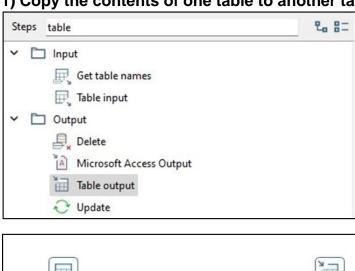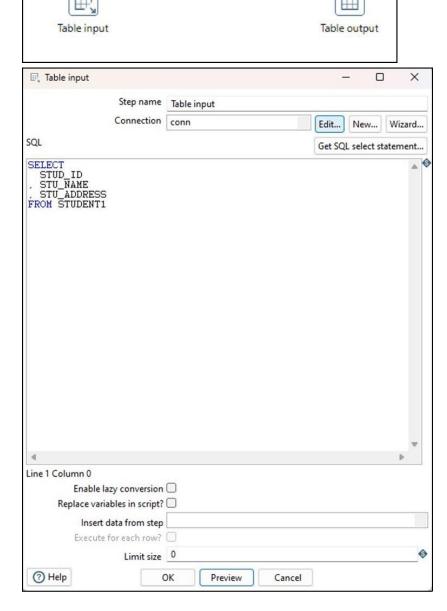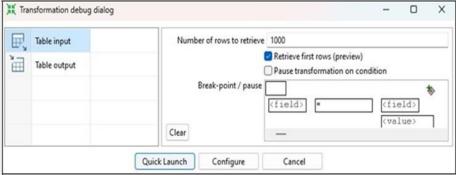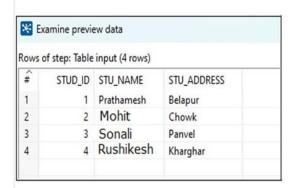# Implementation ETL transformation with Pentaho

1) Copy the contents of one table to another table

Output Fields...

Sniff Test During Execution   >

Check Selected Step(s)

Error Handling...

Preview...

Transformation debug dialog   —   □   ×

| Table input | Number of rows to retrieve | 1000 |
| Table output | | |

☑ Retrieve first rows (preview)
☐ Pause transformation on condition

Break-point / pause [   ]

`<field>` `=` `<field>`
`<value>`

Clear   —

Quick Launch   Configure   Cancel

Examine preview data

Rows of step: Table input (4 rows)

#	STUD_ID	STU_NAME	STU_ADDRESS
1	1	Prathamesh	Belapur
2	2	Mohit	Chowk
3	3	Sonali	Panvel
4	4	Rushikesh	Kharghar

Table input   Table output

Table output

Step name	Table output	
Connection	conn	Edit... New... Wizard...
Target schema	SYSTEM	Browse...
Target table	studentoutput1	Browse...
Commit size	1000	
Truncate table	☐	
Ignore insert errors	☐	
Specify database fields	☑	

Main options | Database fields

Fields to insert:

#	Table field	Stream field		
1	STUD_ID	STUD_ID		Get fields
2	STU_NAME	STU_NAME		Enter field mapping
3	STU_ADDRESS	STU_ADDRESS		

? Help   OK   Cancel   SQL

Transformation debug dialog

Table input	
Table output	

Number of rows to retrieve 1000

☑ Retrieve first rows (preview)
☐ Pause transformation on condition

Break-point / pause [   ]

`<field>` = [   ] `<field>`
`<value>`

Clear   ——

Quick Launch   Configure   Cancel

Examine preview data

Rows of step: Table output (4 rows)

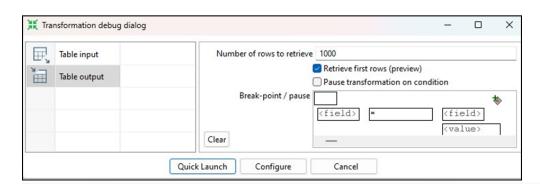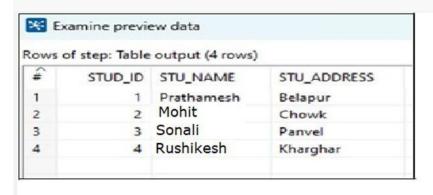| #	STUD_ID	STU_NAME	STU_ADDRESS
1	1	Prathamesh	Belapur
2	2	Mohit	Chowk
3	3	Sonali	Panvel
4	4	Rushikesh	Kharghar

```
SQL> select * from studentoutput1;

     STUD_ID STU_NAME    STU_ADDRESS
---------- ---------- --------------------
          1 Prathamesh Belapur
          2 Mohit       Chowk
          3 Sonali      Panvel
          4 Rushikesh   Kharghar
```
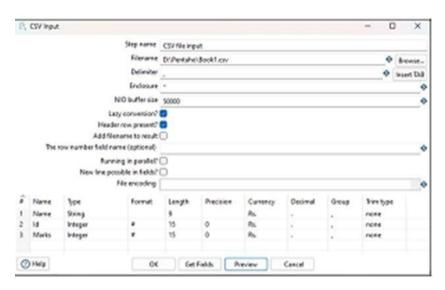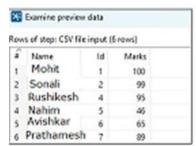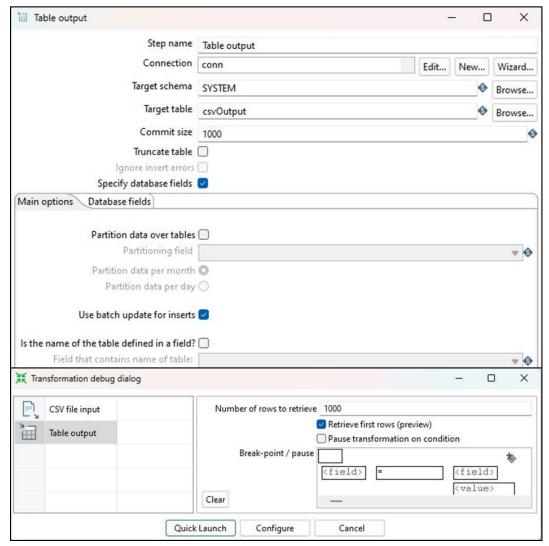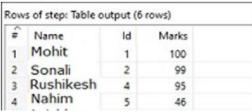
2) Copy contents of .CSV file to target table

Examine preview data

Rows of step: CSV file input (6 rows)

#	Name	Id	Marks
1	Mohit	1	100
2	Sonali	2	99
3	Rushikesh	4	95
4	Nahim	5	46
5	Avishkar	6	65
6	Prathamesh	7	89

CSV file input

CSV file input   Table output

Table output

Step name	Table output
Connection	conn Edit... New... Wizard...
Target schema	SYSTEM Browse...
Target table	csvOutput Browse...
Commit size	1000
Truncate table	☐
Ignore insert errors	☐
Specify database fields	☑

Main options \ Database fields

Partition data over tables ☐
Partitioning field [   ]

Partition data per month ◉
Partition data per day ○

Use batch update for inserts ☑

Is the name of the table defined in a field? ☐
Field that contains name of table: [   ]

Transformation debug dialog

📄	CSV file input
▦	Table output

Number of rows to retrieve 1000
☑ Retrieve first rows (preview)
☐ Pause transformation on condition

Break-point / pause [   ]

<field> = <field>
<value>

Clear   —

Quick Launch   Configure   Cancel

Rows of step: Table output (6 rows)

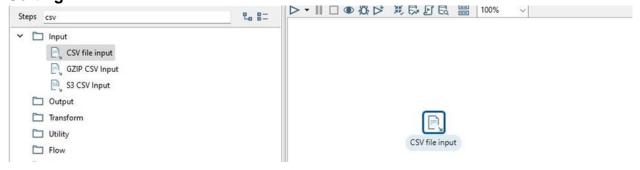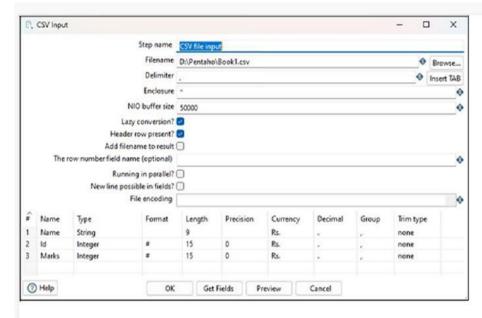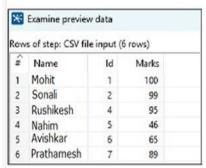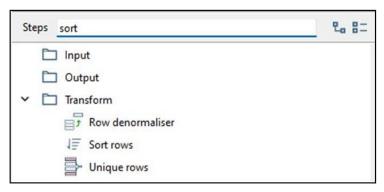| #	Name	Id	Marks
1	Mohit	1	100
2	Sonali	2	99
3	Rushikesh	4	95
4	Nahim	5	46

3) Implement ETL using the following transformations a) Sorting

CSV Input

Step name	CSV file input	
Filename	D:\Pentaho\Book1.csv	Browse...
Delimiter	,	Insert TAB
Enclosure	"	
NIO buffer size	50000	
Lazy conversion?	☑	
Header row present?	☑	
Add filename to result	☐	
The row number field name (optional)		
Running in parallel?	☐	
New line possible in fields?	☐	
File encoding		

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	Name	String		9		Rs.	.	,	none
2	Id	Integer	#	15	0	Rs.	.	,	none
3	Marks	Integer	#	15	0	Rs.	.	,	none

Help   OK   Get Fields   Preview   Cancel

Examine preview data

Rows of step: CSV file input (6 rows)

#	Name	Id	Marks
1	Mohit	1	100
2	Sonali	2	99
3	Rushikesh	4	95
4	Nahim	5	46
5	Avishkar	6	65
6	Prathamesh	7	89

CSV file input

Transformation debug dialog — □ ×

CSV file input	Number of rows to retrieve 1000
Sort rows	☑ Retrieve first rows (preview)
	☐ Pause transformation on condition
	Break-point / pause
	‹field› = ‹field›
	‹value›
	Clear

Quick Launch   Configure   Cancel

Examine preview data

Rows of step: Sort rows (6 rows)

#	Name	Id	Marks
1	Prathamesh	7	89
2	Mohit	1	100
3	Avishkar	6	65
4	Sonali	2	99
5	Rushikesh	4	95
6	Nahim	5	46

Steps output

- ∨ 📁 Output
  - Automatic Documentation Output
  - JSON Output
  - LDAP Output
  - Microsoft Access Output
  - Microsoft Excel Output
  - Pentaho Reporting Output
  - Properties Output
  - RSS Output

▷ ▾ II □ ◉ ⚙ ▷ ⚞ ⚟ 🔍 100% ∨

CSV file input → Sort rows → Microsoft Excel Output

Microsoft Excel output — □ ×

Step name Microsoft Excel Output

File \ Content \ Custom \ Fields

Filename D:\Pentaho\csvOutput.xls ◈ Browse...
Create Parent folder ☐
Do not create file at start ☐
Extension xls ◈
Include stepnr in ☐
Include date in ☐
Include time in ☐
Specify Date time ☐
Date time format ▼

Show filename(s)...

Add filenames to result ☑

OK   Cancel

⑦ Help

Microsoft Excel output

Step name   Microsoft Excel Output

File | Content | Custom | Fields

#	Name	Type	Format	
1	Name	String		
2	Id	Integer		
3	Marks	Integer		

Get Fields   Minimal width

OK   Get the fields as defined in previous steps.

? Help

Transformation debug dialog

CSV file input	
Sort rows	
Microsoft Excel Output	

Number of rows to retrieve   1000

☑ Retrieve first rows (preview)
☐ Pause transformation on condition

Break-point / pause

`<field>`   =   `<field>`
`<value>`

Clear

Quick Launch   Configure   Cancel

Examine preview data

Rows of step: Microsoft Excel Output (6 ro

#	Name	Id	Marks
1	Prathamesh	7	89
2	Mohit	1	100
3	Avishkar	6	65
4	Sonali	2	99
5	Rushikesh	4	95
6	Nahim	5	46

	A	B	C	D
1	Name	Id	Marks	
2	Prathamesh	7.	89.	
3	Mohit	1.	100.	
4	Avishkar	6.	65.	
5	Sonali	2.	99.	
6	Rushikesh	4.	95.	
7	Nahim	5.	46.	
8				

b) Adding a sequence

Examine preview data

Rows of step: Add sequence (6 rows)

#	Name	Id	Marks	Sequence
1	Mohit	1	100	1
2	Sonali	2	99	2
3	Rushikesh	4	95	3
4	Nahim	5	46	4
5	Avishkar	6	65	5
6	Prathamesh	7	89	6

Steps seq

- Input
- Output
- ∨ Transform
  - Add sequence
  - Add value fields changing sequence
  - Get ID from slave server
- Utility
- ∨ Flow
  - Blocking Step

100%

CSV file input ➝ Add sequence

c) Concat Field

Steps data

- ∨ Input
  - Data Grid
  - De-serialize from file
  - ESRI Shapefile Reader

Add constant rows   — □ ✕

Step name Data Grid

Meta \ Data

#	Name	Type	Format	Length
1	ID	Integer		
2	FNAME	String		
3	LNAME	String		
4	ADDRESS	String		

OK   Preview   Cancel

⊘ Help

Add constant rows

Step name: Data Grid

Meta | Data

#	ID	FNAME	LNAME	ADDRESS
1	1	Mohit	SINGH	PUNJAB
2	2	Sonali	GHEMUD	PANVEL
3	3	Rushikesh	KHOSE	VASHI
4	4	Nahim	SHINDE	ANDHERI

OK   Preview   Cancel

⑦ Help

Examine preview data

Rows of step: Data Grid (4 rows)

#	ID	FNAME	LNAME	ADDRESS
1	1	Mohit	SINGH	PUNJAB
2	2	Sonali	GHEMUD	PANVEL
3	3	Rushikesh	KHOSE	VASHI
4	4	Prathamesh	SHINDE	ANDHERI

☐ Input
☐ Output
⌄ ☐ Transform
    ☐ Concat Fields

Data Grid → Concat Fields

Concat Fields

Step name: Concat Fields
Target Field Name: FULL NAME
Length of Target Field: 0
Separator: _
Enclosure: 1   Insert TAB

Fields | Advanced

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim Type	Null
1	FNAME	String							none	
2	LNAME	String							none	

Get Fields   Minimal width

OK   Cancel

⑦ Help

Examine preview data

Rows of step: Concat Fields (4 rows)

#	ID	FNAME	LNAME	ADDRESS	FULL NAME
1	1	MOHIT	DANDGE	PUNJAB	MOHIT_DANDGE
2	2	SONALI	JADHAV	PANVEL	SONALI_JADHAV

Data Grid → Concat Fields

Data Grid → Concat Fields → Microsoft Excel Output

d) Split Fields

Field splitter

Step name	Split Fields
Field to split	FULL NAME
Delimiter	_
Enclosure	

Fields

#	New field	ID	Remove ID?	Type	Length	Precision	Format
1	FIRST		N	String			
2	LAST		N	String			

Help   OK   Cancel

Rows of step: Number range (4 rows)

#	ID	NAME	SALARY	WITHOUT TAX	RANGE
1	1.0	MOHIT	2000.0	3400	Less than 5000
2	2.0	SONALI	5000.0	8500	5000 TO 10000
3	3.0	RUSHIKESH	7000.0	11900	5000 TO 10000
4	4.0	NAHIM	10000.0	17000	5000 TO 10000

Examine preview data

Rows of step: Split Fields (4 rows)

#	ID	FNAME	LNAME	ADDRESS	FIRST	LAST
1	1	MOHIT	DANDGE	PUNJAB	MOHIT	DANDGE
2	2	SONALI	JADHAV	PANVEL	SONALI	JADHAV
3	3	RUSHIKESH	TURE	VASHI	RUSHIKESH	TURE
4	4	NAHIM	PINJARI	ANDHERI	NAHIM	PINJARI

Data Grid → Concat Fields → Microsoft Excel Output → Split Fields

e) Adding Formula

Add constant rows

Step name   Data Grid

Meta   Data

#	Name	Type	Format
1	ID	Number	
2	NAME	String	
3	SALARY	Number	

? Help   OK   Preview   Cancel

Data Grid

Add constant rows — □ ✕

Step name Data Grid

Meta / Data

#	ID	NAME	SALARY
1	1	MOHIT	2000
2	2	SONALI	5000
3	3	RUSHIKESH	4000
4	4	NAHIM	2500

OK   Preview   Cancel
? Help

fx **Formula** — □ ✕

Step name Formula

Fields:

#	New field	Formula	Value type	Length	Precision	Replace valu
1	WITHOUT TAX	[SALARY]*1.70	Number			
2						

? Help   OK   Cancel

Examine preview data

Rows of step: Formula (4 rows)

#	ID	NAME	SALARY	WITHOUT TAX
1	1.0	MOHIT	2000.0	3400
2	2.0	SONALI	5000.0	8500
3	3.0	RUSHIKESH	4000.0	6800
4	4.0	NAHIM	2500.0	4250

Data Grid → Formula → Microsoft Excel Output

f) Adding Number Range

Number ranges

Step name:	Number range
Input field:	SALARY
Output field:	RANGE
Default value(if no range matches):	unknown

Ranges (min <= x< max):

#	Lower Bound	Upper Bound	Value
1	0.0	5000.0	Less than 5000
2	6000.0	10000.0	5000 TO 6000

OK   Cancel

? Help

g) String Operation

String operations

Step name   String operations

The fields to process:

#	In stream field	Out stream field	Trim type	Lower/Upper	Padding
1	NAME	NAME1	none	lower	none

? Help   OK   Get fields   Cancel

Rows of step: String operations (4 rows)

#	ID	NAME	SALARY	NAME1
1	1.0	MOHIT	2000.0	mohit
2	2.0	SONALI	5000.0	sonali
3	3.0	RUSHIKESH	7000.0	rushikesh
4	4.0	NAHIM	10000.0	nahim

4) Implement the merge join transformation on tables.

Add constant rows

Step name: EMPLOYEE

Meta | Data

#	Name	Type	Format
1	ID	Number	
2	NAME	String	
3	DEPT_ID	String	
4			

OK   Preview   Cancel

? Help

Examine preview data

Rows of step: EMPLOYEE (4 rows)

#	ID	NAME	DEPT_ID
1	1.0	Mohit	D1
2	2.0	Sonali	D2
3	3.0	Rushikesh	D2
4	4.0	Nahim	D3

Add constant rows — □ ×

Step name DEPARTMENT

Meta Data

#	Name	Type	Form
1	DEPT_ID	String	
2	DEPT_NAME	String	

OK   Preview   Cancel

? Help

Examine preview data

Rows of step: DEPARTMENT (4 rows)

#	DEPT_ID	DEPT_NAME
1	D1	SALES
2	D2	TECHNICAL
3	D3	MARKETING

Merge Join — □ ×

Step name Merge Join

First Step: Sort rows

Second Step: Sort rows 2

Join Type: INNER

Keys for 1st step:

#	Key field
1	DEPT_ID

Keys for 2nd step:

#	Key field
1	DEPT_ID

Get key fields   Get key fields

? Help   OK   Cancel

Examine preview data

Rows of step: Merge Join (4 rows)

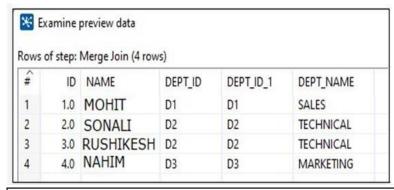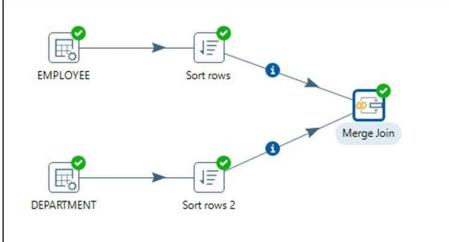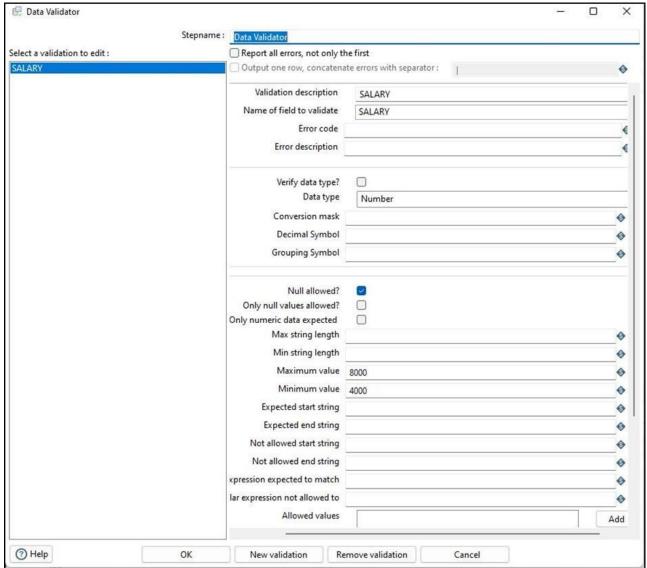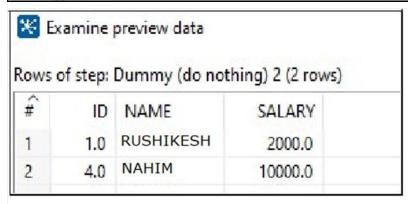| #	ID	NAME	DEPT_ID	DEPT_ID_1	DEPT_NAME
1	1.0	MOHIT	D1	D1	SALES
2	2.0	SONALI	D2	D2	TECHNICAL
3	3.0	RUSHIKESH	D2	D2	TECHNICAL
4	4.0	NAHIM	D3	D3	MARKETING

5) Implement data validations on the table data.
(Perform the transformations such that the validated data is transferred to a table and the invalid data is transferred to another table)

Data Validator

Stepname : Data Validator

☐ Report all errors, not only the first

☐ Output one row, concatenate errors with separator : |

Select a validation to edit :

SALARY

Field	Value
Validation description	SALARY
Name of field to validate	SALARY
Error code	
Error description	
Verify data type?	☐
Data type	Number
Conversion mask	
Decimal Symbol	
Grouping Symbol	
Null allowed?	☑
Only null values allowed?	☐
Only numeric data expected	☐
Max string length	
Min string length	
Maximum value	8000
Minimum value	4000
Expected start string	
Expected end string	
Not allowed start string	
Not allowed end string	
xpression expected to match	
lar expression not allowed to	
Allowed values	Add

ⓘ Help   OK   New validation   Remove validation   Cancel

Examine preview data

Rows of step: Dummy (do nothing) 2 (2 rows)

#	ID	NAME	SALARY
1	1.0	RUSHIKESH	2000.0
2	4.0	NAHIM	10000.0

Examine preview data

Rows of step: Dummy (do nothing) (2 rows)

#	ID	NAME	SALARY
1	2.0	MOHIT	5000.0
2	3.0	SONALI	7000.0