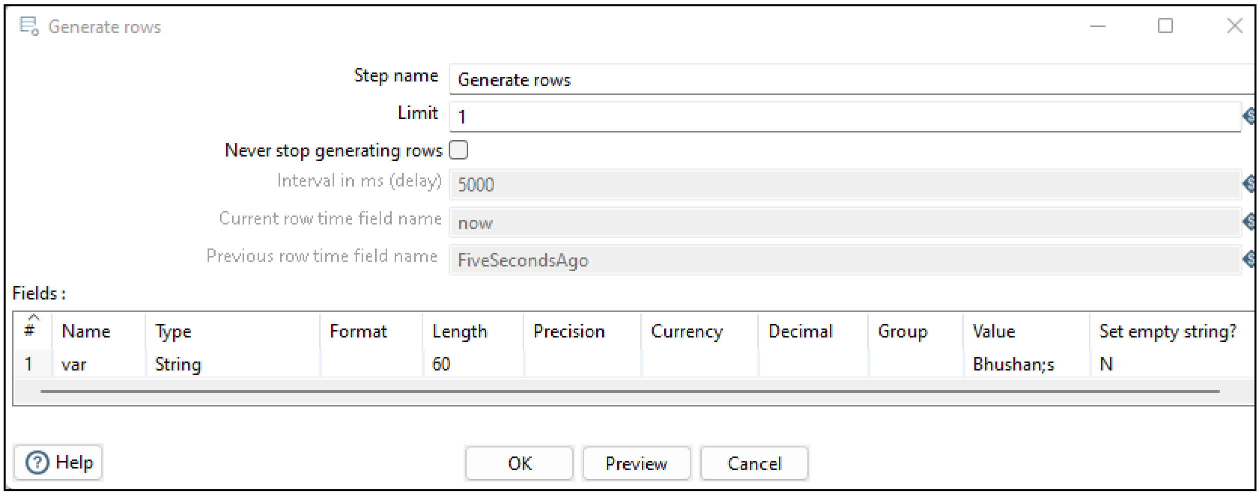
PDI is a java-based tool (Uses the Apache Java application server)

SPLit Field :-



Execution Results

```
2023/10/31 10:53:49 - Write to log.0 - ------------> Linenr 1------------------------------
2023/10/31 10:53:49 - Write to log.0 - var11 = Bhushan
2023/10/31 10:53:49 - Write to log.0 - var12 = s
2023/10/31 10:53:49 - Write to log.0 -
2023/10/31 10:53:49 - Write to log.0 - ====================
2023/10/31 10:53:49 - Split fields.0 - Finished processing (I=0, O=0, R=1, W=1, U=0, E=0)
2023/10/31 10:53:49 - Write to log.0 - Finished processing (I=0, O=0, R=1, W=1, U=0, E=0)
2023/10/31 10:53:49 - Spoon - The transformation has finished!!
```



Generate rows

Step name: Generate rows

Limit: 1

Never stop generating rows ☐

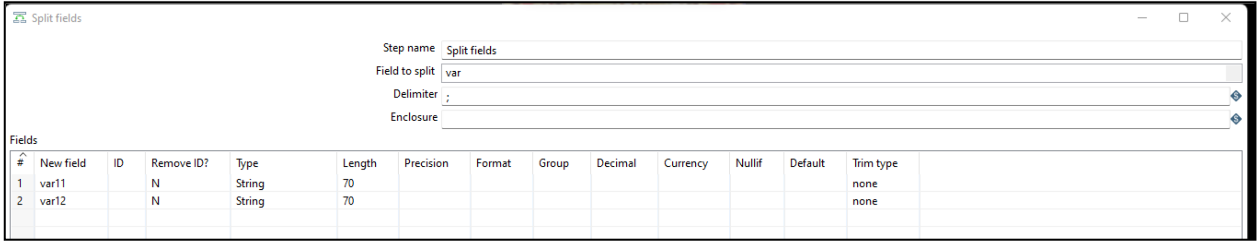
Interval in ms (delay): 5000

Current row time field name: now

Previous row time field name: FiveSecondsAgo

Fields :

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Value	Set empty string?
1	var	String		60					Bhushan;s	N



Split fields

Step name: Split fields

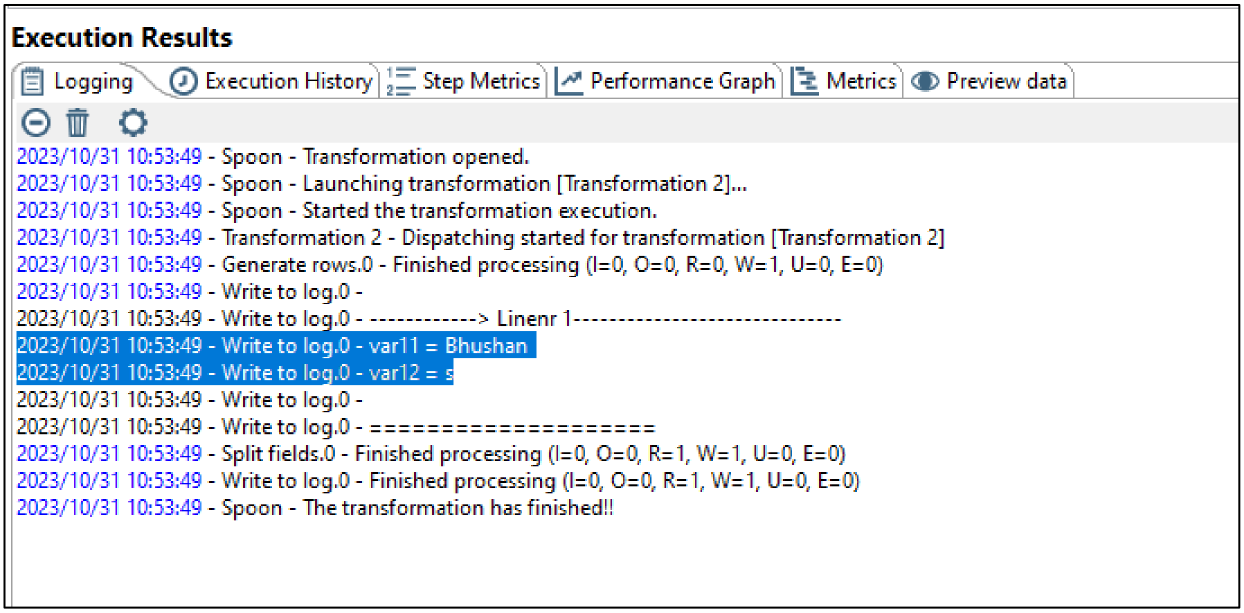
Field to split: var

Delimiter: ;

Enclosure:

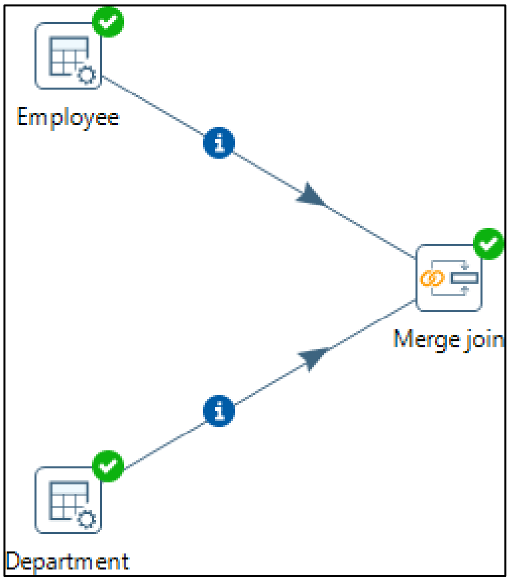
Fields

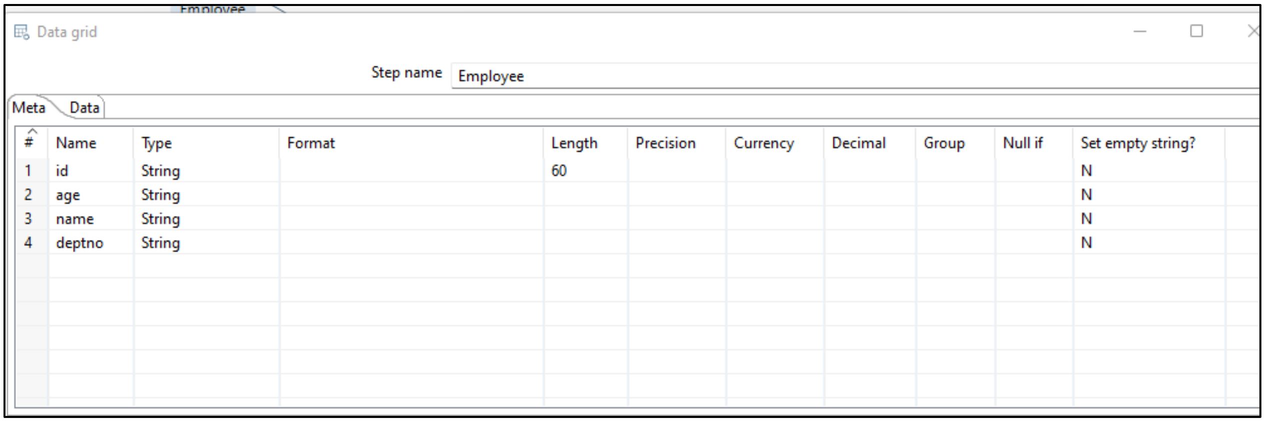
#	New field	ID	Remove ID?	Type	Length	Precision	Format	Group	Decimal	Currency	Nullif	Default	Trim type
1	var11		N	String	70								none
2	var12		N	String	70								none



```
2023/10/31 10:53:49 - Spoon - Transformation opened.
2023/10/31 10:53:49 - Spoon - Launching transformation [Transformation 2]...
2023/10/31 10:53:49 - Spoon - Started the transformation execution.
2023/10/31 10:53:49 - Transformation 2 - Dispatching started for transformation [Transformation 2]
2023/10/31 10:53:49 - Generate rows.0 - Finished processing (I=0, O=0, R=0, W=1, U=0, E=0)
2023/10/31 10:53:49 - Write to log.0 -
2023/10/31 10:53:49 - Write to log.0 - ------------> Linenr 1------------------------------
2023/10/31 10:53:49 - Write to log.0 - var11 = Bhushan
2023/10/31 10:53:49 - Write to log.0 - var12 = s
2023/10/31 10:53:49 - Write to log.0 -
2023/10/31 10:53:49 - Write to log.0 - ====================
2023/10/31 10:53:49 - Split fields.0 - Finished processing (I=0, O=0, R=1, W=1, U=0, E=0)
2023/10/31 10:53:49 - Write to log.0 - Finished processing (I=0, O=0, R=1, W=1, U=0, E=0)
2023/10/31 10:53:49 - Spoon - The transformation has finished!!
```

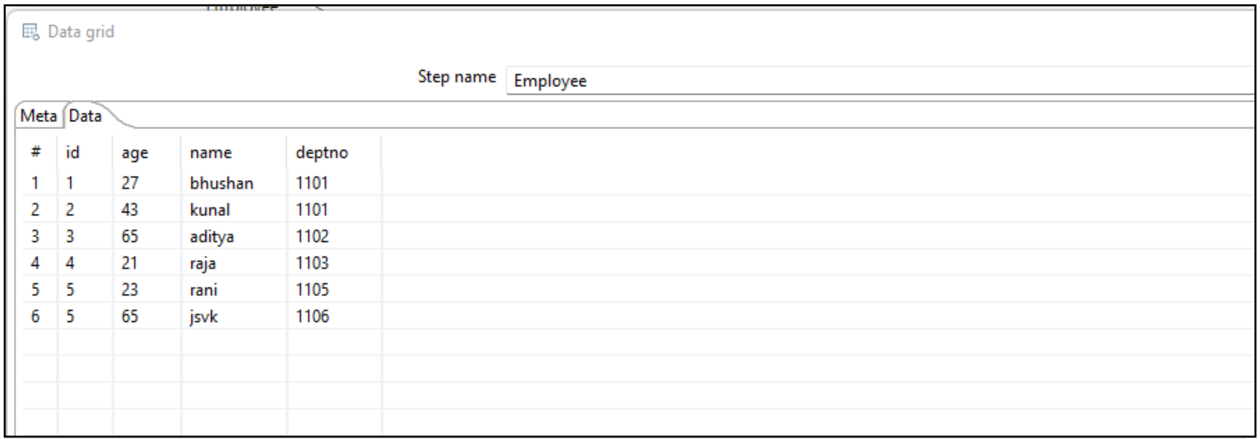
Merge Joint





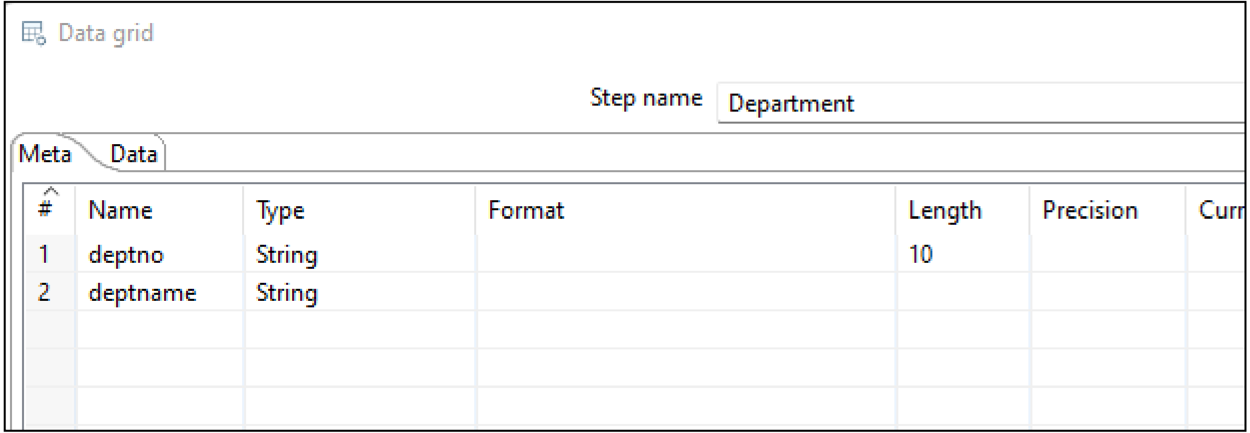
Step name Employee

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Null if	Set empty string?
1	id	String		60						N
2	age	String								N
3	name	String								N
4	deptno	String								N



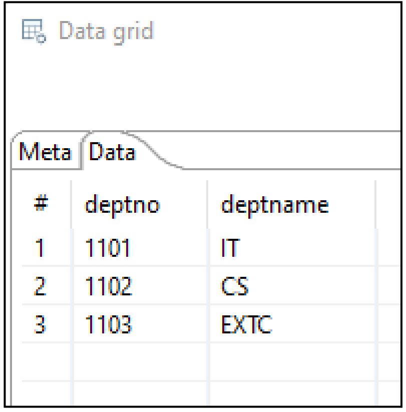
Step name Employee

#	id	age	name	deptno
1	1	27	bhushan	1101
2	2	43	kunal	1101
3	3	65	aditya	1102
4	4	21	raja	1103
5	5	23	rani	1105
6	5	65	jsvk	1106

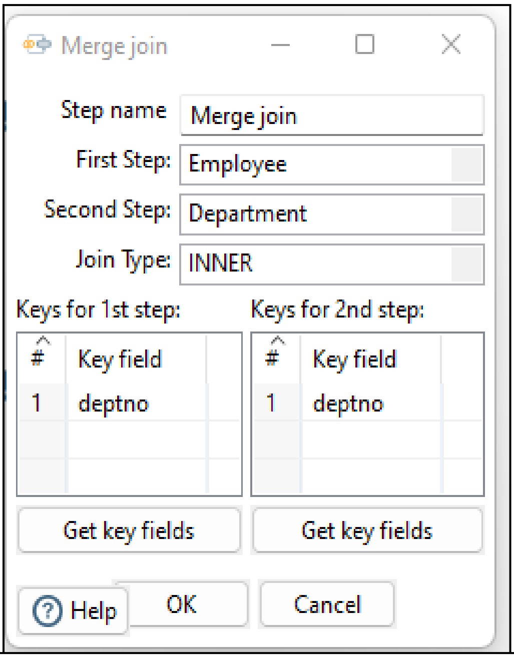


Step name Department

#	Name	Type	Format	Length	Precision	Curr
1	deptno	String		10		
2	deptname	String				



#	deptno	deptname
1	1101	IT
2	1102	CS
3	1103	EXTC



Step name: Merge join

First Step: Employee

Second Step: Department

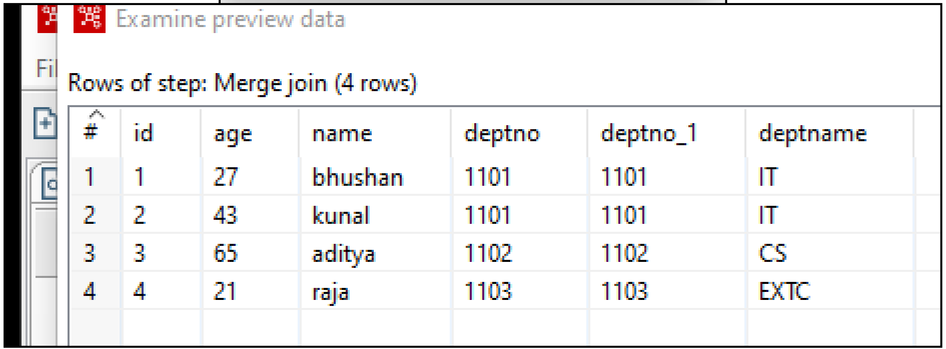
Join Type: INNER

Keys for 1st step:

#	Key field
1	deptno

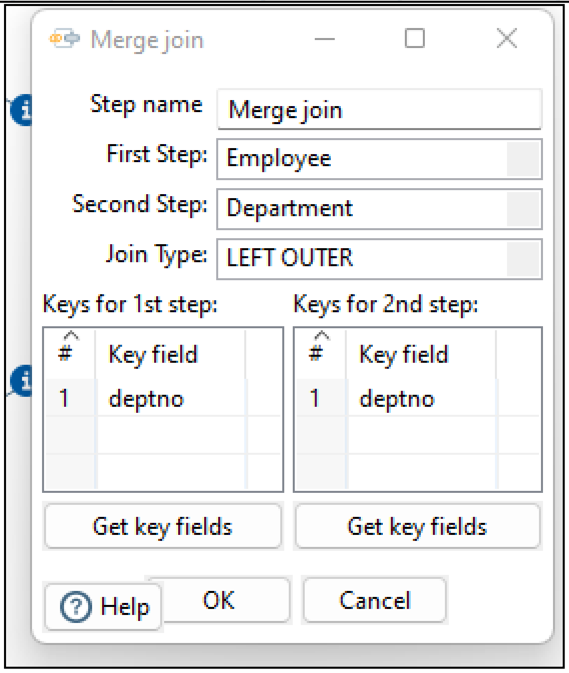
Keys for 2nd step:

#	Key field
1	deptno



Rows of step: Merge join (4 rows)

#	id	age	name	deptno	deptno_1	deptname
1	1	27	bhushan	1101	1101	IT
2	2	43	kunal	1101	1101	IT
3	3	65	aditya	1102	1102	CS
4	4	21	raja	1103	1103	EXTC



Step name: Merge join

First Step: Employee

Second Step: Department

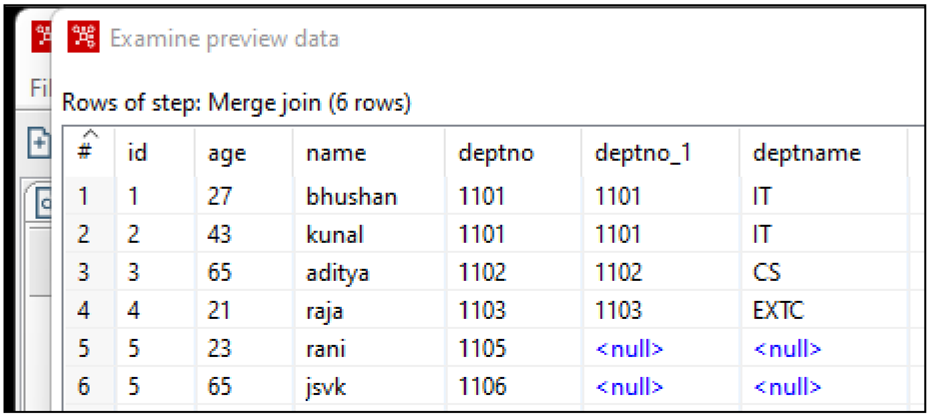
Join Type: LEFT OUTER

Keys for 1st step:

#	Key field
1	deptno

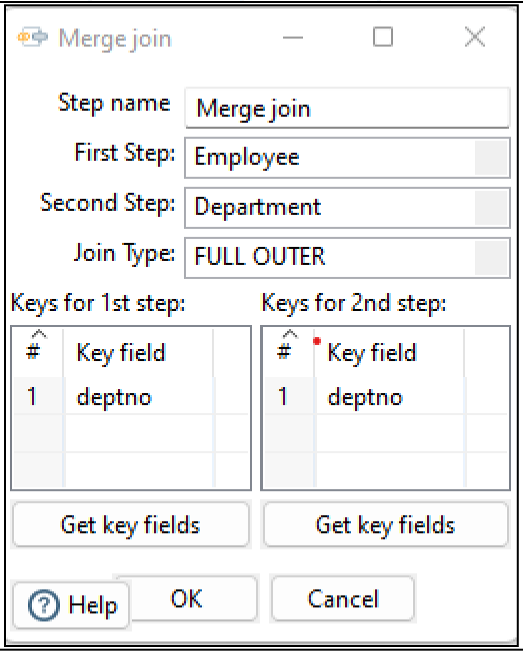
Keys for 2nd step:

#	Key field
1	deptno


Examine preview data

Rows of step: Merge join (6 rows)

#	id	age	name	deptno	deptno_1	deptname
1	1	27	bhushan	1101	1101	IT
2	2	43	kunal	1101	1101	IT
3	3	65	aditya	1102	1102	CS
4	4	21	raja	1103	1103	EXTC
5	5	23	rani	1105	<null>	<null>
6	5	65	jsvk	1106	<null>	<null>


Merge join

Step name: Merge join

First Step: Employee

Second Step: Department

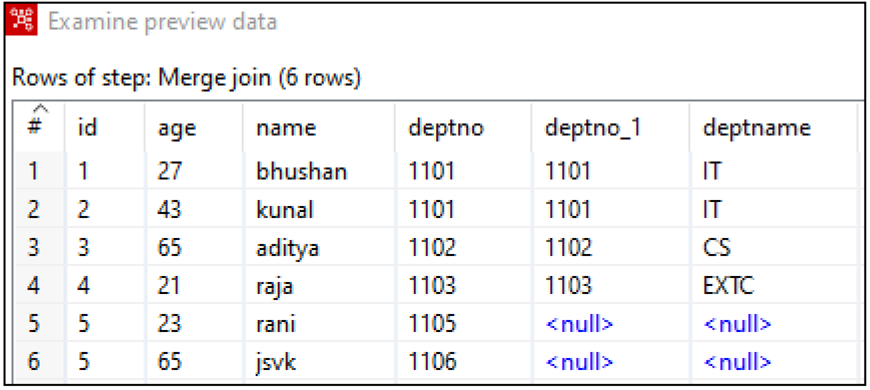
Join Type: FULL OUTER

Keys for 1st step:

#	Key field
1	deptno

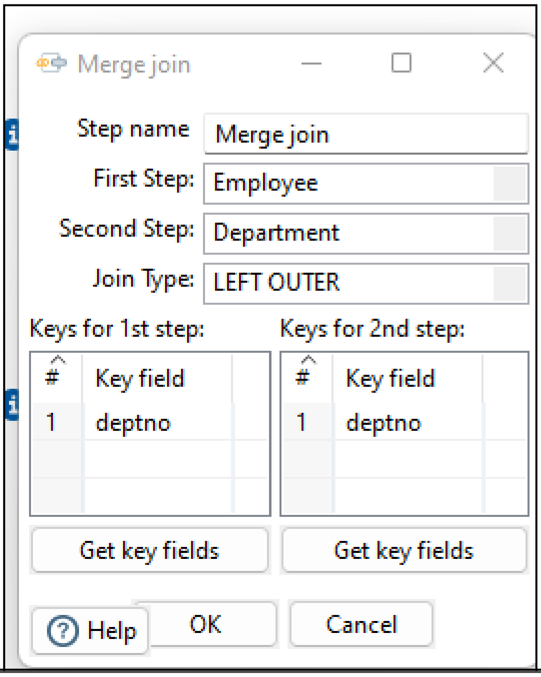
Keys for 2nd step:

#	Key field
1	deptno


Examine preview data

Rows of step: Merge join (6 rows)

#	id	age	name	deptno	deptno_1	deptname
1	1	27	bhushan	1101	1101	IT
2	2	43	kunal	1101	1101	IT
3	3	65	aditya	1102	1102	CS
4	4	21	raja	1103	1103	EXTC
5	5	23	rani	1105	<null>	<null>
6	5	65	jsvk	1106	<null>	<null>



Step name: Merge join

First Step: Employee

Second Step: Department

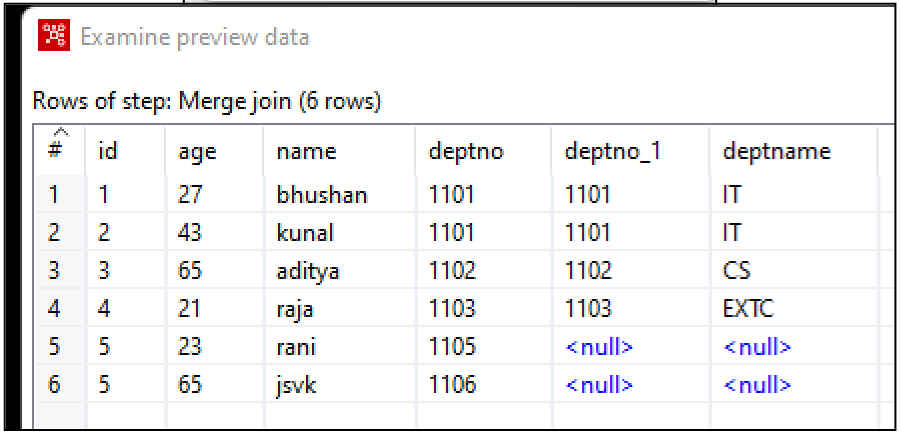
Join Type: LEFT OUTER

Keys for 1st step:

#	Key field
1	deptno

Keys for 2nd step:

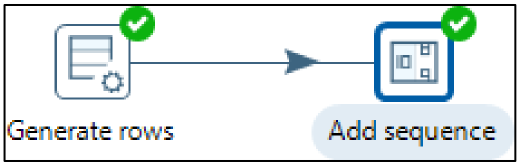
#	Key field
1	deptno

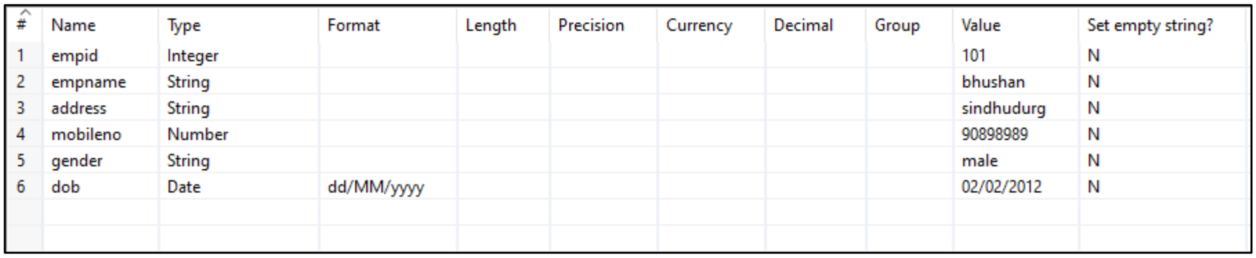


Rows of step: Merge join (6 rows)

#	id	age	name	deptno	deptno_1	deptname
1	1	27	bhushan	1101	1101	IT
2	2	43	kunal	1101	1101	IT
3	3	65	aditya	1102	1102	CS
4	4	21	raja	1103	1103	EXTC
5	5	23	rani	1105	<null>	<null>
6	5	65	jsvk	1106	<null>	<null>

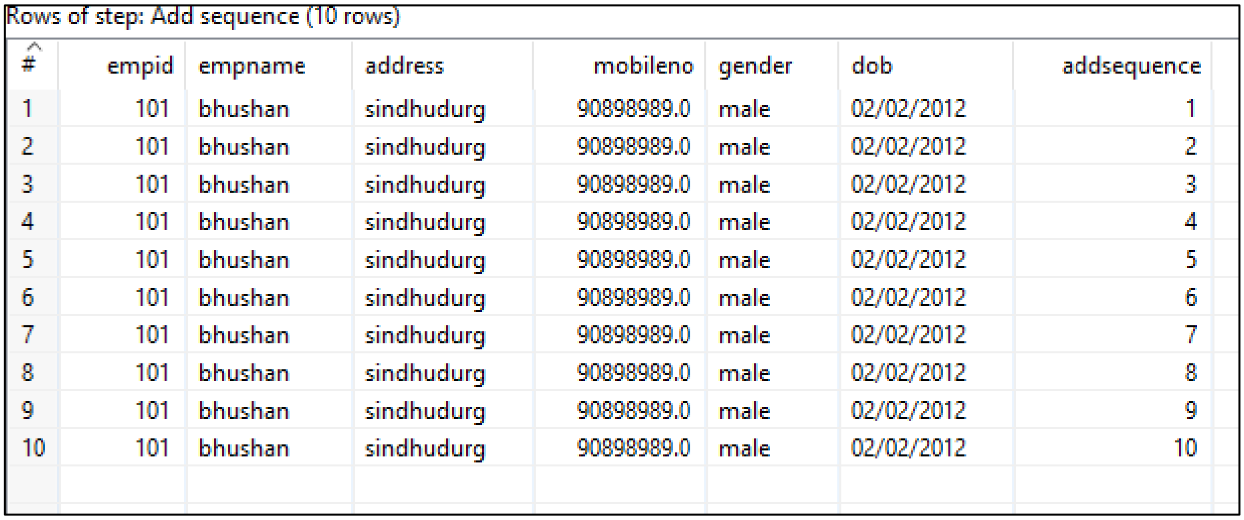
Adding Sequence





#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Value	Set empty string?
1	empid	Integer							101	N
2	empname	String							bhushan	N
3	address	String							sindhudurg	N
4	mobileno	Number							90898989	N
5	gender	String							male	N
6	dob	Date	dd/MM/yyyy						02/02/2012	N

Result preview

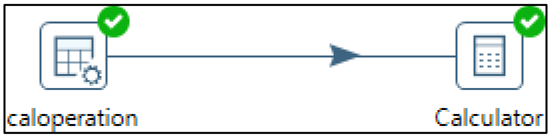


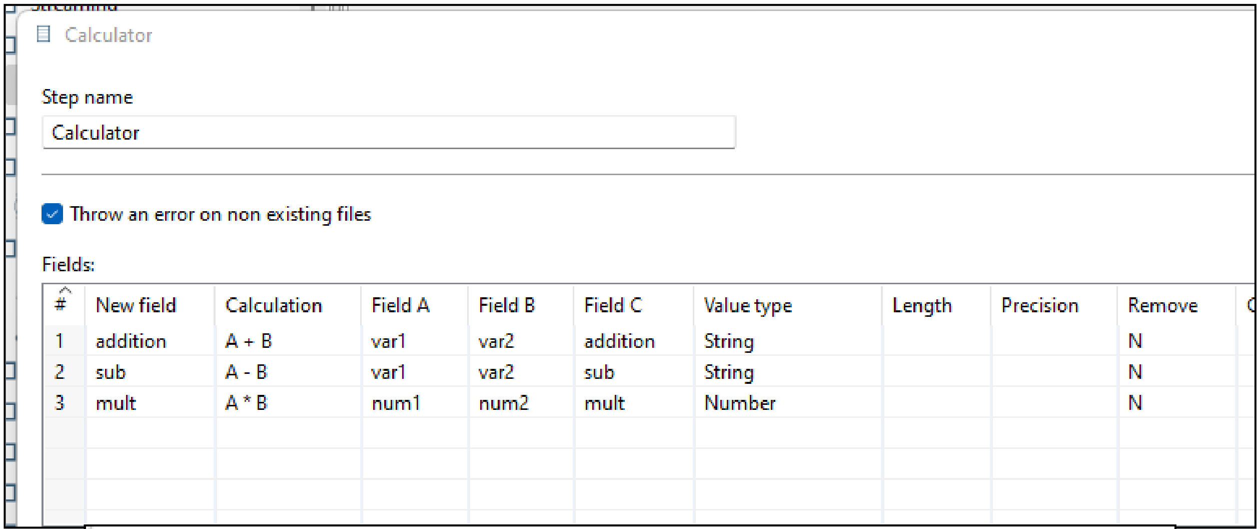
Rows of step: Add sequence (10 rows)

#	empid	empname	address	mobileno	gender	dob	addsequence
1	101	bhushan	sindhudurg	90898989.0	male	02/02/2012	1
2	101	bhushan	sindhudurg	90898989.0	male	02/02/2012	2
3	101	bhushan	sindhudurg	90898989.0	male	02/02/2012	3
4	101	bhushan	sindhudurg	90898989.0	male	02/02/2012	4
5	101	bhushan	sindhudurg	90898989.0	male	02/02/2012	5
6	101	bhushan	sindhudurg	90898989.0	male	02/02/2012	6
7	101	bhushan	sindhudurg	90898989.0	male	02/02/2012	7
8	101	bhushan	sindhudurg	90898989.0	male	02/02/2012	8
9	101	bhushan	sindhudurg	90898989.0	male	02/02/2012	9
10	101	bhushan	sindhudurg	90898989.0	male	02/02/2012	10

On data grid

Calc Operation



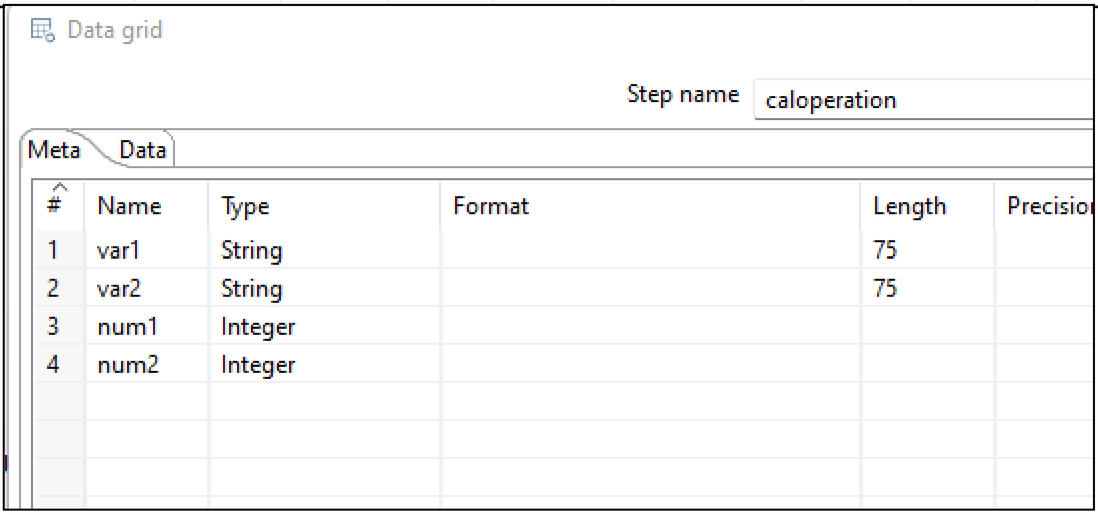


Step name: Calculator

☑ Throw an error on non existing files

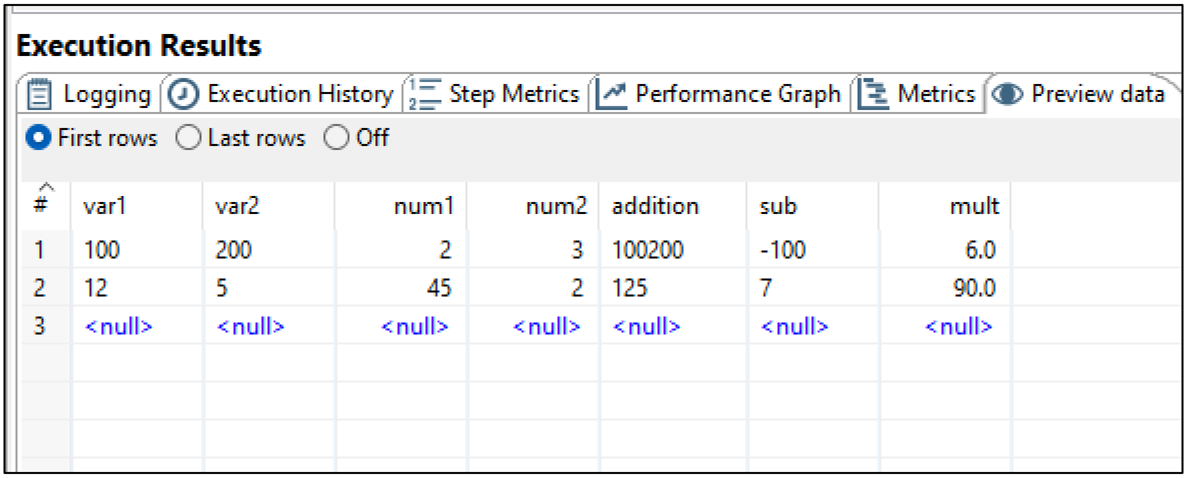
Fields:

#	New field	Calculation	Field A	Field B	Field C	Value type	Length	Precision	Remove
1	addition	A + B	var1	var2	addition	String			N
2	sub	A - B	var1	var2	sub	String			N
3	mult	A * B	num1	num2	mult	Number			N



Step name: caloperation

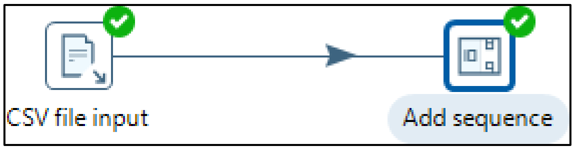
#	Name	Type	Format	Length	Precisio
1	var1	String		75	
2	var2	String		75	
3	num1	Integer			
4	num2	Integer			

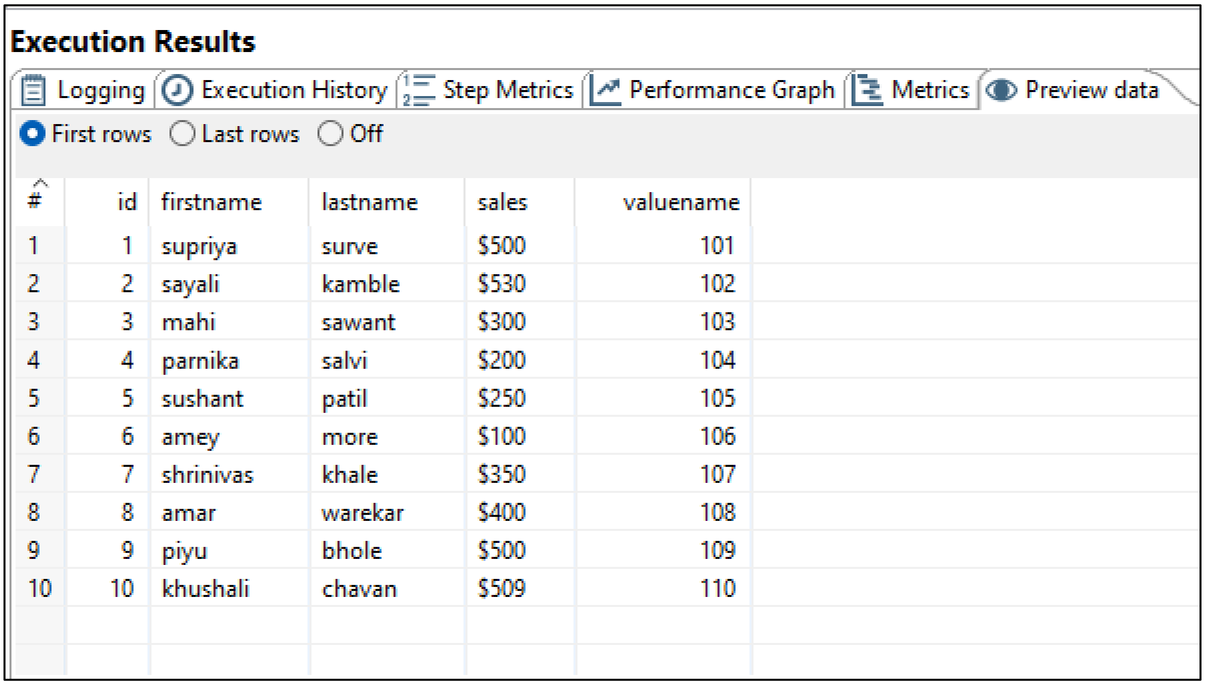


● First rows ○ Last rows ○ Off

#	var1	var2	num1	num2	addition	sub	mult
1	100	200	2	3	100200	-100	6.0
2	12	5	45	2	125	7	90.0
3	<null>	<null>	<null>	<null>	<null>	<null>	<null>

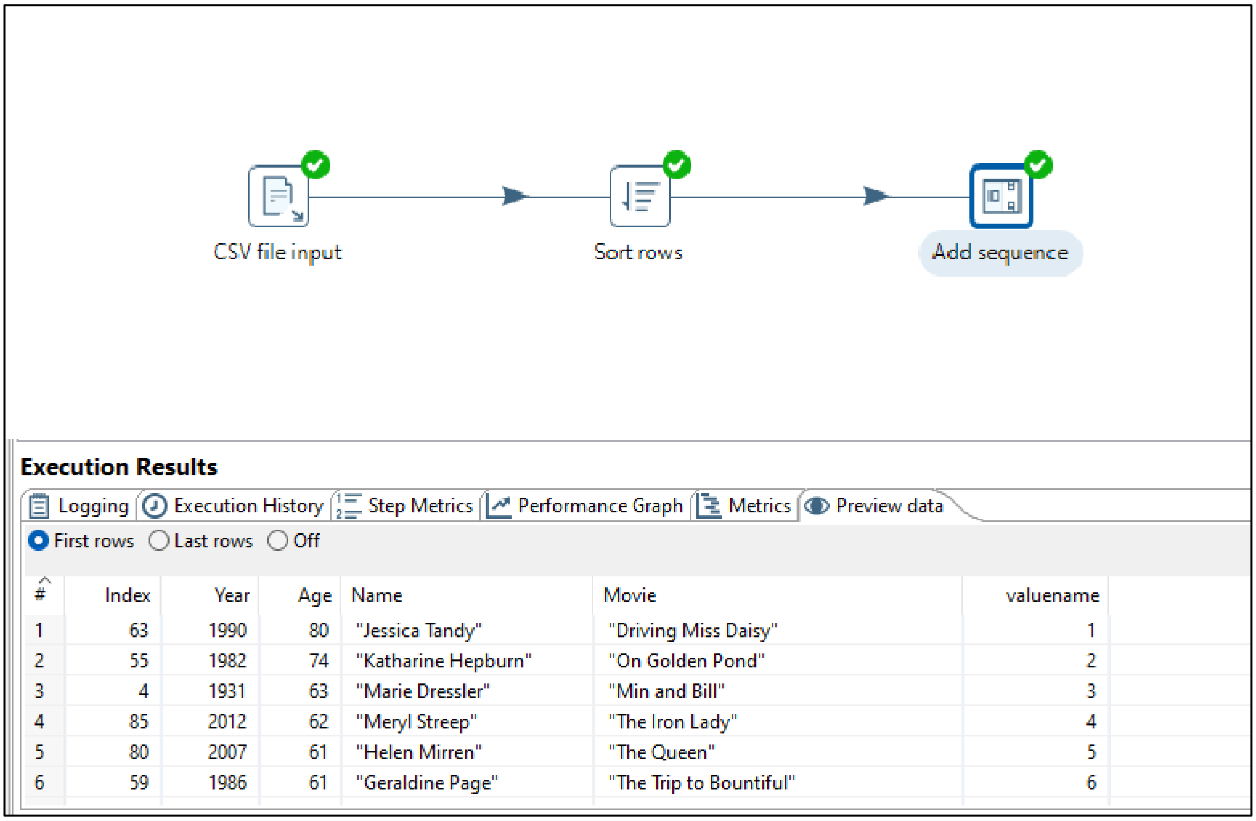
CSV INCLUDE





● First rows ○ Last rows ○ Off

#	id	firstname	lastname	sales	valuename
1	1	supriya	surve	$500	101
2	2	sayali	kamble	$530	102
3	3	mahi	sawant	$300	103
4	4	parnika	salvi	$200	104
5	5	sushant	patil	$250	105
6	6	amey	more	$100	106
7	7	shrinivas	khale	$350	107
8	8	amar	warekar	$400	108
9	9	piyu	bhole	$500	109
10	10	khushali	chavan	$509	110

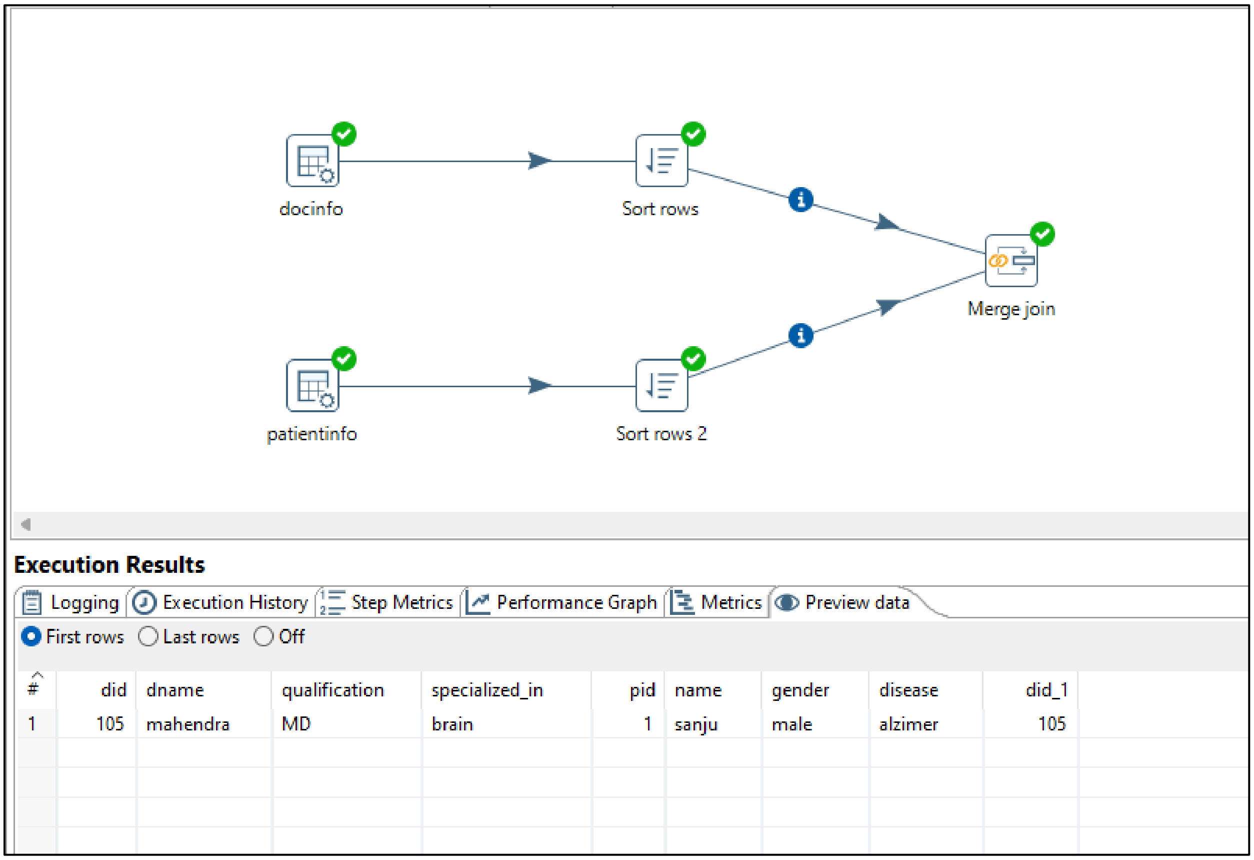


Execution Results

● First rows ○ Last rows ○ Off

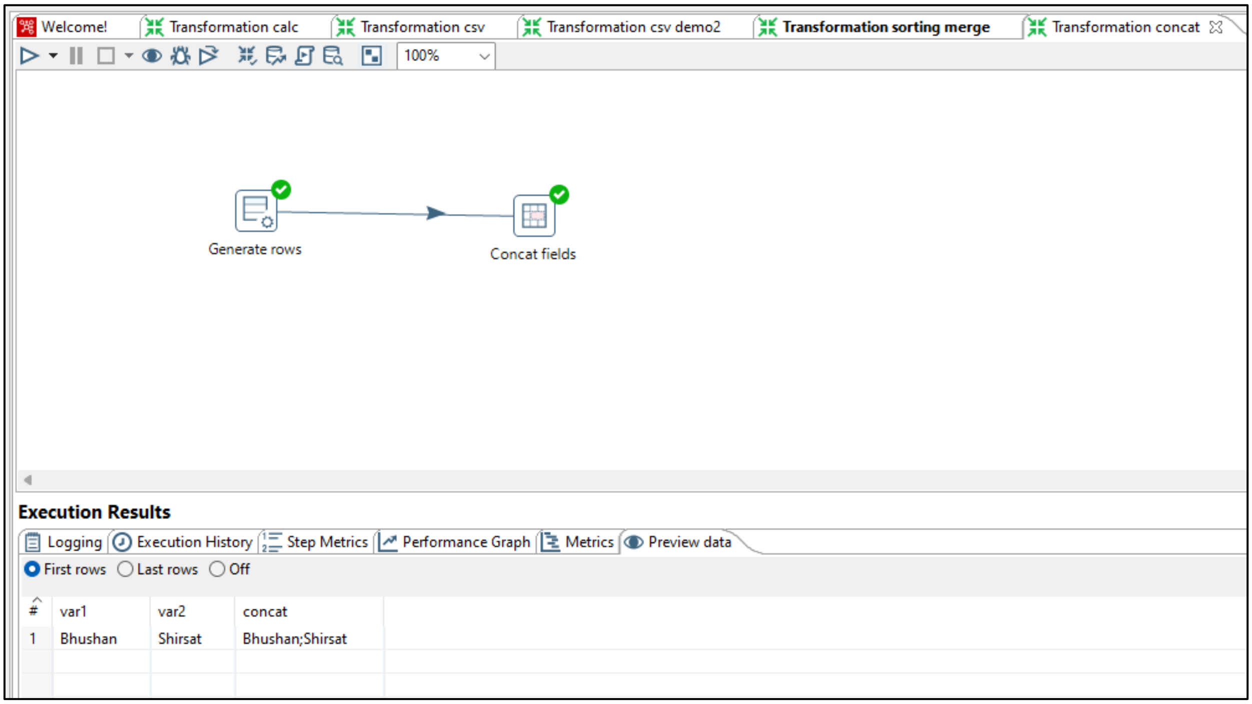
#	Index	Year	Age	Name	Movie	valuename
1	63	1990	80	"Jessica Tandy"	"Driving Miss Daisy"	1
2	55	1982	74	"Katharine Hepburn"	"On Golden Pond"	2
3	4	1931	63	"Marie Dressler"	"Min and Bill"	3
4	85	2012	62	"Meryl Streep"	"The Iron Lady"	4
5	80	2007	61	"Helen Mirren"	"The Queen"	5
6	59	1986	61	"Geraldine Page"	"The Trip to Bountiful"	6

SORT ROW



#	did	dname	qualification	specialized_in	pid	name	gender	disease	did_1
1	105	mahendra	MD	brain	1	sanju	male	alzimer	105

Concat



#	var1	var2	concat
1	Bhushan	Shirsat	Bhushan;Shirsat