
**Course:-MCAL13 Advanced Database Management System
Lab**

Practical – 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.

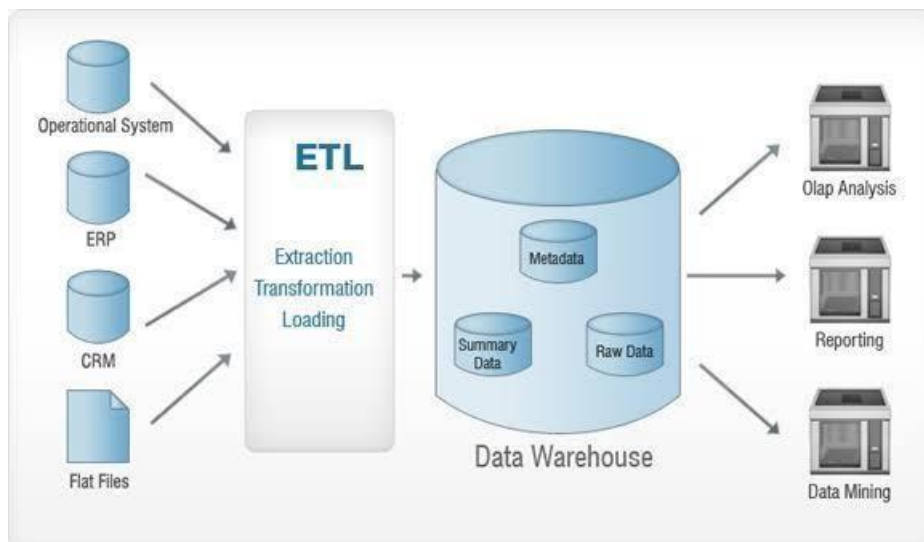
Pentaho, a subsidiary of Hitachi Vantara, is free and open- source platform for data integration and analytics. The software comes in a free community edition and a subscription-based enterprise edition.

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)



Finolex Academy of Management & Technology, Ratnagiri
Department of MCA

**Course:-MCAL13 Advanced Database Management System
Lab**

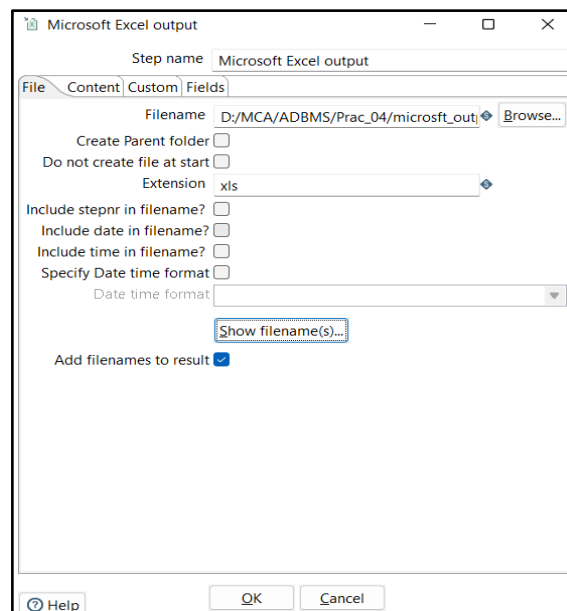
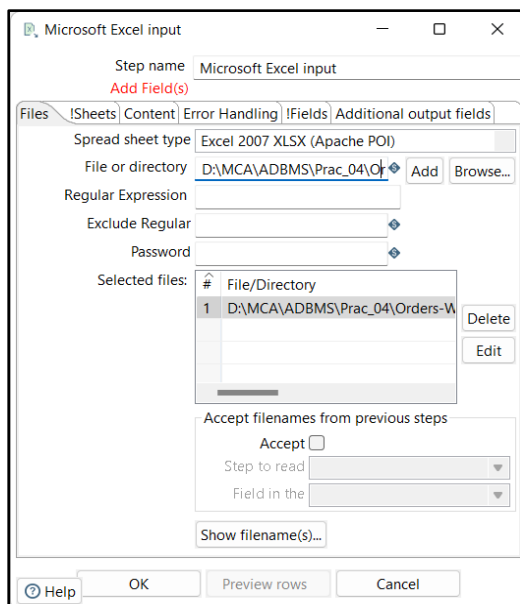
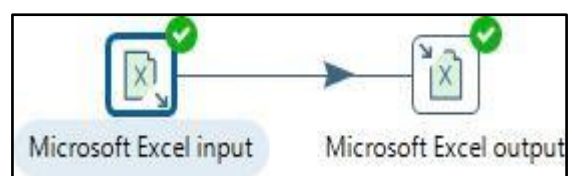
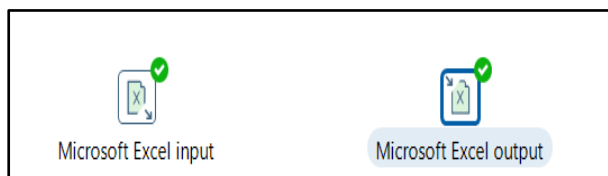
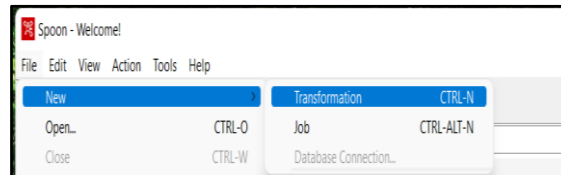
PDI Architecture

Components of PDI (KETTLE)

| Tool | Windows command | Linux/ MacOS command |
|---|--------------------|-------------------------|
| Spoon is the graphical interface used to create transformations and jobs. | spoon.bat | spoon.sh |
| Pan is a batch-style command line tool used to run transformations | pan.bat | pan.sh |
| Kitchen is a batch-style command line tool used to run jobs. | Kitchen.bat | Kitchen.sh |
| Carte is a web server that can be used to run jobs on remote servers. | Carte.bat | Carte.sh |

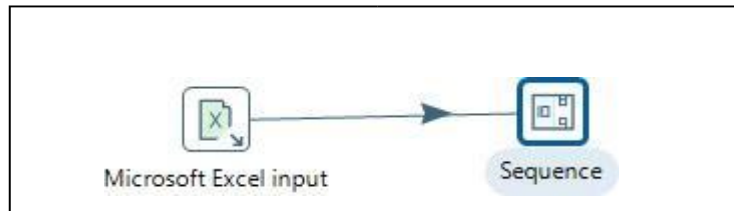
Course:-MCAL13 Advanced Database Management System Lab

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



Course:-MCAL13 Advanced Database Management System Lab

Sequence



Add sequence

Step name:

Name of value:

Use a database to generate the sequence

Use DB to get sequence? ☐

Connection: Edit... New... Wizard...

Schema name: Schemas...

Sequence name: Sequences...

Use a transformation counter to generate the sequence

Use counter to calculate sequence? ☒

Counter name (optional):

Start at value:

Increment by:

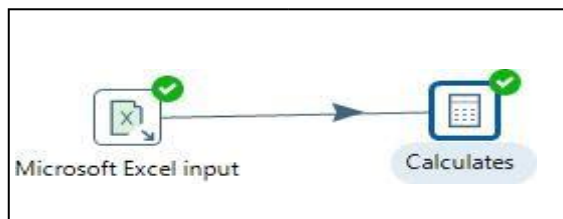
Maximum value:

Help OK Cancel

☒ First rows ☐ Last rows ☐ Off

| Order Date | Order Quantity | Sales | Ship Mode | Profit | Unit Price | Customer Name | Customer Segment | Product Category | Sequence |
|-------------------------|----------------|-----------|----------------|----------|------------|--------------------|------------------|------------------|----------|
| 2010/10/13 00:00:00.000 | 6.0 | 261.54 | Regular Air | -213.25 | 38.94 | Muhammed MacIntyre | Small Business | Office Supplies | 1 |
| 2012/02/20 00:00:00.000 | 2.0 | 6.93 | Regular Air | -4.64 | 2.08 | Ruben Dartt | Corporate | Office Supplies | 2 |
| 2011/07/15 00:00:00.000 | 26.0 | 2808.08 | Regular Air | 1054.82 | 107.53 | Liz Pelletier | Corporate | Furniture | 3 |
| 2011/07/15 00:00:00.000 | 24.0 | 1761.4 | Delivery Truck | -1748.56 | 70.89 | Liz Pelletier | Corporate | Furniture | 4 |
| 2011/07/15 00:00:00.000 | 23.0 | 160.2335 | Regular Air | -85.129 | 7.99 | Liz Pelletier | Corporate | Technology | 5 |
| 2011/07/15 00:00:00.000 | 15.0 | 140.56 | Regular Air | -128.38 | 8.46 | Liz Pelletier | Corporate | Technology | 6 |
| 2011/10/22 00:00:00.000 | 30.0 | 288.56 | Regular Air | 60.72 | 9.11 | Julie Creighton | Corporate | Office Supplies | 7 |
| 2011/10/22 00:00:00.000 | 14.0 | 1892.848 | Regular Air | 48.987 | 155.99 | Julie Creighton | Corporate | Technology | 8 |
| 2011/11/02 00:00:00.000 | 46.0 | 2484.7455 | Regular Air | 657.477 | 65.99 | Sample Company A | Home Office | Technology | 9 |

Adding Calculator



Calculator

Step name:

☒ Throw an error on non existing files

Fields:

| # | New field | Calculation | Field A | Field B | Field C | Value type |
|---|-----------|-------------|------------|---------|---------|------------|
| 1 | Calculate | A * B | Unit Price | Profit | | Number |
| | | | | | | |
| | | | | | | |
| | | | | | | |

Finolex Academy of Management & Technology, Ratnagiri

Department of MCA

Course:-MCAL13 Advanced Database Management System Lab

| <input checked="" type="radio"/> First rows <input type="radio"/> Last rows <input type="radio"/> Off | | | | | | | | | |
|---|-----------|----------------|----------|------------|--------------------|------------------|------------------|--------------|--|
| Order Quantity | Sales | Ship Mode | Profit | Unit Price | Customer Name | Customer Segment | Product Category | Calculate | |
| 6.0 | 261.54 | Regular Air | -213.25 | 38.94 | Muhammed MacIntyre | Small Business | Office Supplies | -8303.955 | |
| 2.0 | 6.93 | Regular Air | -4.64 | 2.08 | Ruben Dartt | Corporate | Office Supplies | -9.6512 | |
| 26.0 | 2808.08 | Regular Air | 1054.82 | 107.53 | Liz Pelletier | Corporate | Furniture | 113424.7946 | |
| 24.0 | 1761.4 | Delivery Truck | -1748.56 | 70.89 | Liz Pelletier | Corporate | Furniture | -123955.4184 | |
| 23.0 | 160.2335 | Regular Air | -85.129 | 7.99 | Liz Pelletier | Corporate | Technology | -680.18071 | |
| 15.0 | 140.56 | Regular Air | -128.38 | 8.46 | Liz Pelletier | Corporate | Technology | -1086.0948 | |
| 30.0 | 288.56 | Regular Air | 60.72 | 9.11 | Julie Creighton | Corporate | Office Supplies | 553.1592 | |
| 14.0 | 1892.848 | Regular Air | 48.987 | 155.99 | Julie Creighton | Corporate | Technology | 7641.48213 | |
| 46.0 | 2484.7455 | Regular Air | 657.477 | 65.99 | Sample Company A | Home Office | Technology | 43386.90723 | |

Concatenation Two field

Concat fields

Step name

Concat fields

Target Field Name

Concat

Length of Target Field

0

Separator

:

Insert TAB

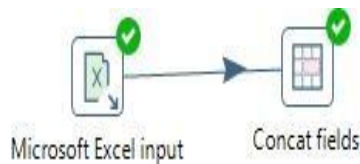
Enclosure

"

Fields

Advanced

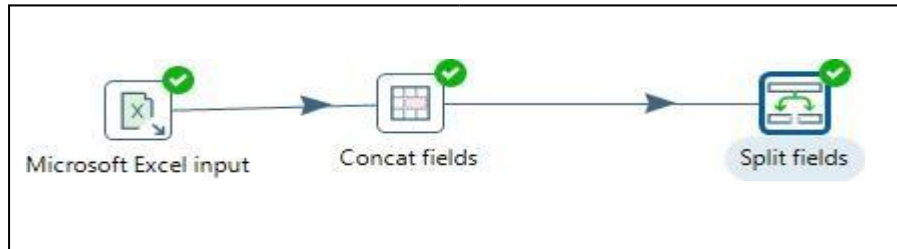
| # | Name | Type | Format | Length | Precision | Currency | Decimal | Group | Trim Type |
|---|------------------|--------|--------|--------|-----------|----------|---------|-------|-----------|
| 1 | Customer Name | String | | | | | | | |
| 2 | Customer Segment | String | | | | | | | |



| <input checked="" type="radio"/> First rows <input type="radio"/> Last rows <input type="radio"/> Off | | | | | | | | | |
|---|-----------|----------------|----------|------------|--------------------|------------------|------------------|-----------------------------------|--|
| Order Quantity | Sales | Ship Mode | Profit | Unit Price | Customer Name | Customer Segment | Product Category | Concat | |
| 6.0 | 261.54 | Regular Air | -213.25 | 38.94 | Muhammed MacIntyre | Small Business | Office Supplies | Muhammed MacIntyre:Small Business | |
| 2.0 | 6.93 | Regular Air | -4.64 | 2.08 | Ruben Dartt | Corporate | Office Supplies | Ruben Dartt:Corporate | |
| 26.0 | 2808.08 | Regular Air | 1054.82 | 107.53 | Liz Pelletier | Corporate | Furniture | Liz Pelletier:Corporate | |
| 24.0 | 1761.4 | Delivery Truck | -1748.56 | 70.89 | Liz Pelletier | Corporate | Furniture | Liz Pelletier:Corporate | |
| 23.0 | 160.2335 | Regular Air | -85.129 | 7.99 | Liz Pelletier | Corporate | Technology | Liz Pelletier:Corporate | |
| 15.0 | 140.56 | Regular Air | -128.38 | 8.46 | Liz Pelletier | Corporate | Technology | Liz Pelletier:Corporate | |
| 30.0 | 288.56 | Regular Air | 60.72 | 9.11 | Julie Creighton | Corporate | Office Supplies | Julie Creighton:Corporate | |
| 14.0 | 1892.848 | Regular Air | 48.987 | 155.99 | Julie Creighton | Corporate | Technology | Julie Creighton:Corporate | |
| 46.0 | 2484.7455 | Regular Air | 657.477 | 65.99 | Sample Company A | Home Office | Technology | Sample Company A:Home Office | |

Course:-MCAL13 Advanced Database Management System Lab

Splitting Two Fields



Split fields

Step name: Split fields

Field to split: Customer Name

Delimiter: ,

Enclosure: "

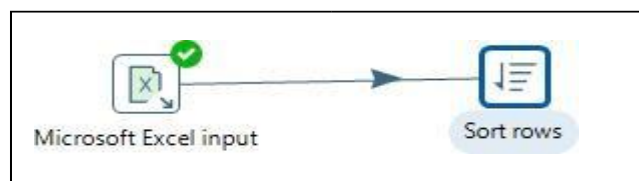
| # | New field Name | ID | Remove ID? | Type | Length | Precision | Format | Group | Decimal | Currency | Nullif | Default | Trim type |
|---|----------------|----|------------|--------|--------|-----------|--------|-------|---------|----------|--------|---------|-----------|
| 1 | Name | | | String | | | | | | | | | |
| 2 | segment | | | String | | | | | | | | | |

Help OK Cancel

☒ First rows ☐ Last rows ☐ Off

| Order Quantity | Sales | Ship Mode | Profit | Unit Price | Customer Name | Customer Segment | Product Category | Name | segment |
|----------------|-----------|----------------|----------|------------|--------------------|------------------|------------------|--------------------|----------------|
| 6.0 | 261.54 | Regular Air | -213.25 | 38.94 | Muhammed MacIntyre | Small Business | Office Supplies | Muhammed MacIntyre | Small Business |
| 2.0 | 6.93 | Regular Air | -4.64 | 2.08 | Ruben Dartt | Corporate | Office Supplies | Ruben Dartt | Corporate |
| 26.0 | 2808.08 | Regular Air | 1054.82 | 107.53 | Liz Pelletier | Corporate | Furniture | Liz Pelletier | Corporate |
| 24.0 | 1761.4 | Delivery Truck | -1748.56 | 70.89 | Liz Pelletier | Corporate | Furniture | Liz Pelletier | Corporate |
| 23.0 | 160.2335 | Regular Air | -85.129 | 7.99 | Liz Pelletier | Corporate | Technology | Liz Pelletier | Corporate |
| 15.0 | 140.56 | Regular Air | -128.38 | 8.46 | Liz Pelletier | Corporate | Technology | Liz Pelletier | Corporate |
| 30.0 | 288.56 | Regular Air | 60.72 | 9.11 | Julie Creighton | Corporate | Office Supplies | Julie Creighton | Corporate |
| 14.0 | 1892.848 | Regular Air | 48.987 | 155.99 | Julie Creighton | Corporate | Technology | Julie Creighton | Corporate |
| 46.0 | 2484.7455 | Regular Air | 657.477 | 65.99 | Sample Company A | Home Office | Technology | Sample Company A | Home Office |

Sorting Data



Sort rows

Step name: Sort rows

Sort directory: %%java.io.tmpdir%%

TMP-file prefix: out

Sort size (rows in memory): 1000000

Free memory threshold (in %):

Compress TMP Files? ☐

Only pass unique rows? (verifies keys only) ☐

| # | Fieldname | Ascending | Case sensitive compare? | Sort based on current locale? | Collator Strength | Presorted? |
|---|---------------|-----------|-------------------------|-------------------------------|-------------------|------------|
| 1 | Customer Name | N | N | N | 0 | N |
| 2 | Order ID | Y | N | N | 0 | N |

Course:-MCAL13 Advanced Database Management System Lab

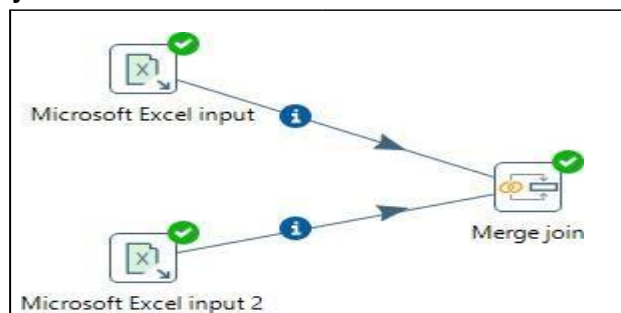
Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

First rows Last rows Off

| # | Order ID | Order Date | Order Quantity | Sales | Ship Mode | Profit | Unit Price | Customer Name | Customer Segment | Product Category |
|---|----------|-------------------------|----------------|-----------|----------------|---------|------------|----------------|------------------|------------------|
| 1 | 2882.0 | 2011/08/21 00:00:00.000 | 23.0 | 3872.87 | Delivery Truck | 565.34 | 160.98 | Yoseph Carroll | Consumer | Furniture |
| 2 | 2882.0 | 2011/08/21 00:00:00.000 | 9.0 | 356.72 | Regular Air | 12.61 | 40.98 | Yoseph Carroll | Consumer | Technology |
| 3 | 4514.0 | 2009/04/29 00:00:00.000 | 28.0 | 2841.4395 | Regular Air | 374.625 | 125.99 | Yoseph Carroll | Consumer | Technology |
| 4 | 4935.0 | 2010/05/24 00:00:00.000 | 30.0 | 106.64 | Regular Air | -31.95 | 3.68 | Xylona Price | Corporate | Office Supplies |
| 5 | 5254.0 | 2012/07/25 00:00:00.000 | 31.0 | 1735.3515 | Regular Air | 258.624 | 65.99 | William Brown | Corporate | Technology |

Implements the merge join transformation on tables



Merge join

Step name: Merge join

First Step: Microsoft Excel input

Second Step: Microsoft Excel input 2

Join Type: FULL OUTER

Keys for 1st step:

| # | Key field |
|---|-----------|
| 1 | Name |
| 2 | Address |

Keys for 2nd step:

| # | Key field |
|---|------------|
| 1 | id |
| 2 | name |
| 3 | salary |
| 4 | start_date |
| 5 | dept |

First rows Last rows Off

| # | Name | Address | id | name_1 | salary | start_date | dept |
|----|---------|-----------|--------|----------|--------|------------|------------|
| 1 | <null> | <null> | 1.0 | Rick | 623.3 | 2012-01-01 | IT |
| 2 | <null> | <null> | 2.0 | Dan | 515.2 | 2013-09-23 | Operations |
| 3 | <null> | <null> | 3.0 | Michelle | 611.0 | 2014-11-15 | IT |
| 4 | <null> | <null> | 4.0 | Ryan | 729.0 | 2014-05-11 | HR |
| 5 | <null> | <null> | 5.0 | Gary | 843.25 | 2015-03-27 | Finance |
| 6 | <null> | <null> | 6.0 | Nina | 578.0 | 2013-05-21 | IT |
| 7 | <null> | <null> | 7.0 | Simon | 632.8 | 2013-07-30 | Operations |
| 8 | <null> | <null> | 8.0 | Guru | 722.5 | 2014-06-17 | Finance |
| 9 | Pritesh | Ratnagiri | <null> | <null> | <null> | <null> | <null> |
| 10 | Gaurav | Devrukh | <null> | <null> | <null> | <null> | <null> |
| 11 | Vivek | Mumbai | <null> | <null> | <null> | <null> | <null> |