



ETL PROJECT USING PENTAHO DATA INTEGRATION (SPOON)



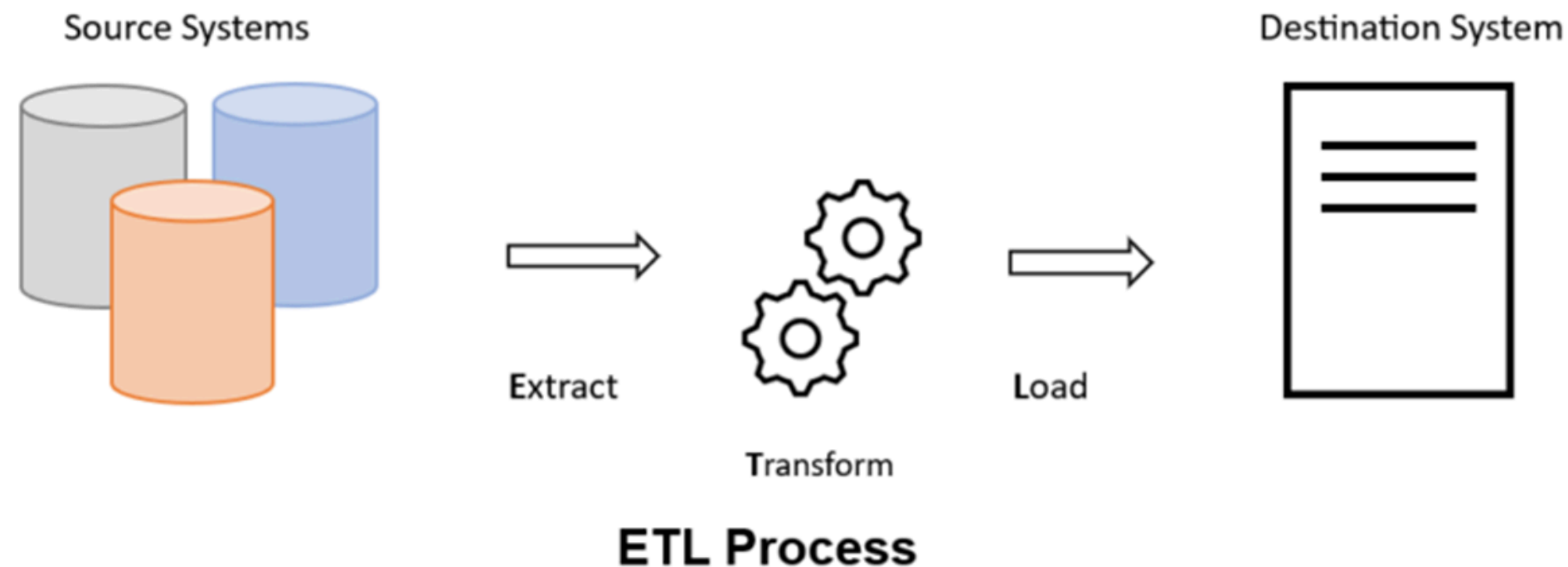
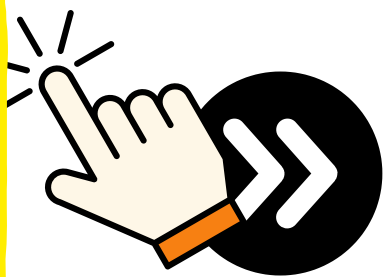
A Complete End-to-End ETL Process:
Data Extraction, Transformation & Loading



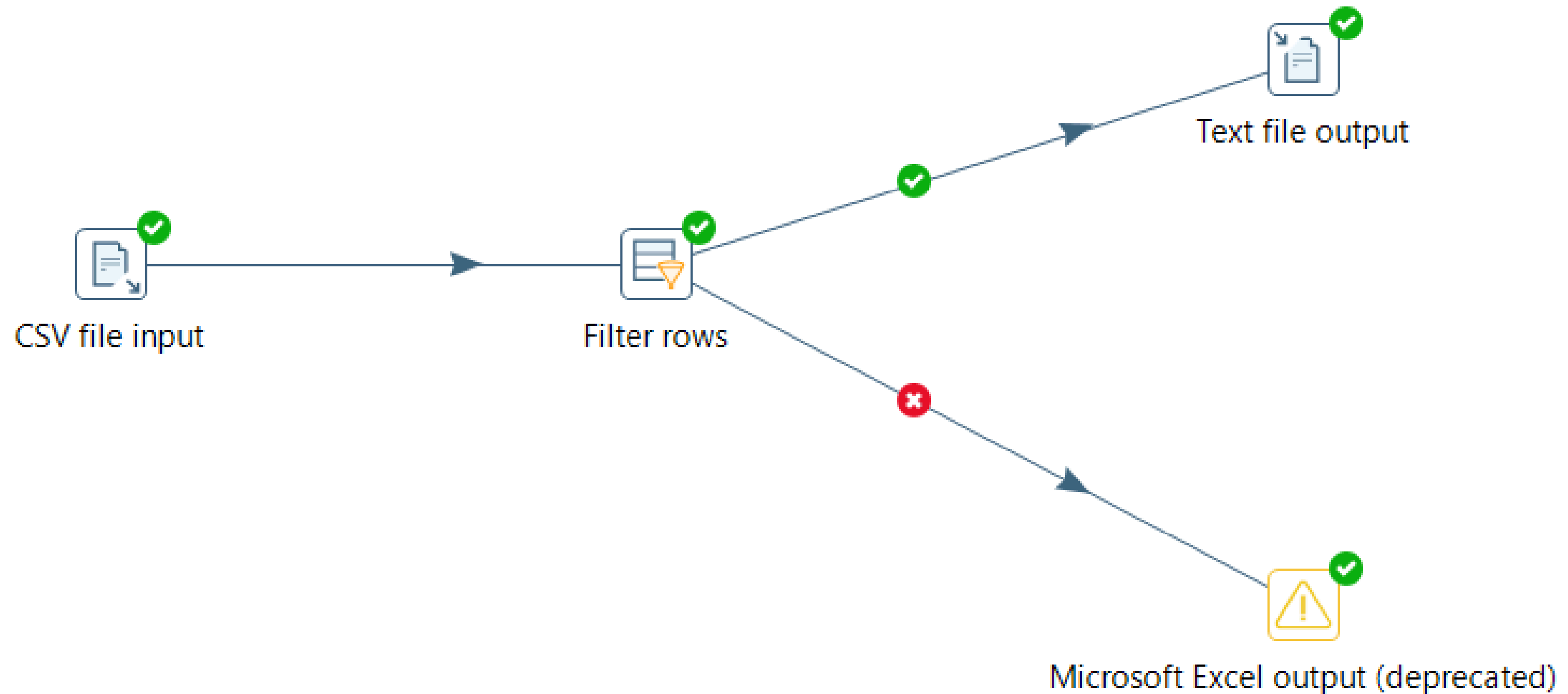
By : Ramesh Kumar Prajapati

Pentaho Data Integration

Spoon, a graphical user interface for designing transformations and jobs in Pentaho Data Integration (PDI) — a popular ETL tool (Extract, Transform, Load)




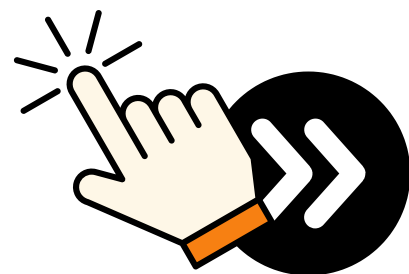
Transformation 1






ETL Flow (Transformation 1)

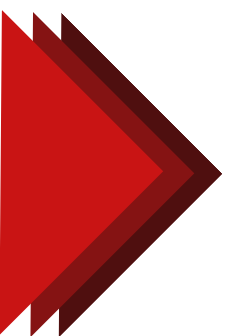
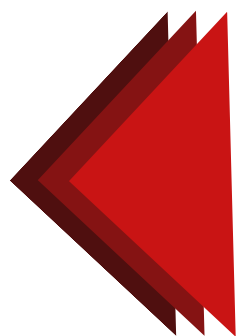
1. CSV File Input

-  Extract Step
- This step reads data from a CSV file (Comma-Separated Values file).
- It is the starting point where raw data is brought into the ETL pipeline.



2. Filter Rows

-  Transform Step
- This step is used to filter data based on a condition.
- It splits the flow into two paths:
 -  Green path: Rows that match the condition (True).
 -  Red path: Rows that don't match the condition (False).




3.Text File Output

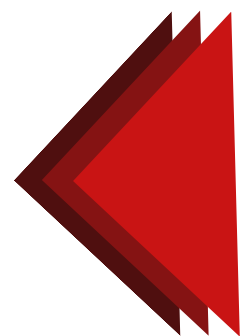
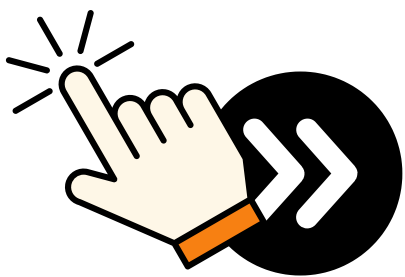
Load Step (for true condition)

- This step writes the filtered (true condition) data to a text file.

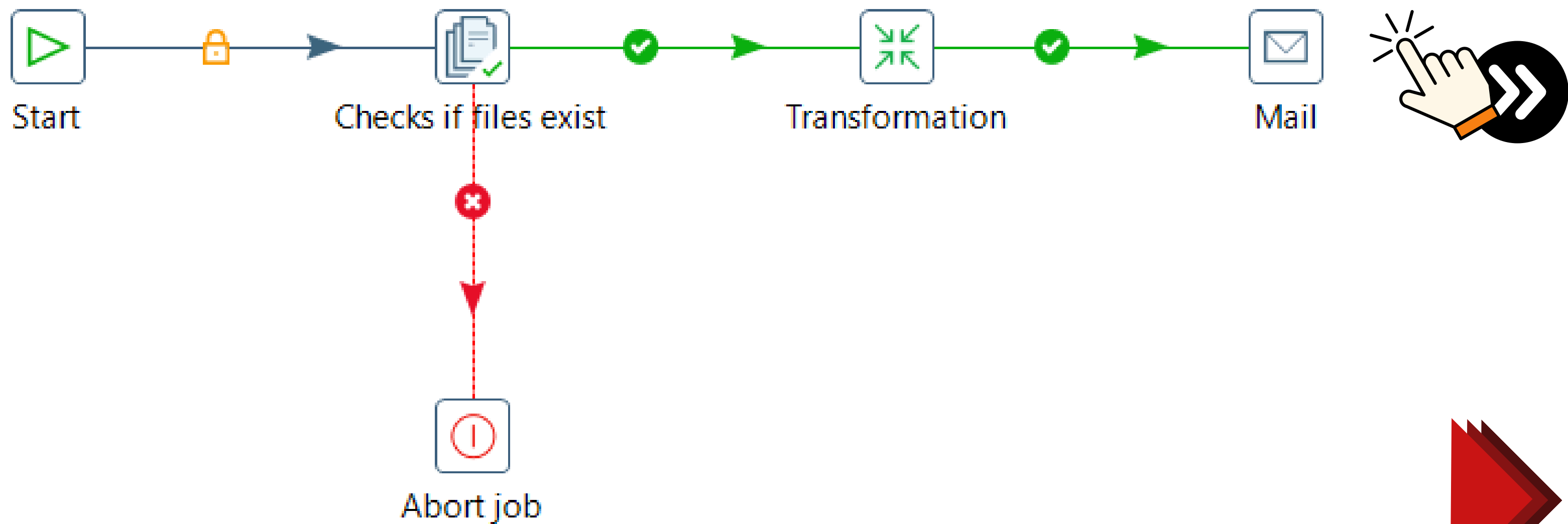
4.Microsoft Excel Output (deprecated)

Load Step (for false condition)

- Writes the false-condition data to an Excel file.
-  **Deprecated:** This means the step is outdated and not recommended for future use. Consider replacing it with a newer Excel writer step like "Excel Writer (Streaming)".



Job 1





Job Workflow (Job 1)

1. ● Start

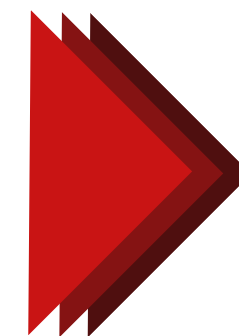
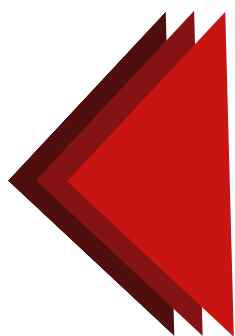
- Entry point of the job.
- The execution begins here.

2. 📁 Checks if files exist

- This step verifies if required input files exist in the file system.

✓ Green arrow (success): If the files are found, proceed to the transformation.

- ✗ Red arrow (failure): If files are not found, go to “Abort job”.

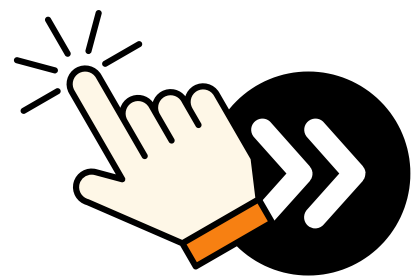


3. ⚠️ Abort job

- If files are missing, this step immediately stops the job.
- This helps prevent downstream processes from running on missing/incomplete data.

4. 🔄 Transformation

- This runs a transformation file (ETL process).
- Usually used to extract, clean, and load data.

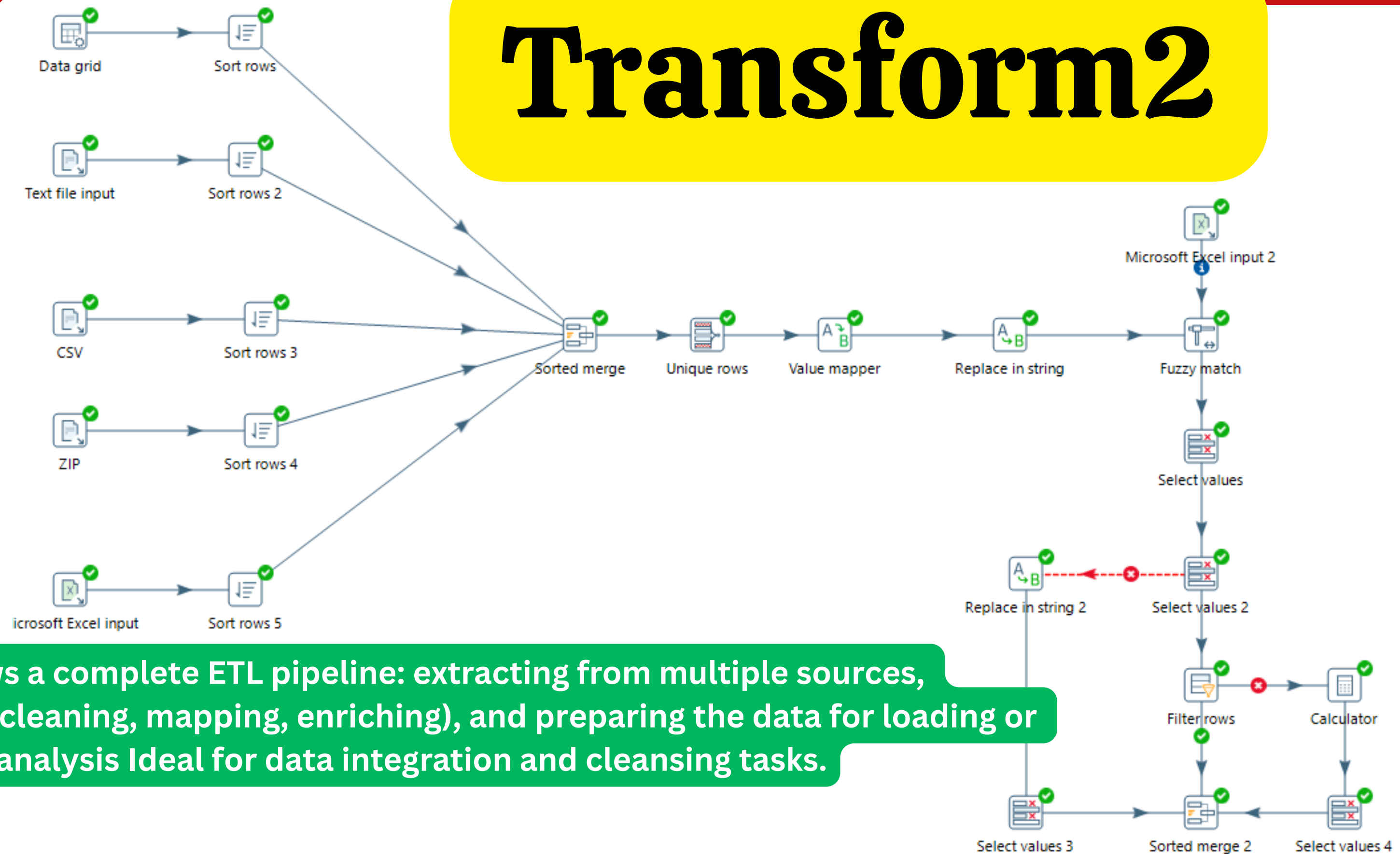


5. ✉️ Mail

- Sends an email notification.
- Typically used to inform stakeholders of job success or failure (can include logs, success messages, file counts, etc.).



Transform2



This flow shows a complete ETL pipeline: extracting from multiple sources, transforming (cleaning, mapping, enriching), and preparing the data for loading or analysis. Ideal for data integration and cleansing tasks.



ETL Flow (Transformation 2)

1.Data Sources (Left side):

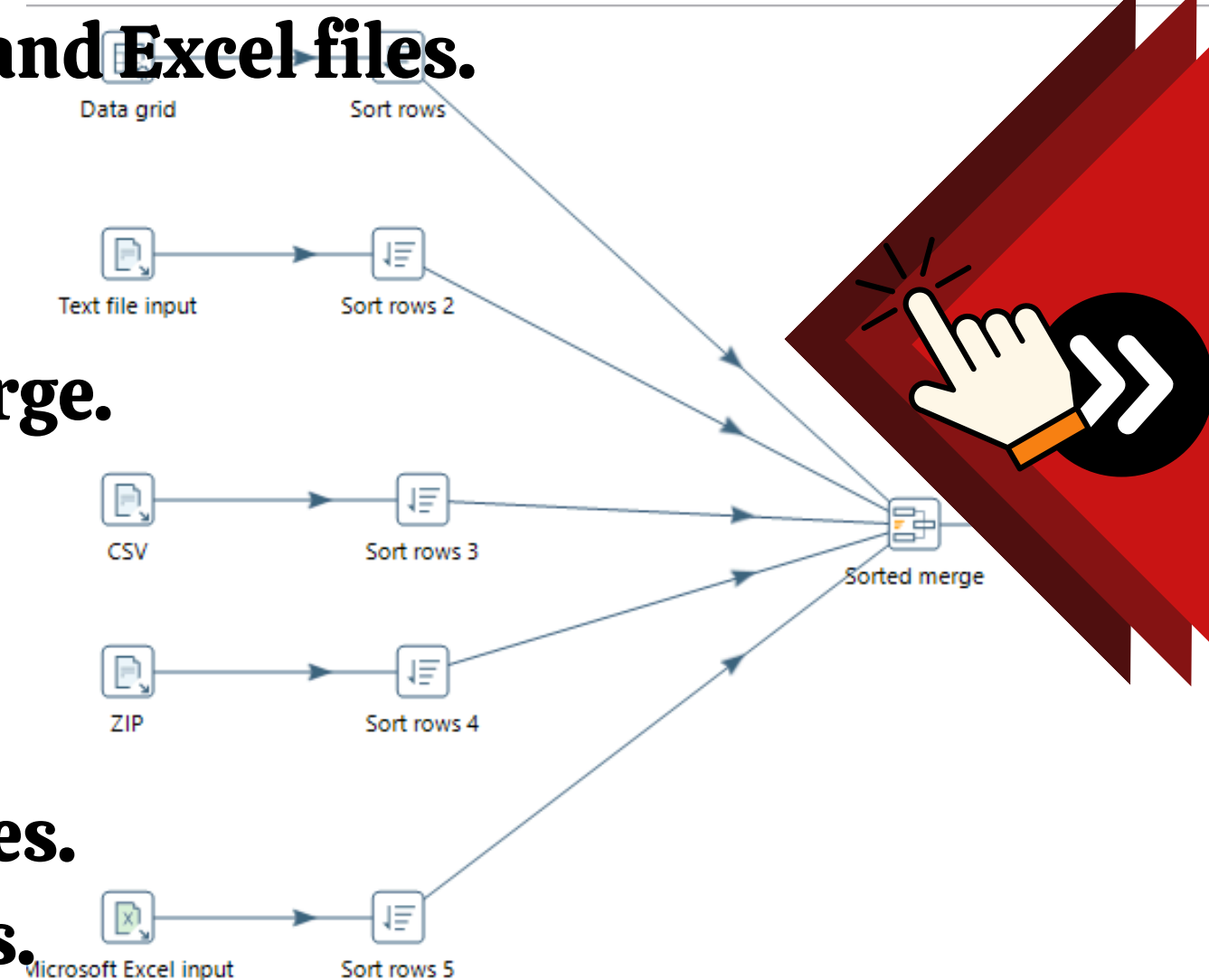
- Inputs include: Data Grid, Text File, CSV, ZIP, and Excel files.
- All data inputs are sorted using Sort rows.

2. Merge Step:

- All sorted inputs are merged using Sorted merge.
- Data Cleaning & Mapping:

3. Unique rows:

- Removes duplicate entries.
- Value mapper: Maps certain values to new ones.
- Replace in string: Replaces specific substrings.



4.Enrichment & Matching:

- Another Excel Input is compared using Fuzzy match.
- Select values: Chooses specific fields.
- Replace in string 2: Further string operations.

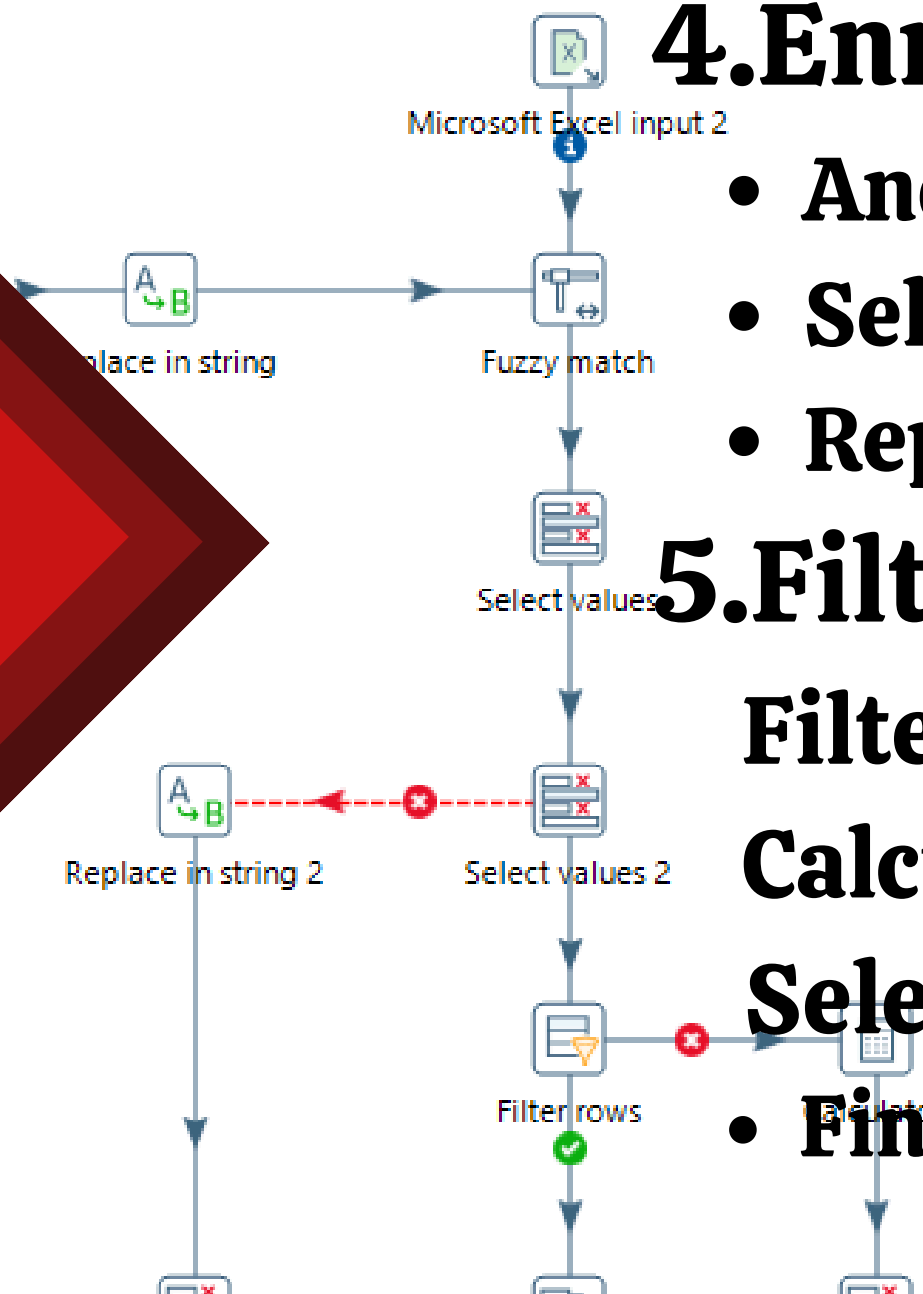
5.Filtering & Calculations (Bottom-right):

Filter rows: Applies conditions to filter data.

Calculator: Performs calculations on columns.

Select values 4 and Sorted merge 2:

- Final selections and merging for output.



**THANK
YOU**

