Data Transformation

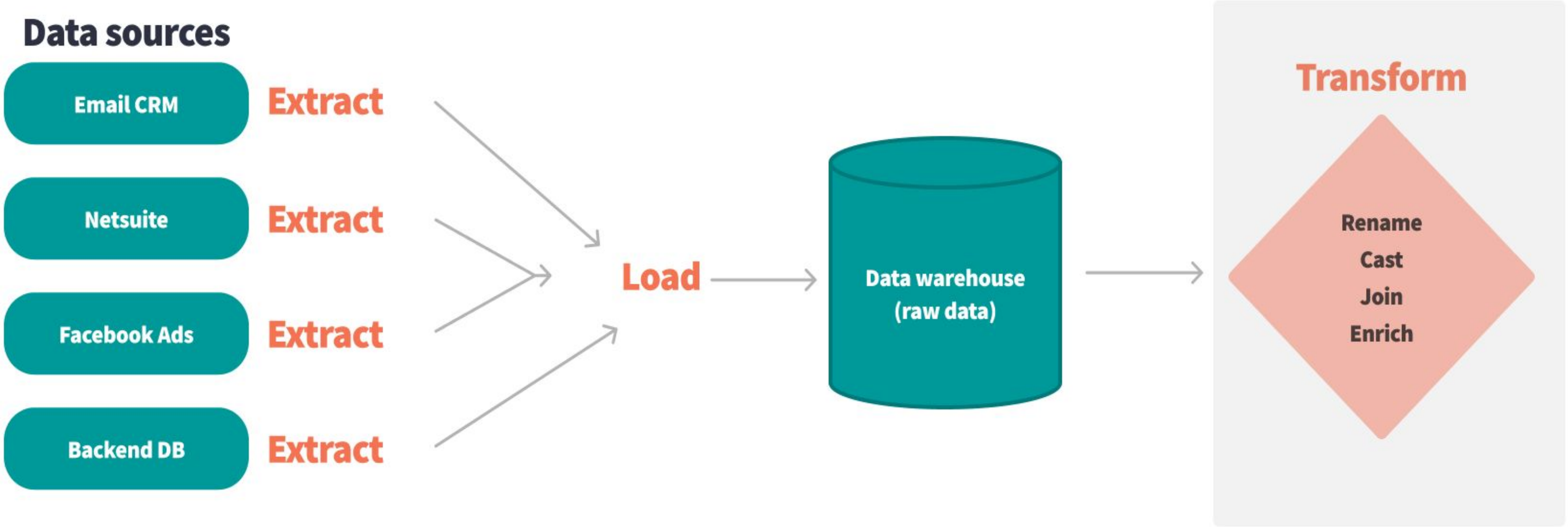
Data Transformation

- Data transformation is one of the techniques that we use in between data processing. This technique lets us convert the raw data into a required format so that the next procedures of data processing and data modelling can be performed efficiently.

- Technically, Data transformation changes the data structure, format and value and makes it clean and usable for the next processes. There can be two stages of data transformation processes because many organisations use data warehouses arranged in the ETL process, where data transformation is an in-between process.

- However, data transformation is an integral part of the entire data processing pipeline because it can be found right from data integration to the final stages of data wrangling.

ELT



Data sources

Email CRM

Netsuite

Facebook Ads

Backend DB

Extract

Extract

Extract

Extract

Load

Data warehouse (raw data)

Transform

Rename

Cast

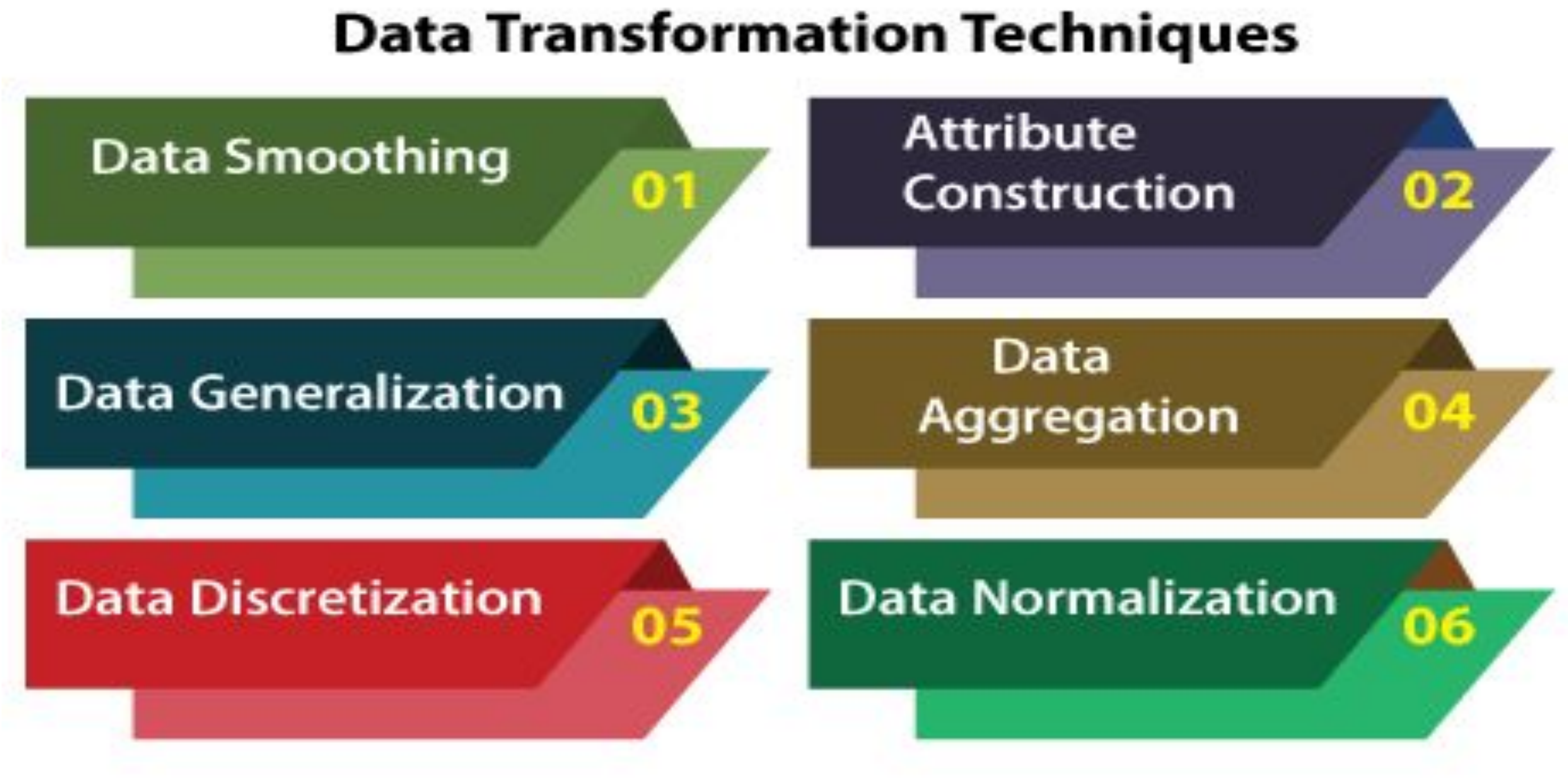
Join

Enrich

Type of data transformation

- **Constructive:** The process of adding, copying or replicating data.
- **Destructive:** deleting the records or field.
- **Aesthetic:** standardising the data to make it valuable.
- **Structural**: reorganising the data by moving, merging and renaming the columns.

Data Transformation Techniques



Data Transformation Techniques

Data Smoothing 01

Attribute Construction 02

Data Generalization 03

Data Aggregation 04

Data Discretization 05

Data Normalization 06

Various general data transformation techniques:

1. Data smoothing

- We can think of data smoothing as the process of removing noise from the data that can also involve some algorithms. This transformation makes the important data feature more visible and helps predict the pattern.

2. Attribute construction

- In this data transformation technique, we add new attributes to the data based on the already existing attributes. This new attribute smoothens the other processes by simplifying the way to get accurate data.

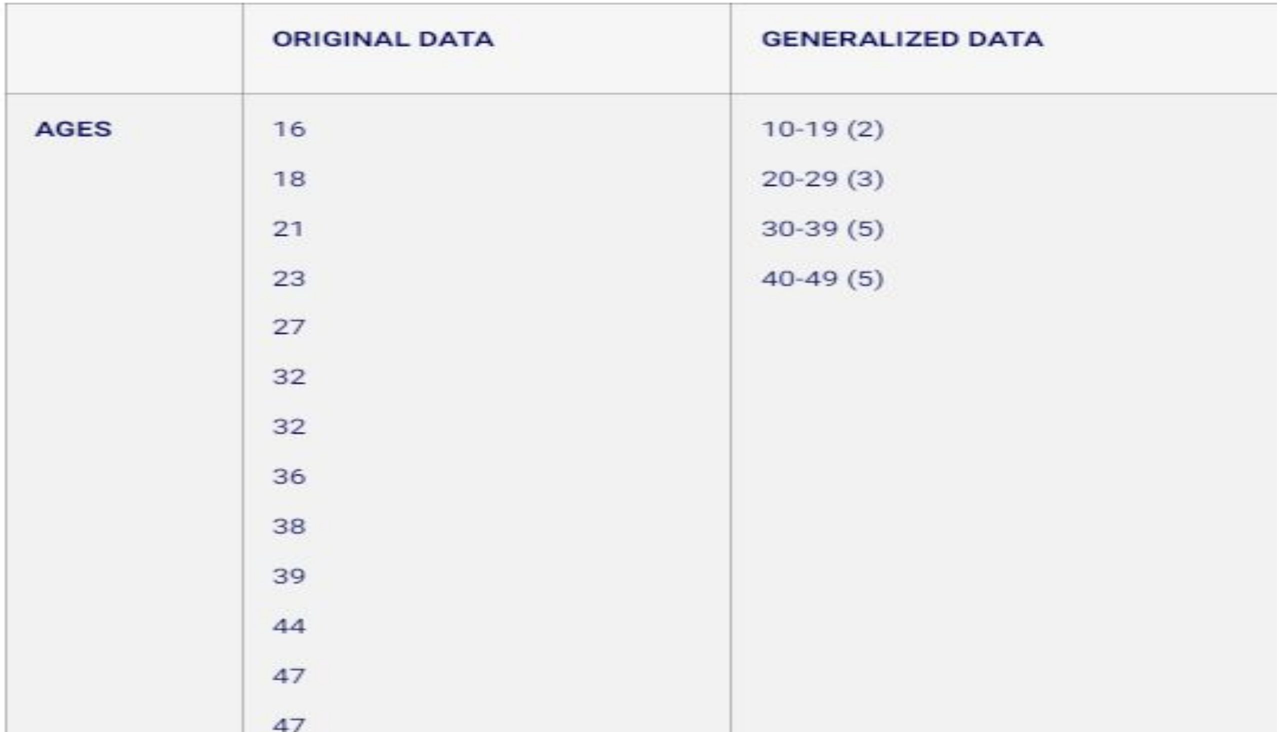
3. Data Generalization

- This data transformation technique depends on the concept of hierarchy to transform low-level data attributes into high-level data attributes.
- Through this transformation, we get a clear picture of the data, and this process can be divided into two different approaches:

  1.Data cube process (OLAP: Online analytical processing) approach.

  2. Attribute-oriented induction (AOI) approach.

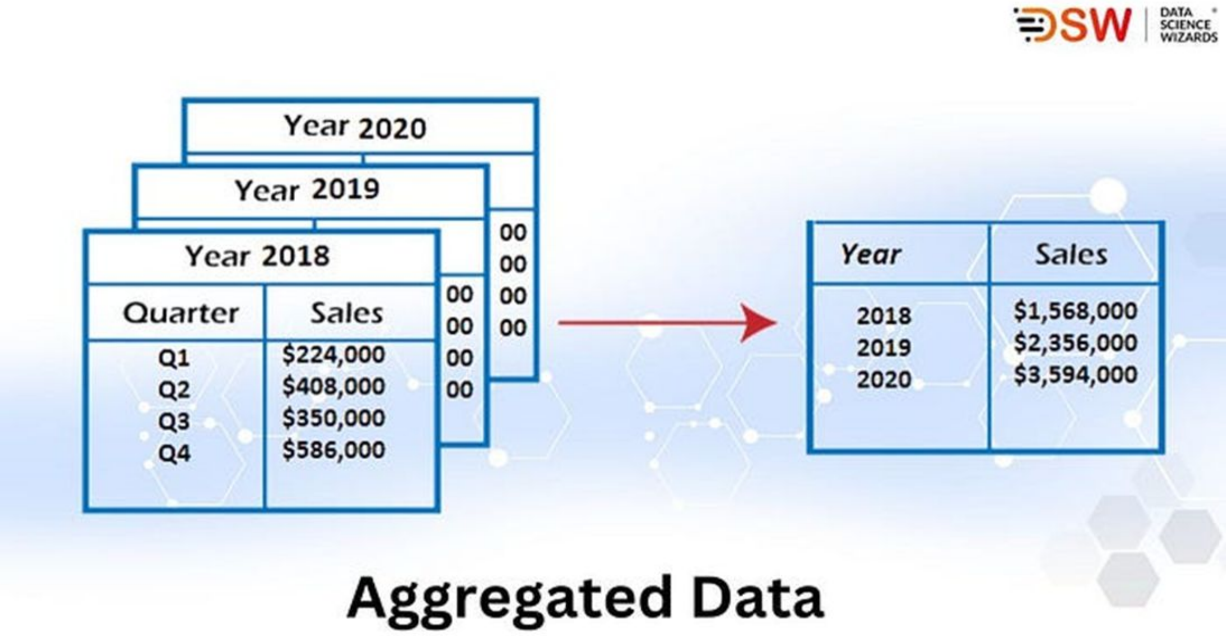
- A simple example of data generalization is to convert age data from numerical to categories.



	ORIGINAL DATA	GENERALIZED DATA
AGES	16	10-19 (2)
	18	20-29 (3)
	21	30-39 (5)
	23	40-49 (5)
	27	
	32	
	32	
	36	
	38	
	39	
	44	
	47	
	47	

4. Data aggregation

- Data aggregation or data collection is a technique of presenting data in a summary form. There is a huge chance of data coming from different sources, and integrating all the coming data into a description is data aggregation. This part of data processing is crucial because it depends on the quality and quantity of the data we use.
- An example of this process is making an annual sales report based on the given quarterly or monthly data.



Year 2018

Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
2018	$1,568,000
2019	$2,356,000
2020	$3,594,000

Aggregated Data

5. Data Discretization

- We can think of data discretization as the process of converting continuous data into a set of intervals so that small interval labels can substitute them. Just like normalization, this technique makes the data more interpretable. If a data analysis work uses continuous values, then discrete forms of values can be replaced by constant quality attributes.
- We can also call this technique the data reduction technique, as it transforms discrete data into a set of categorical data.
- We can classify this technique into supervised and unsupervised discretization. Where supervised data discretization uses class information, on the other hand, unsupervised data discretization uses the processing direction of the involved data process.
- For example, we can classify people in numeric age groups such as(0–5,5–10…) or in classes like kid, youth or adult.

6. Data Normalization

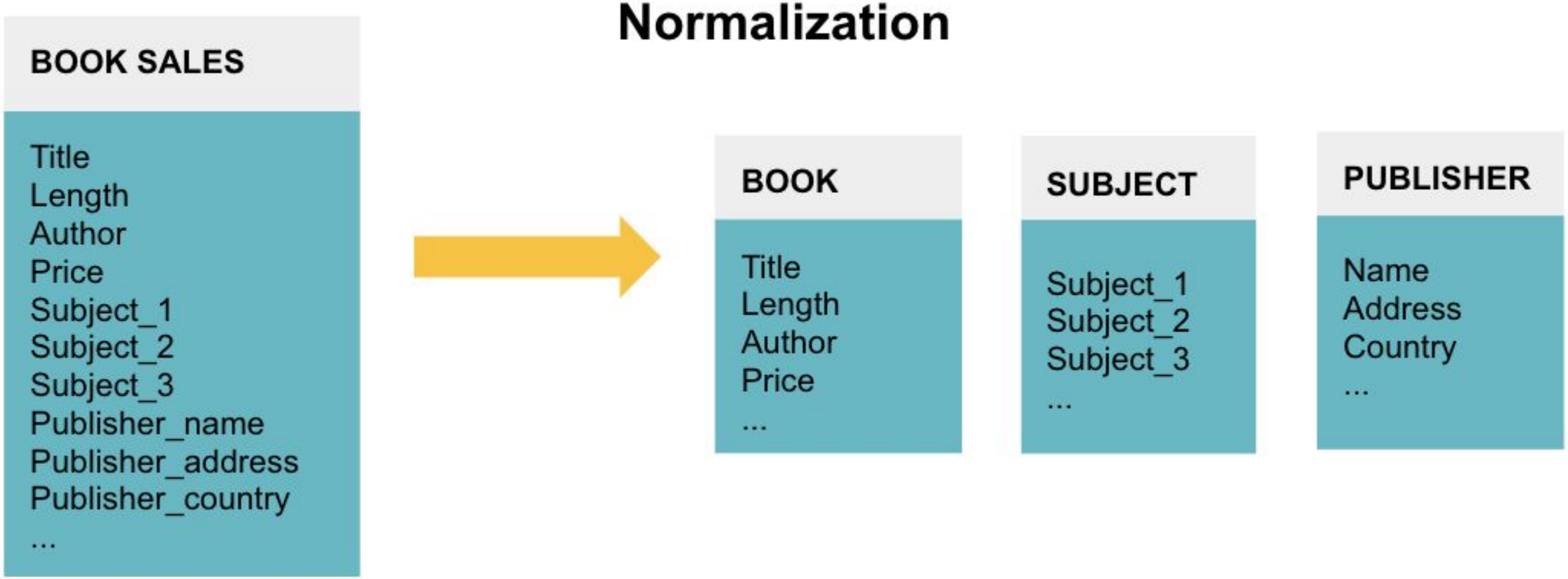
- Data normalization is another technique of data transformation that involves scaling the data in a smaller range, such as scaling data between 0 to 1 or -1 to 1.

- The main objective of data normalization is to eliminate data redundancy and improve data consistency and accuracy. By organizing data into smaller, more manageable ranges, it becomes easier to update and maintain the data.

- Normalization also helps to reduce the likelihood of data inconsistencies, which can arise when the same information is stored in multiple locations.

- There are various types of data normalization techniques we use, such as
  - Min-Max normalization,
  - Z-score normalization,
  - Decimal Scaling etc.,

Normalization



BOOK SALES

Title
Length
Author
Price
Subject_1
Subject_2
Subject_3
Publisher_name
Publisher_address
Publisher_country
...

BOOK

Title
Length
Author
Price
...

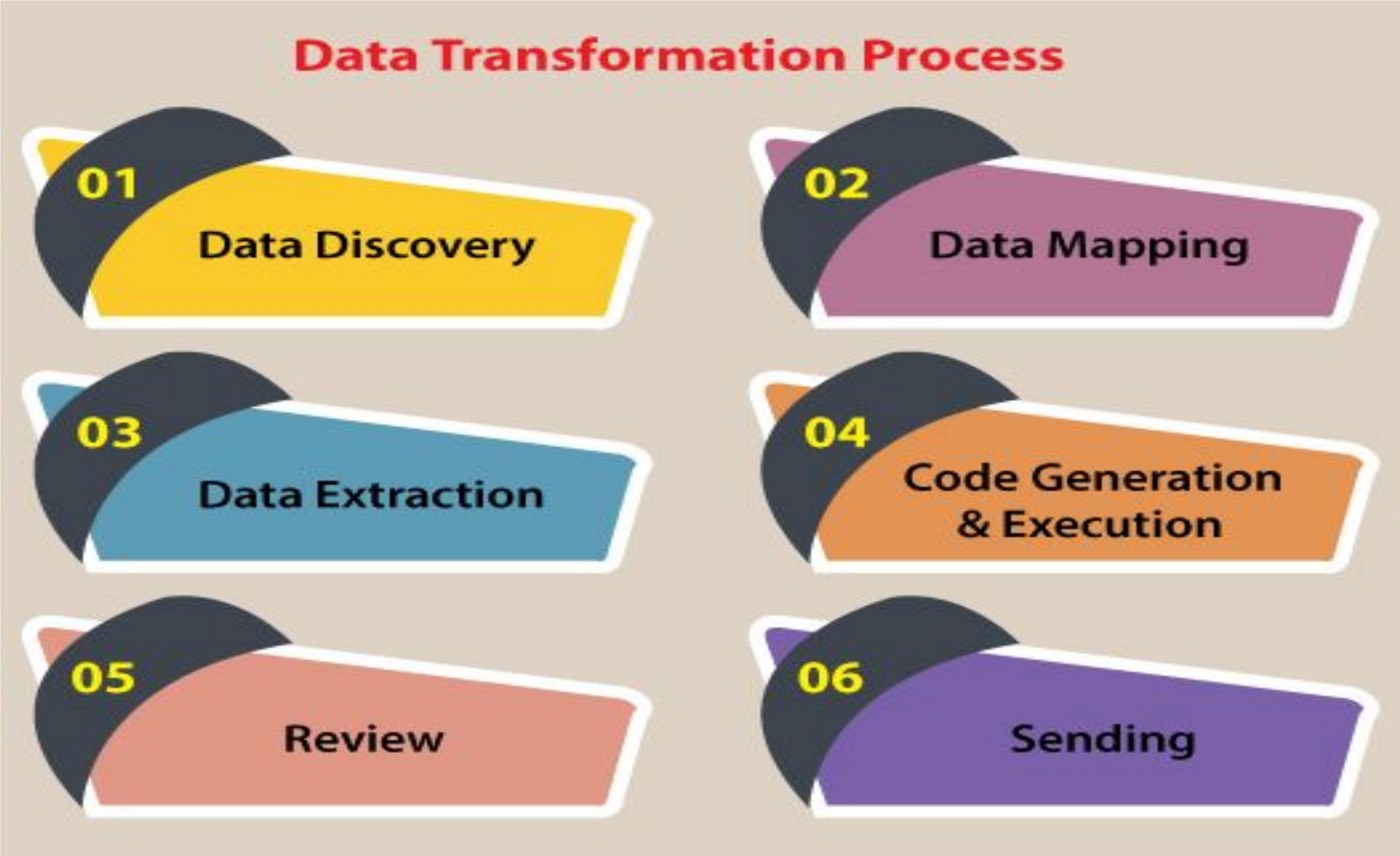
SUBJECT

Subject_1
Subject_2
Subject_3
...

PUBLISHER

Name
Address
Country
...

Data transformation Process



Data Transformation Process

01 Data Discovery

02 Data Mapping

03 Data Extraction

04 Code Generation & Execution

05 Review

06 Sending

Data transformation process

1.Data Discovery: Before applying ETL in any organisation, it is important to know and understand the data source. Usually, data profiling tools we use to do so. With the help of this step, one gets to know what ahead is required to get data in the required format.

2. Data Mapping: This step of the process is performed to determine how any field is mapped, modified, filtered, joined or aggregated.

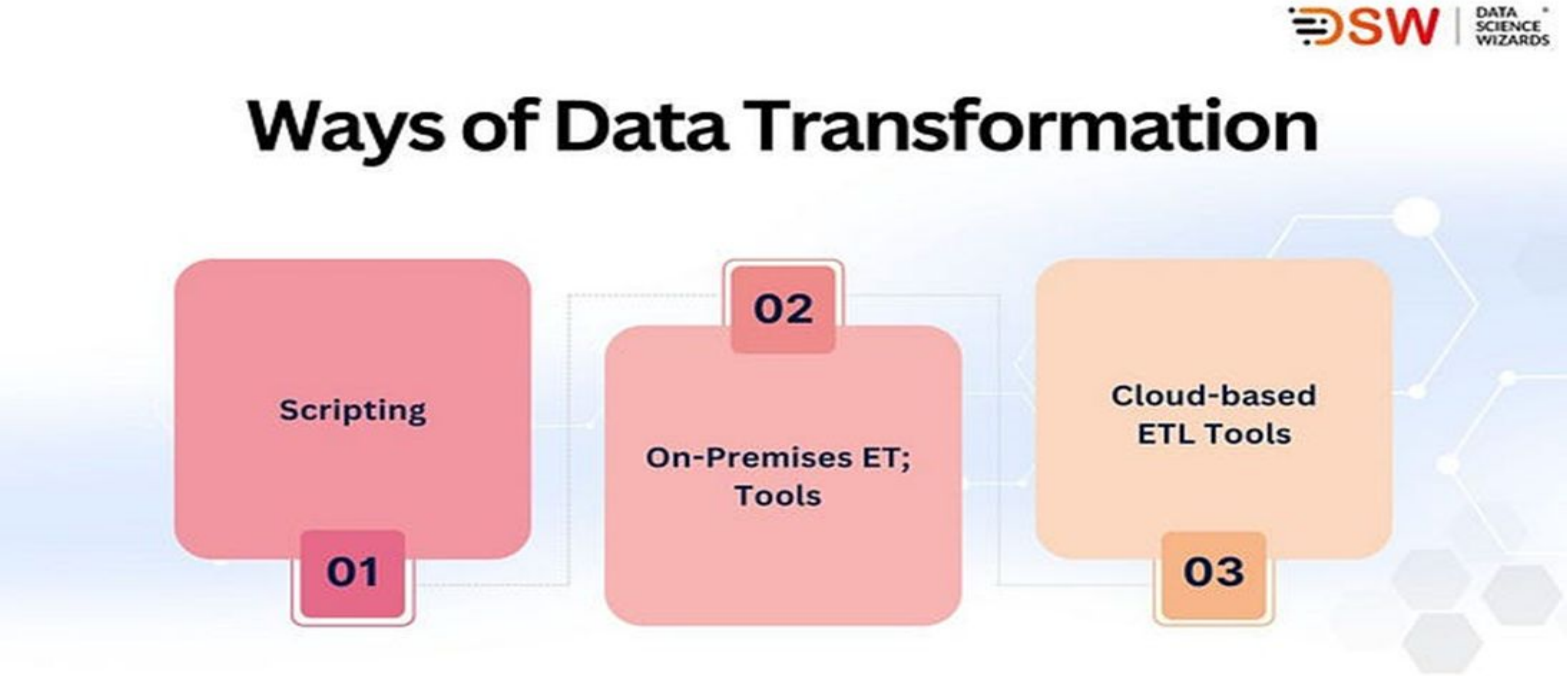
3. Data Extraction: This step involves the extraction of data from its original source. Examples of sources can be databases or streaming sources such as sales log files from web applications.

4. Code Execution: The foremost step that involves actual data transformation is code generation and execution. Here code is generated to transform data in the required format.

5. Review: This step ensures that the code executed on data is transforming the data accurately.

6. Sending: Till here, the data is transformed, but the last thing which remains to complete is to send the data to its target destination. Here the destination can be a relational database or warehouse that handles both unstructured and structured data.

Ways of data transformation



Ways of Data Transformation

01 Scripting

02 On-Premises ET; Tools

03 Cloud-based ETL Tools

Three general ways of data transformation are as follows.

- **By scripting:** this way, we write codes in python, SQL or other languages to query and transform the data. However, using python and SQL has its own benefits because they help automate various tasks, as well as they require less coding to perform than a traditional programming language requires.
- **Using ETL tools:** these tools are designed to make the data extraction and transformation work easier than scripting on-premises. These tools can be hosted on the organisation’s server and can save a lot of time. But these tools often require expertise in using tools and significant infrastructure costs.
- **Cloud-based ETL tools:** these tools are based on a relevantly new technology where it can be hosted in the cloud. These tools are designed to provide the easiest way to extract and transform data even a non-technical person can perform operations on it. Using these tools, we can store the data in any cloud source, and when required easily, we can load the data in a data warehouse. With the help of such ETL tools, we can schedule data pulling while monitoring the usage.

Advantages of Data Transformation

- **Improved data quality:** we all know of the risks and costs associated with low-quality data, so data transformation helps organisations eliminate data quality issues.
- **Faster Queries:** it becomes easy and quick to extract transformed data from its storage or location.
- **Efficient Data Management:** as data is non-stop generating in different sources, it becomes challenging to build and understand metadata. Data transformation helps in refining the metadata, and this is how data management becomes more efficient.
- **Better organisation:** Data transformation makes the process easy because it is more interpretable for both humans and computers.
- **More use of data:** In the process of generating and collecting data continuously, a lot of data gets unanalysed. Data transformation lets the get the most out of data by standardising it and making it more usable.