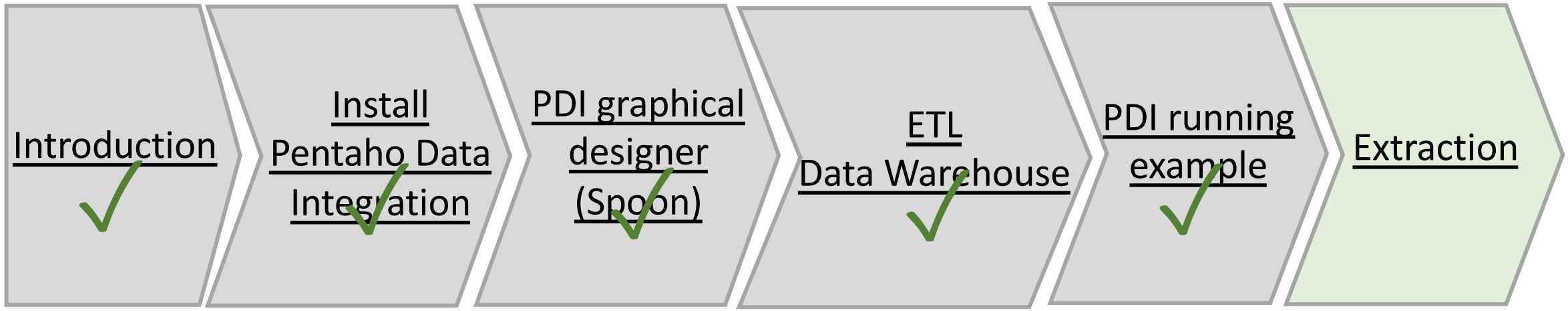
DATA ANALYTICS
(Data Warehouse)
Pentaho Data Integration

Luca Cinelli, PhD

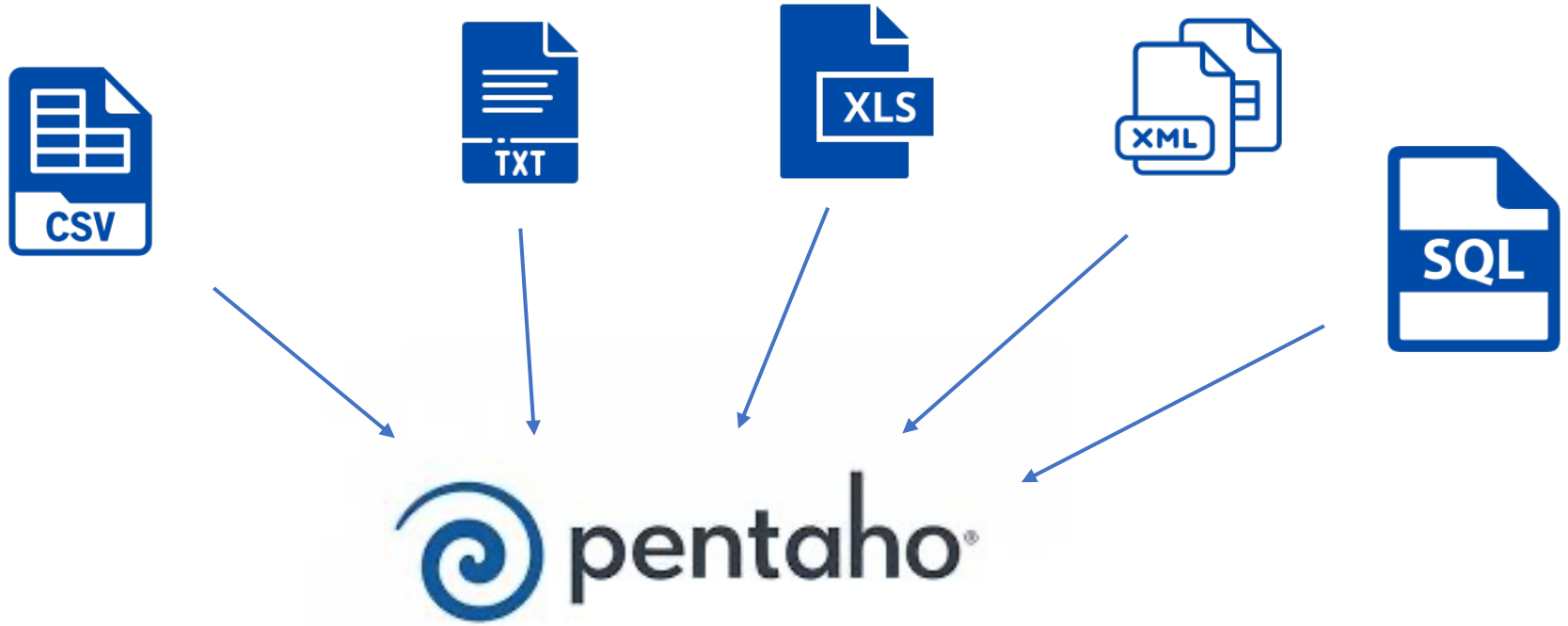
luca.cinelli@unical.it

Outline


Introduction
Install Pentaho Data Integration
PDI graphical designer (Spoon)
ETL Data Warehouse
PDI running example
Extraction

Extraction

Data sources



Example: Datasets

1. Sales Data
2. Customer Data
3. Product Data

Example: ETL problem scenario

1. Extraction

Customer data

CSV

TXT

Excel

Folder

Zip file

Product Data

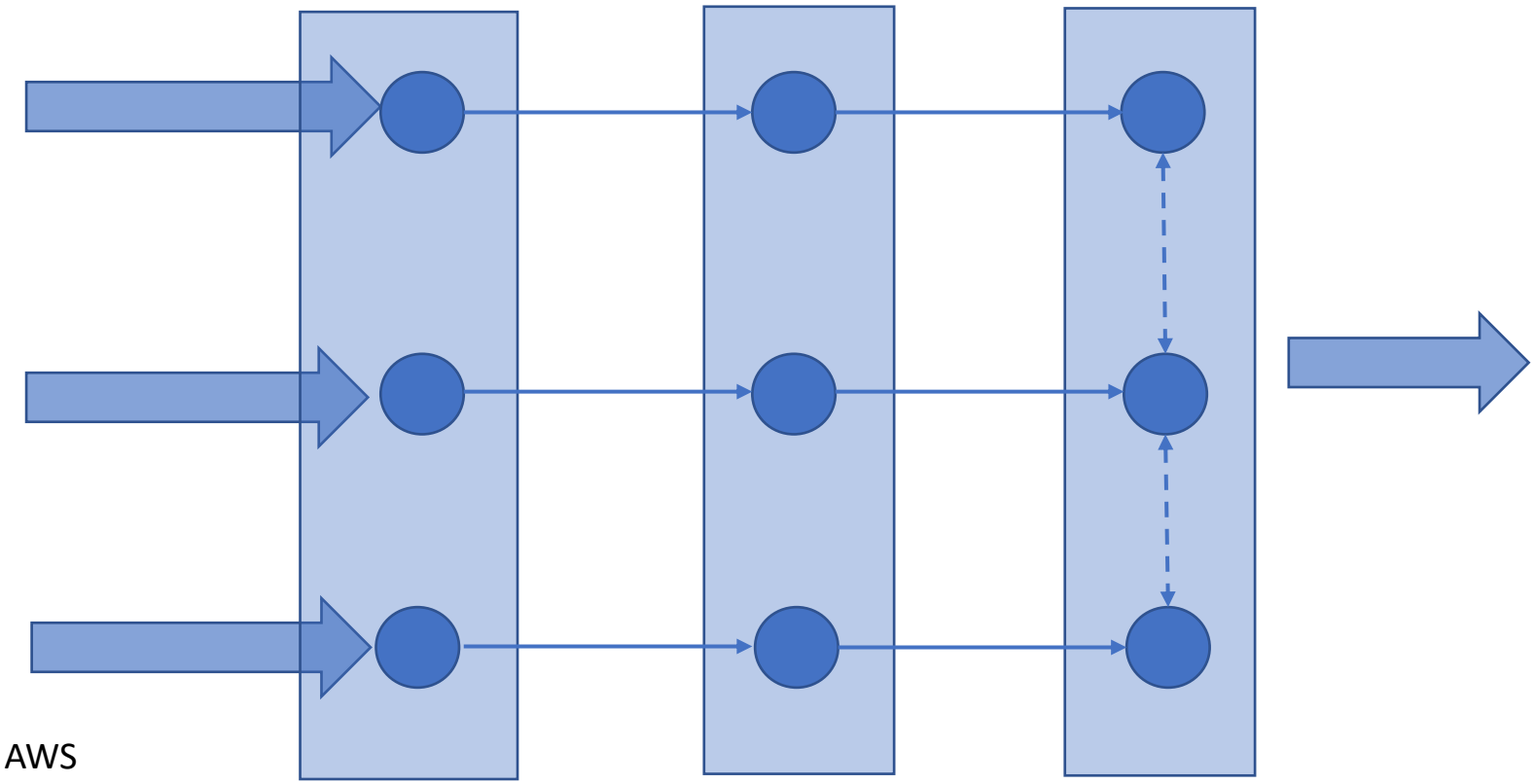
XML

JSON

Sales Data

SQL based DB

Cloud based Storage AWS



3. Cleaning and Validation

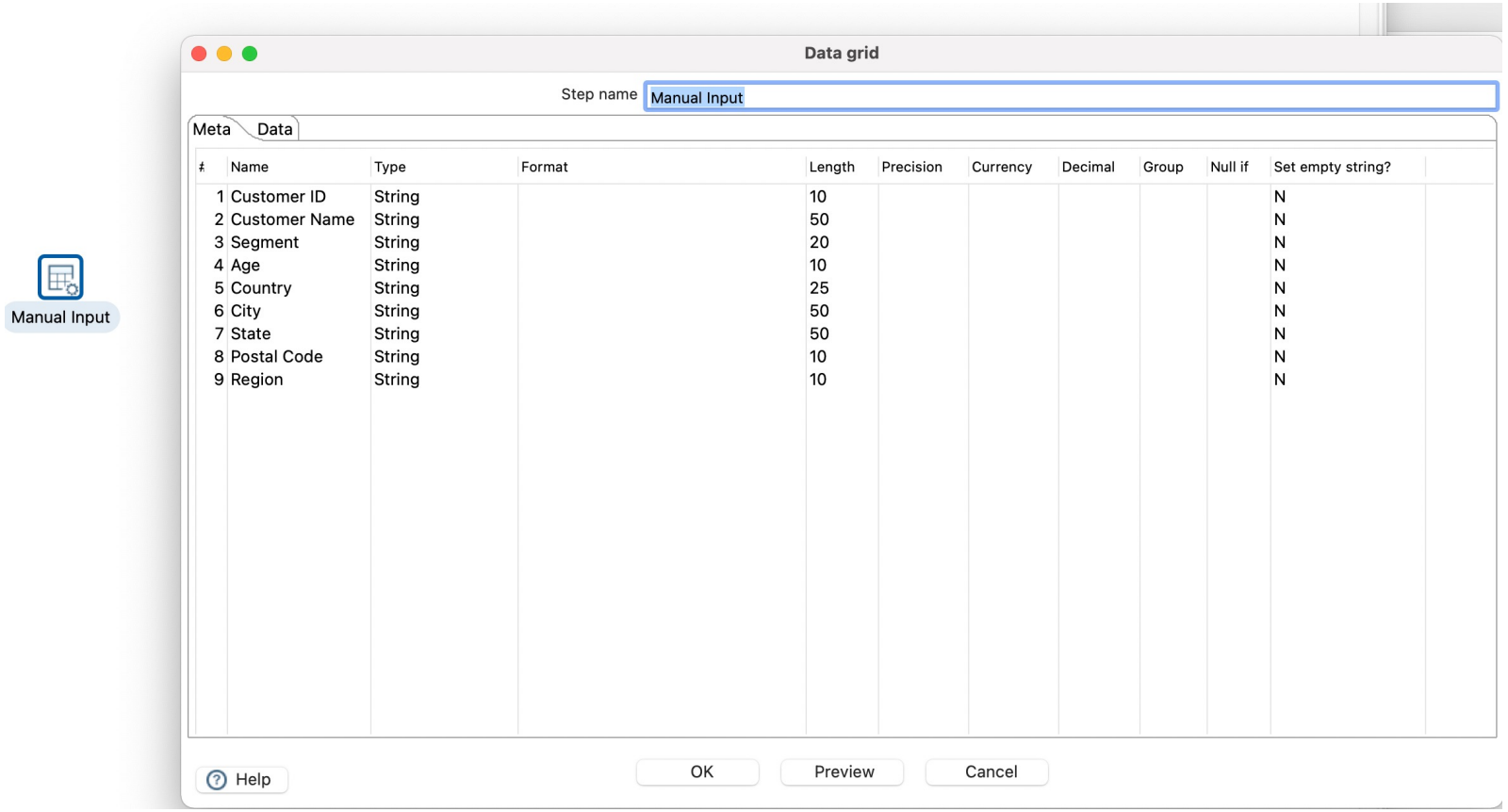
2. Merging

4. Transformation, aggregation and Joining

5. Loading to create a DataMart

Extraction from tabular format

- Manually entering data into PDI (ManualInput.ktr)



Extraction from tabular format

- Manually entering data into PDI
- Import data from text file (TxtInput.ktr, input: CustomerData_Central.txt)

Extraction from tabular format

- Manually entering data into PDI
- Import data from text file
- Import data from multiple CSV file (MultipleFiles.ktr, input: customer_data_multiple_files)

Extraction from tabular format

- Manually entering data into PDI
- Import data from text file
- Import data from multiple CSV file version 2 (MultipleFiles_withGetFileName.ktr, input: customer_data_multiple_files)

Extraction from tabular format

- Manually entering data into PDI
- Import data from text file
- Import data from multiple CSV file
- Import data from excel file (ExcelInput.ktr, input: CustomerData_East.xlsx)

Extraction from tabular format

- Manually entering data into PDI
- Import data from text file
- Import data from multiple CSV file
- Import data from excel file
- Extract data from zip file (ZipInput.ktr, input: CustomerData_South.zip)

Extraction from no-tabular format

- Extract data from XML file (XMLInput.ktr, input: ProductDataAsXML.xml)

Extraction from no-tabular format

- Extract data from XML file
- Extract data from JSON file (JSONInput.ktr, input: ProductDataasJSON.js)