SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution
Coimbatore-35

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

UNIT I – INTRODUCTION TO DATA SCIENCE

CLEANING INTEGRATING AND TRANSFORMING DATA

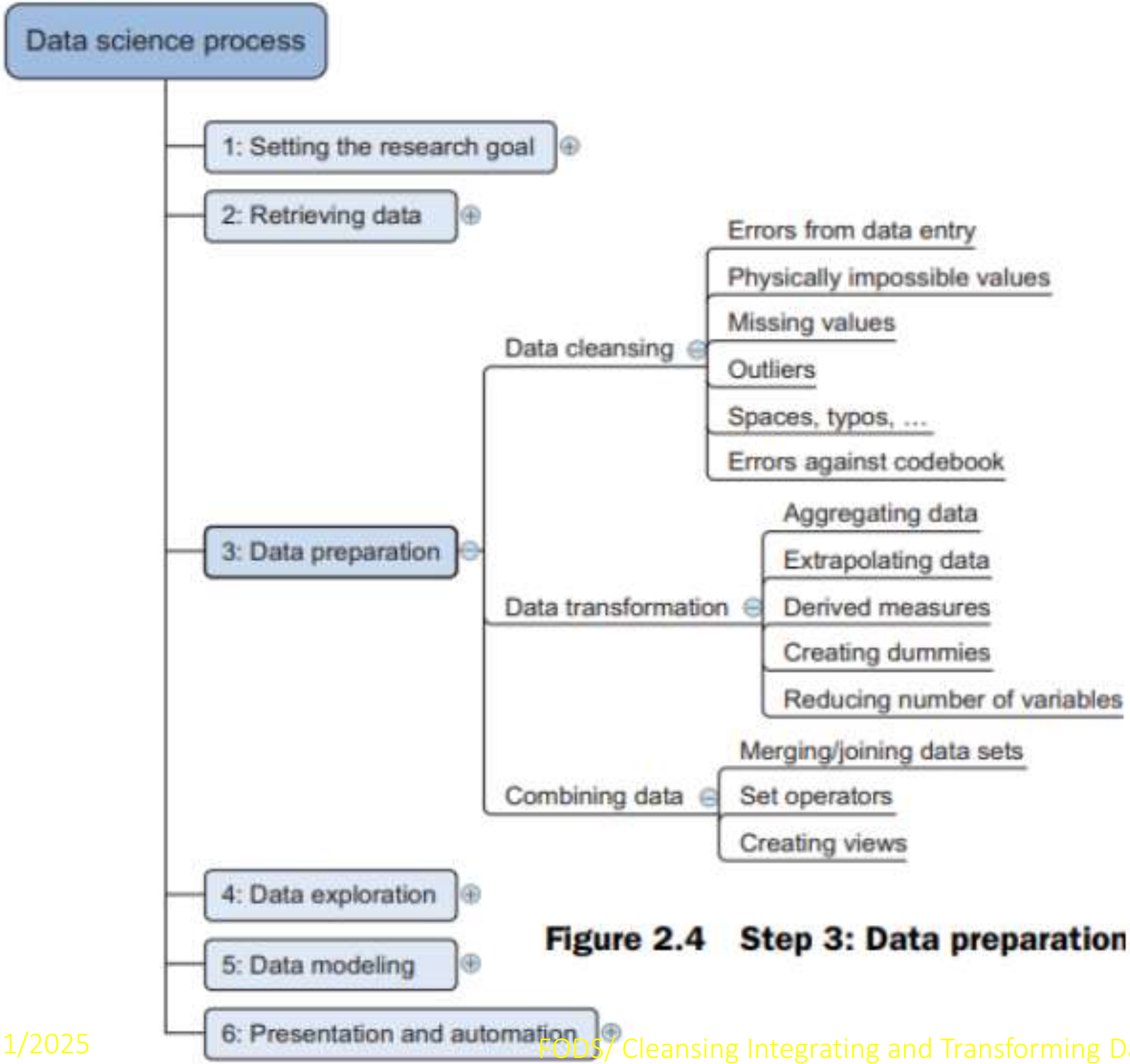
EMPATHY:

- ❖ Cleansing data
- ❖ Correct errors as early as possible, combining data from different data sources
- ❖ Transforming data
- ❖ Exploratory data analysis
- ❖ Build the models:
- ❖ Model and variable selection
- ❖ Model execution
- ❖ Model diagnostics and
- ❖ model comparison

Cleansing, Integrating, and Transforming data

DEFINE:

- The data received from the data retrieval phase is likely to be “a diamond in the rough.”
- Your task now is to sanitise and prepare it for use in the modelling and reporting phase.
- Doing so is tremendously important because your models will perform better and you’ll lose less time trying to fix strange output.
- It can’t be mentioned nearly enough times: garbage in equals garbage out.
- Your model needs the data in a specific format, so data transformation will always come into play.
- It’s a good habit to correct data errors as early on in the process as possible.



Data science process

- 1: Setting the research goal
- 2: Retrieving data
- 3: Data preparation
  - Data cleansing
    - Errors from data entry
    - Physically impossible values
    - Missing values
    - Outliers
    - Spaces, typos, ...
    - Errors against codebook
  - Data transformation
    - Aggregating data
    - Extrapolating data
    - Derived measures
    - Creating dummies
    - Reducing number of variables
  - Combining data
    - Merging/joining data sets
    - Set operators
    - Creating views
- 4: Data exploration
- 5: Data modeling
- 6: Presentation and automation

Figure 2.4 Step 3: Data preparation

Cleansing data: -

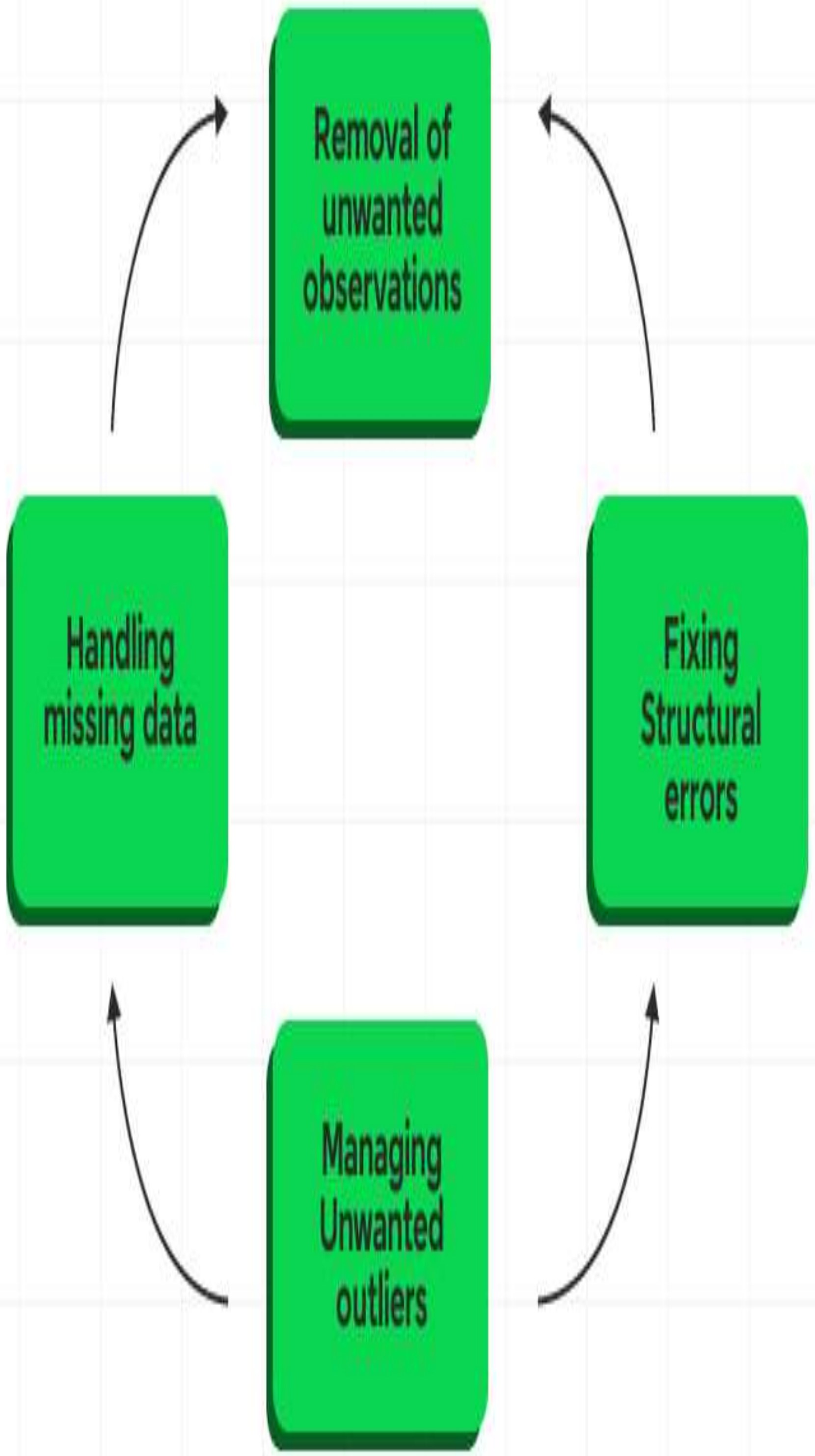
IDEATE:

- Data cleansing is a sub process of the data science process that focuses on removing errors in your data so your data becomes a true and consistent representation of the processes it originates from.

- Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in datasets.

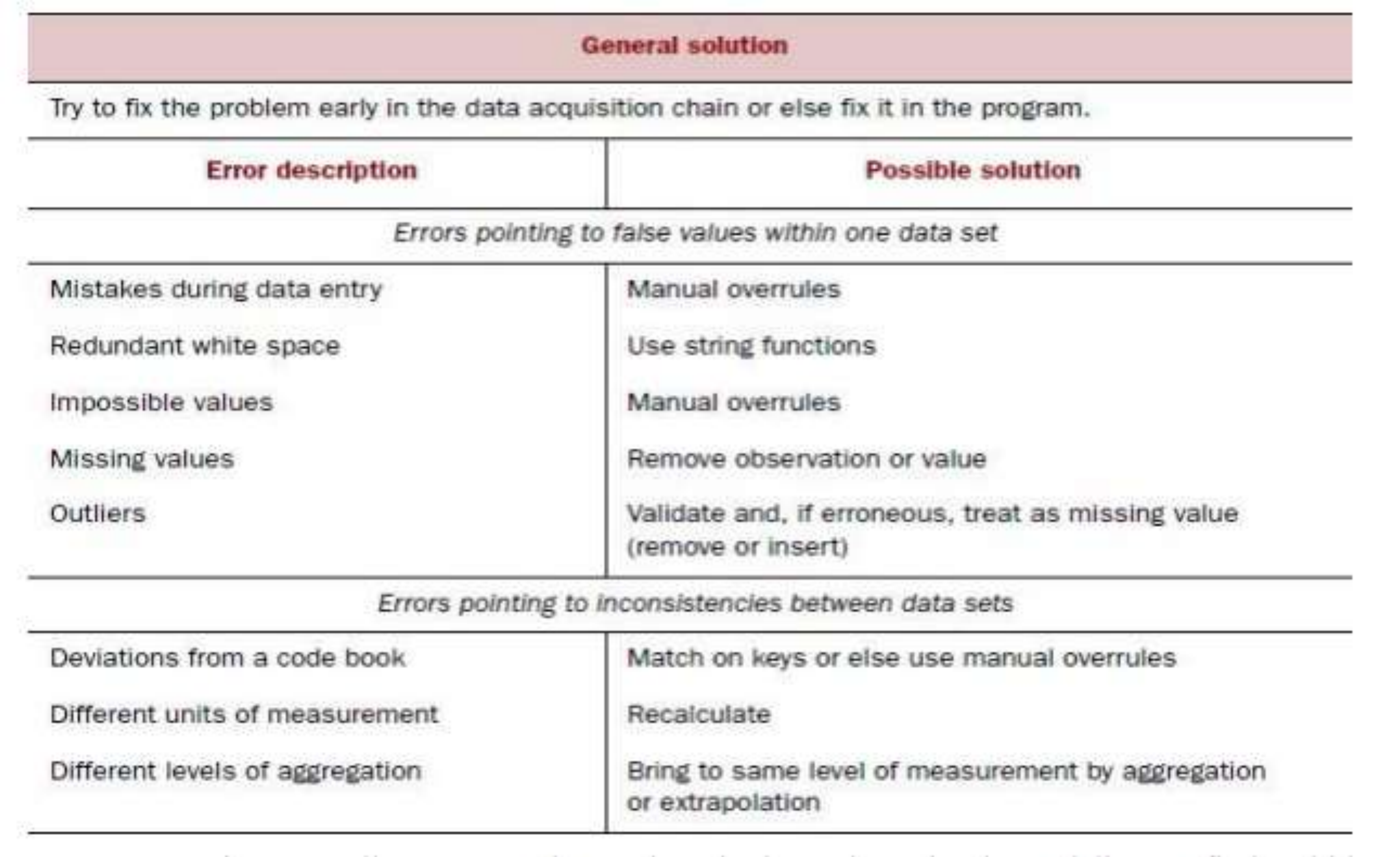
It involves:

- Removing duplicate or irrelevant observations
- Fixing structural errors
- Handling missing data
- Filtering out outliers
- Standardising data formats
- Correcting typos or formatting issues
- This process is important for ensuring data quality and reliability in analysis and decision-making.



TESTING:

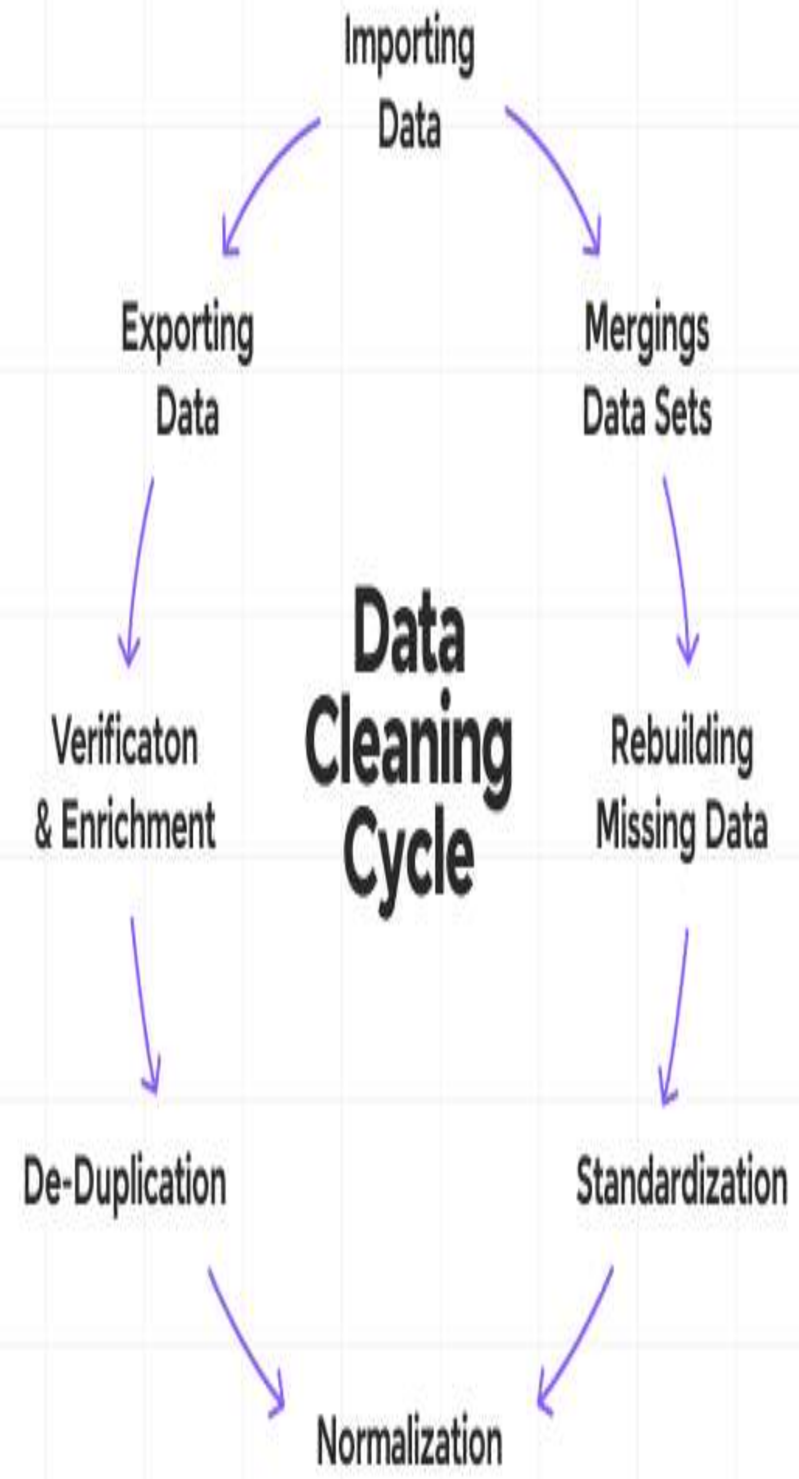
Overview of common errors

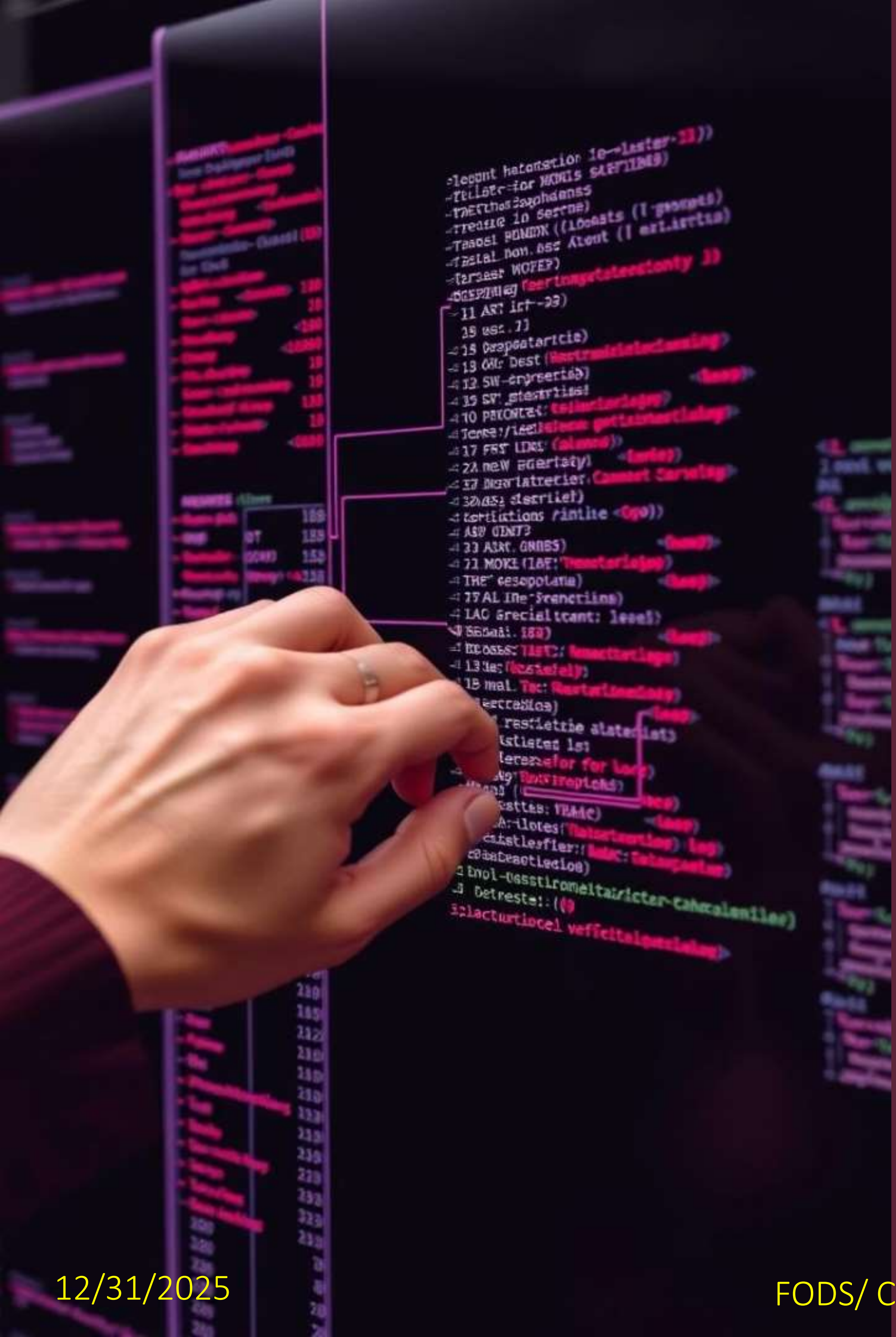


<table>
<tr><th colspan="2">General solution</th></tr>
<tr><td colspan="2">Try to fix the problem early in the data acquisition chain or else fix it in the program.</td></tr>
<tr><th>Error description</th><th>Possible solution</th></tr>
<tr><td colspan="2"><i>Errors pointing to false values within one data set</i></td></tr>
<tr><td>Mistakes during data entry</td><td>Manual overrules</td></tr>
<tr><td>Redundant white space</td><td>Use string functions</td></tr>
<tr><td>Impossible values</td><td>Manual overrules</td></tr>
<tr><td>Missing values</td><td>Remove observation or value</td></tr>
<tr><td>Outliers</td><td>Validate and, if erroneous, treat as missing value (remove or insert)</td></tr>
<tr><td colspan="2"><i>Errors pointing to inconsistencies between data sets</i></td></tr>
<tr><td>Deviations from a code book</td><td>Match on keys or else use manual overrules</td></tr>
<tr><td>Different units of measurement</td><td>Recalculate</td></tr>
<tr><td>Different levels of aggregation</td><td>Bring to same level of measurement by aggregation or extrapolation</td></tr>
</table>

Step-by-Step Data Cleaning Process

Step 1: Understand the Data
Step 2: Make a Copy of the Raw Data
Step 3: Perform Initial Data Exploration
Step 4: Check Data Types and Structures
Step 5: Handle Missing Data
Step 6: Remove Duplicates
Step 7: Handle Outliers
Identify and address extreme values:
Step 8: Standardize and Normalize
Step 9: Correct Invalid Values
Step 10: Handle Structural Errors
Step 11: Validate and Cross-Check
Step 12: Handle Special Cases
Step 13: Document the Cleaning Process
Step 14: Create Automated Cleaning Scripts
Step 15: Perform a Final Review





Data Integration

PROTOTYPING:

What is Data Integration?

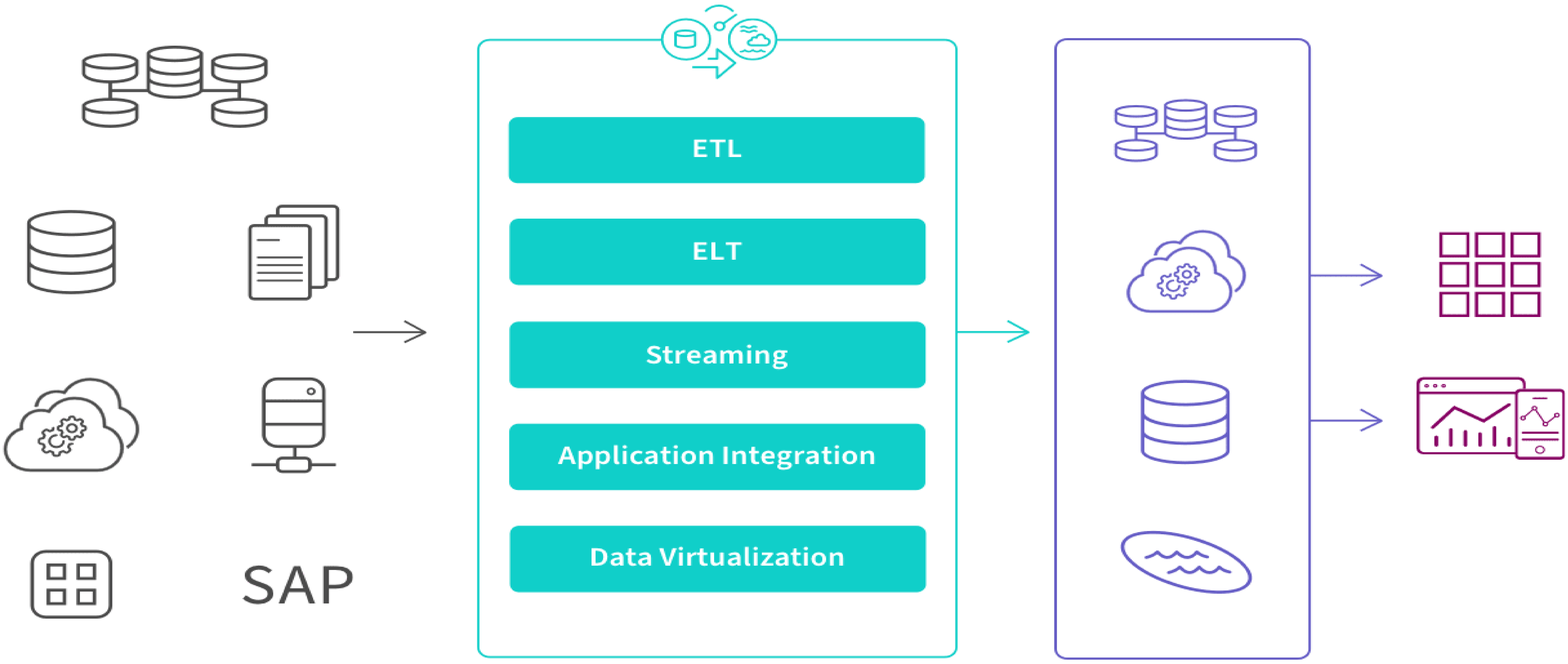
- ❖ Data integration refers to the process of bringing together data from multiple sources across an organization to provide a complete, accurate, and up-to-date dataset for BI, data analysis and other applications and business processes.

- ❖ It includes data replication, ingestion and transformation to combine different types of data into standardized formats to be stored in a target repository such as a data warehouse, data lake or data lakehouse.

Five Approaches

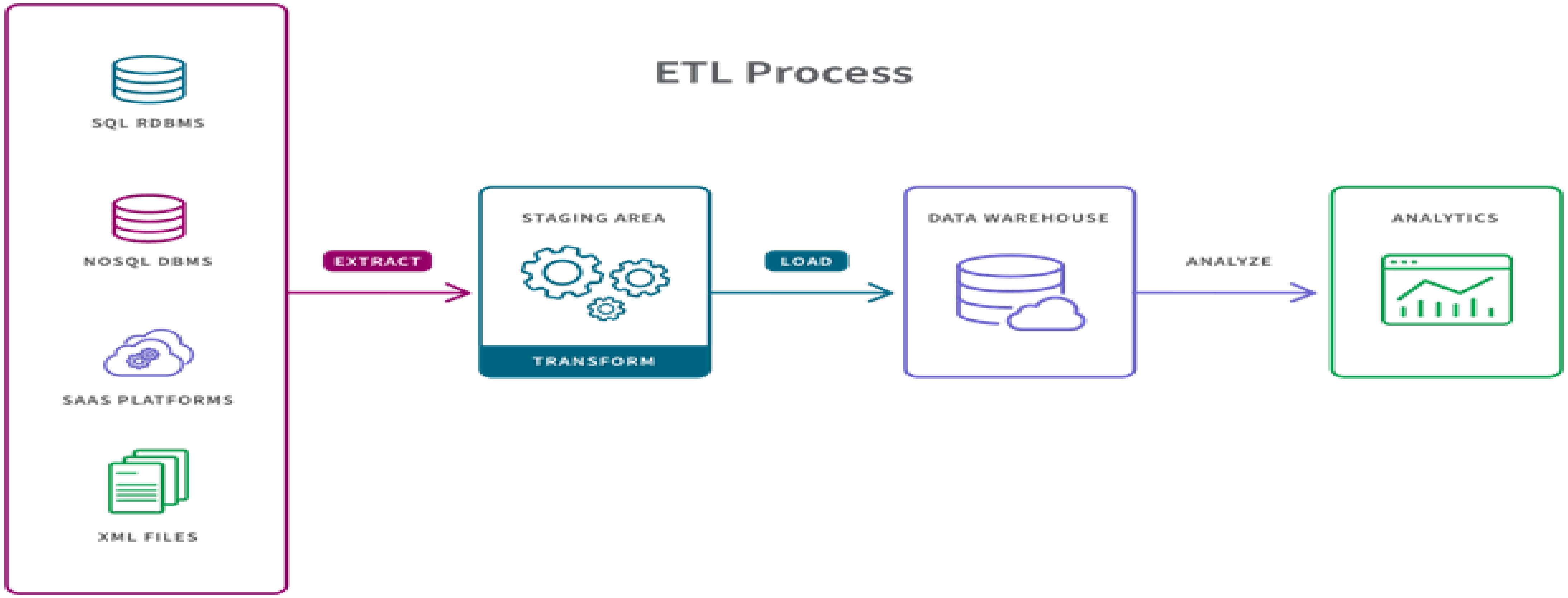
- There are five different approaches, or patterns, for executing data integration: ETL, ELT, streaming, application integration (API), and data virtualisation.

- To implement these processes, data engineers, architects and developers can either manually code an architecture using SQL or, more often, they set up and manage a data integration tool, which streamlines development and automates the system.

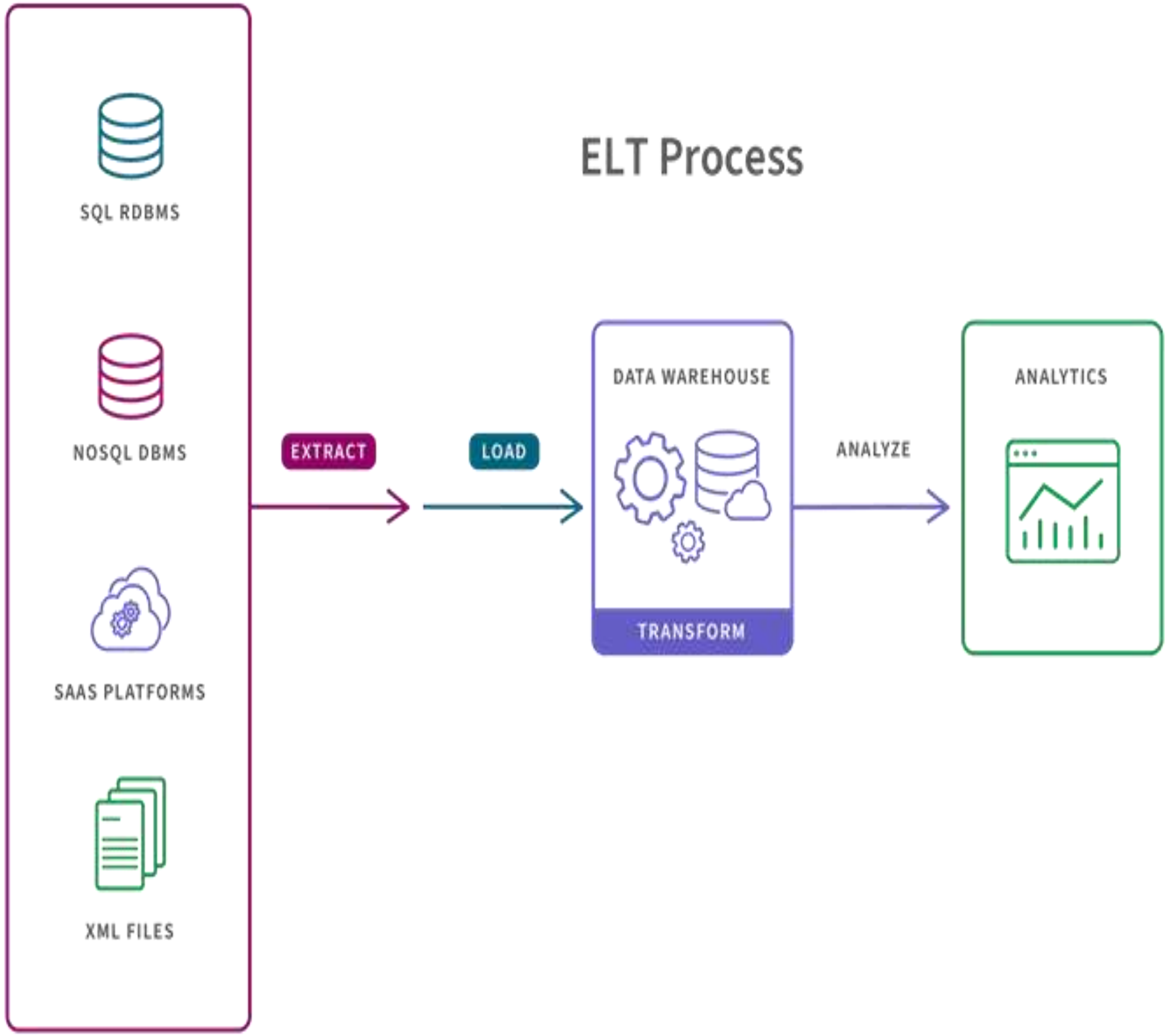

ETL
ELT
Streaming
Application Integration
Data Virtualization
SAP

What is an ETL Pipeline?

- An ETL pipeline is a set of processes to extract data from one system, transform it, and load it into a target repository.
- ETL is an acronym for "Extract, Transform, and Load" and describes the three stages of the process.

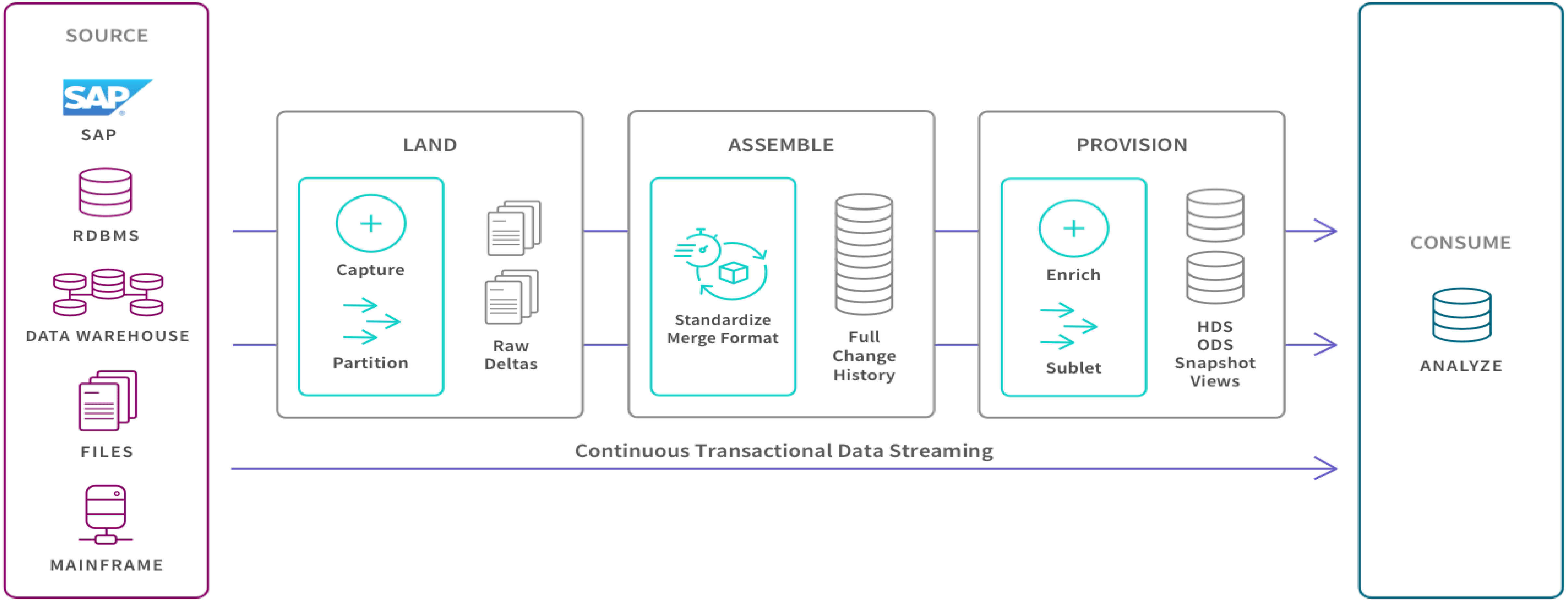


- The ETL and ELT acronyms both describe processes of cleaning, enriching, and transforming data from a variety of sources before integrating it for use in data analytics, business intelligence and data science.



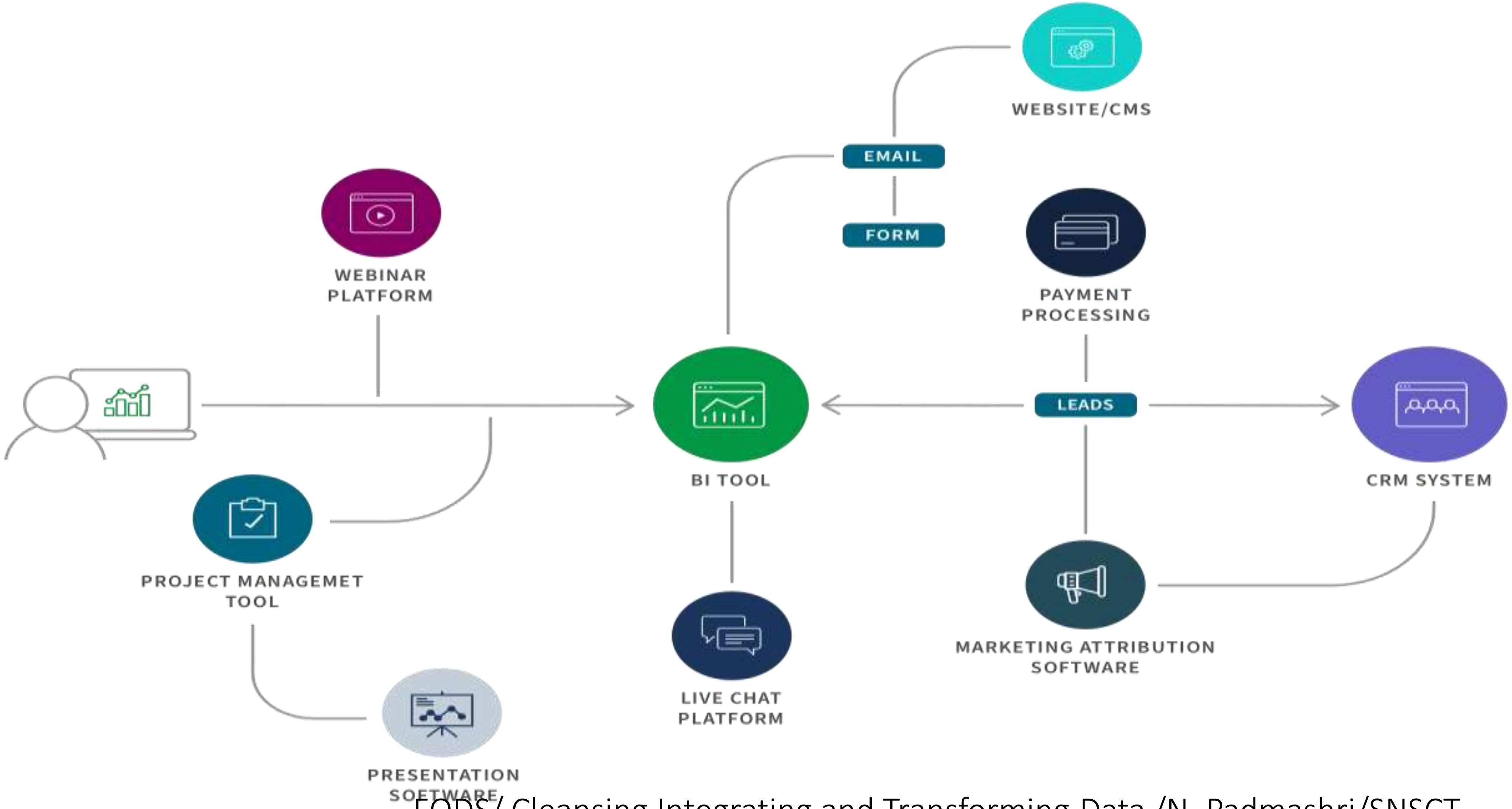
Data Streaming

- Instead of loading data into a new repository in batches, streaming data integration moves data continuously in real-time from source to target.
- Modern data integration (DI) platforms can deliver analytics-ready data into streaming and cloud platforms, data warehouses, and data lakes.



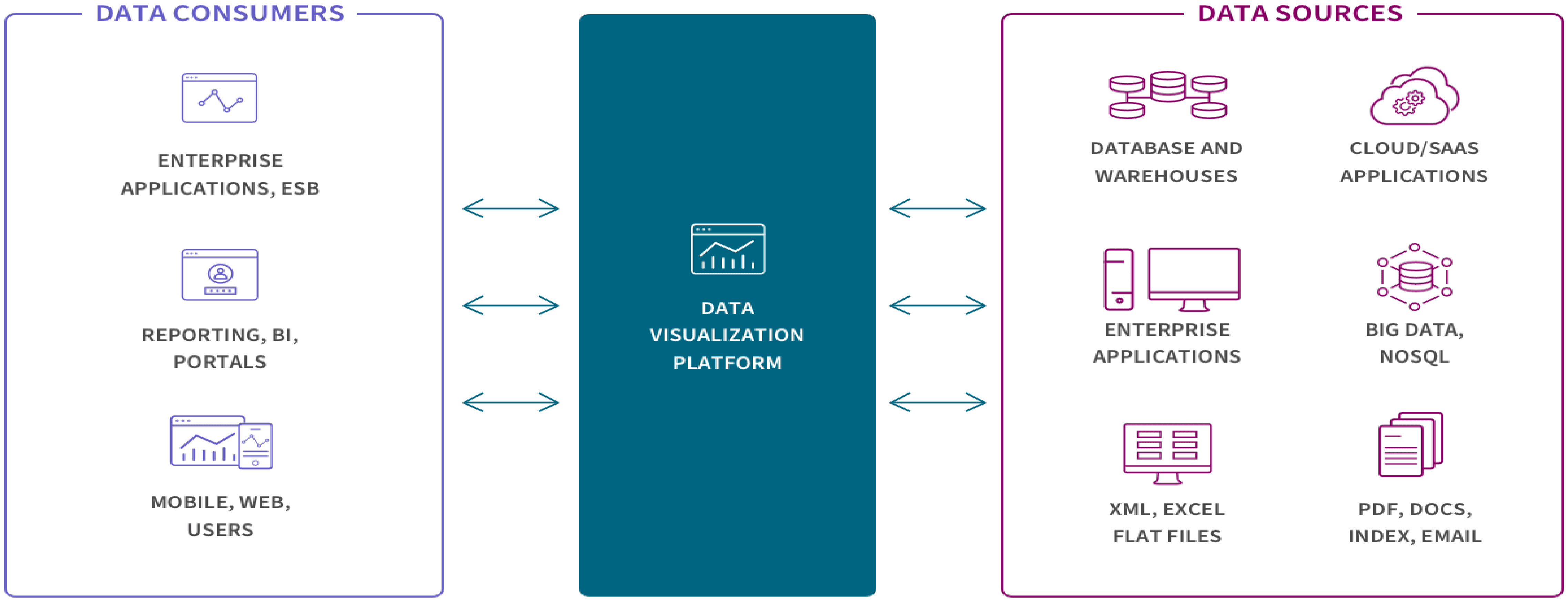
Application Integration

- Application integration (API) allows separate applications to work together by moving and syncing data between them.
- The most typical use case is to support operational needs such as ensuring that your HR system has the same data as your finance system.



Data Virtualization

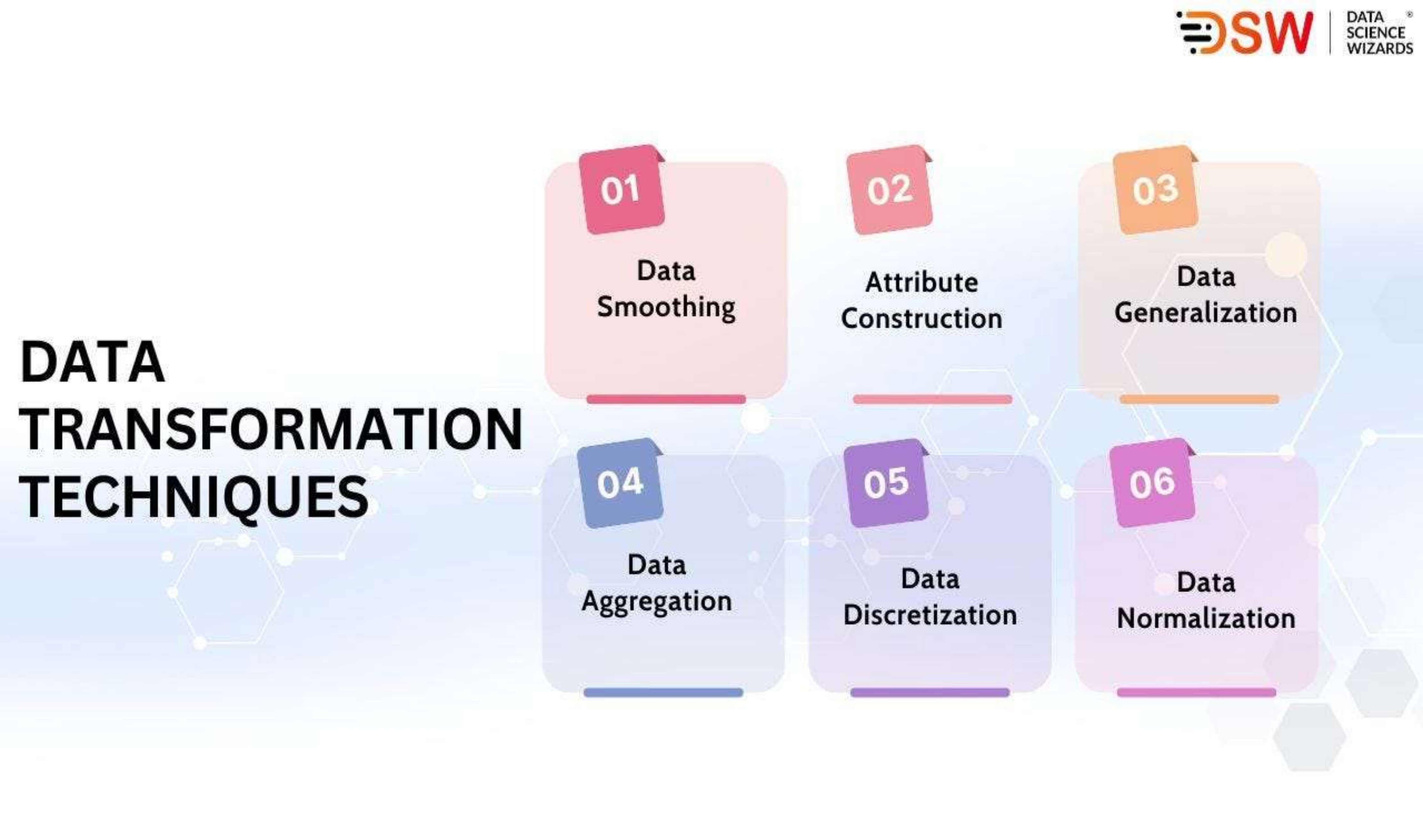
- Like streaming, data virtualisation also delivers data in real time, but only when it is requested by a user or application.
- Still, this can create a unified view of data and make data available on demand by virtually combining data from different systems.
- Virtualisation and streaming are well-suited for transactional systems built for high-performance queries.



What is data transformation?

- In the process of data transformation data is converted or manipulated into a different format to make it more suitable for analysis, visualization, or storage.
- This process typically involves cleaning, aggregating, integrating, and enriching data so that it can be used effectively by organizations.

Data transformation techniques:



DSW | DATA SCIENCE WIZARDS

DATA TRANSFORMATION TECHNIQUES

01 Data Smoothing

02 Attribute Construction

03 Data Generalization

04 Data Aggregation

05 Data Discretization

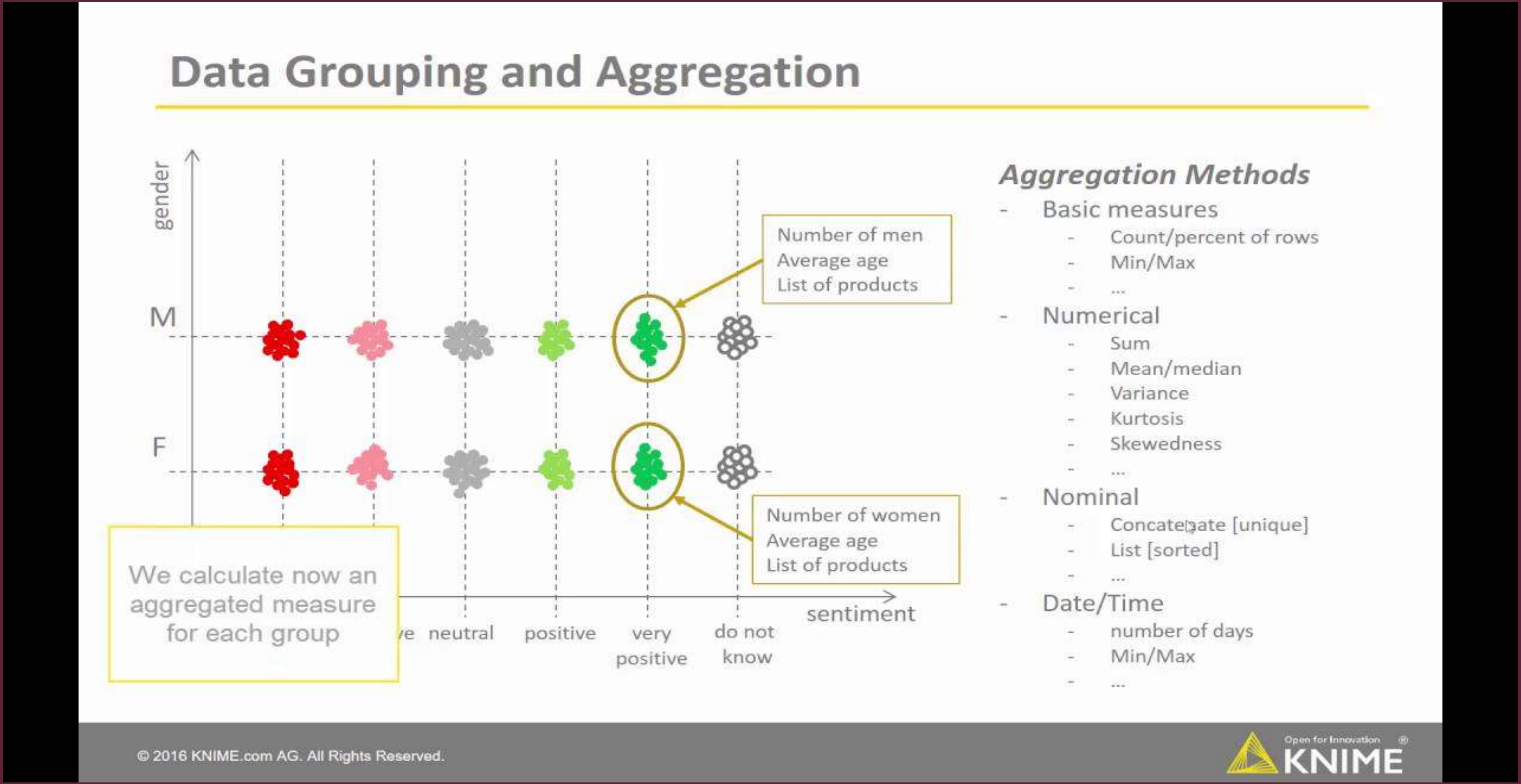
06 Data Normalization

1. Data Cleaning:

This technique involves removing or correcting any errors or inconsistencies in the data, such as missing values, duplicate entries, or outliers. Data cleaning ensures that the dataset is accurate and reliable for analysis.

2. Data Aggregation:

Aggregating data involves combining multiple rows of data into a single summary value, such as calculating averages, sums, or counts. This technique is useful for simplifying large datasets and summarising key information.



Data Grouping and Aggregation

Number of men
Average age
List of products

Number of women
Average age
List of products

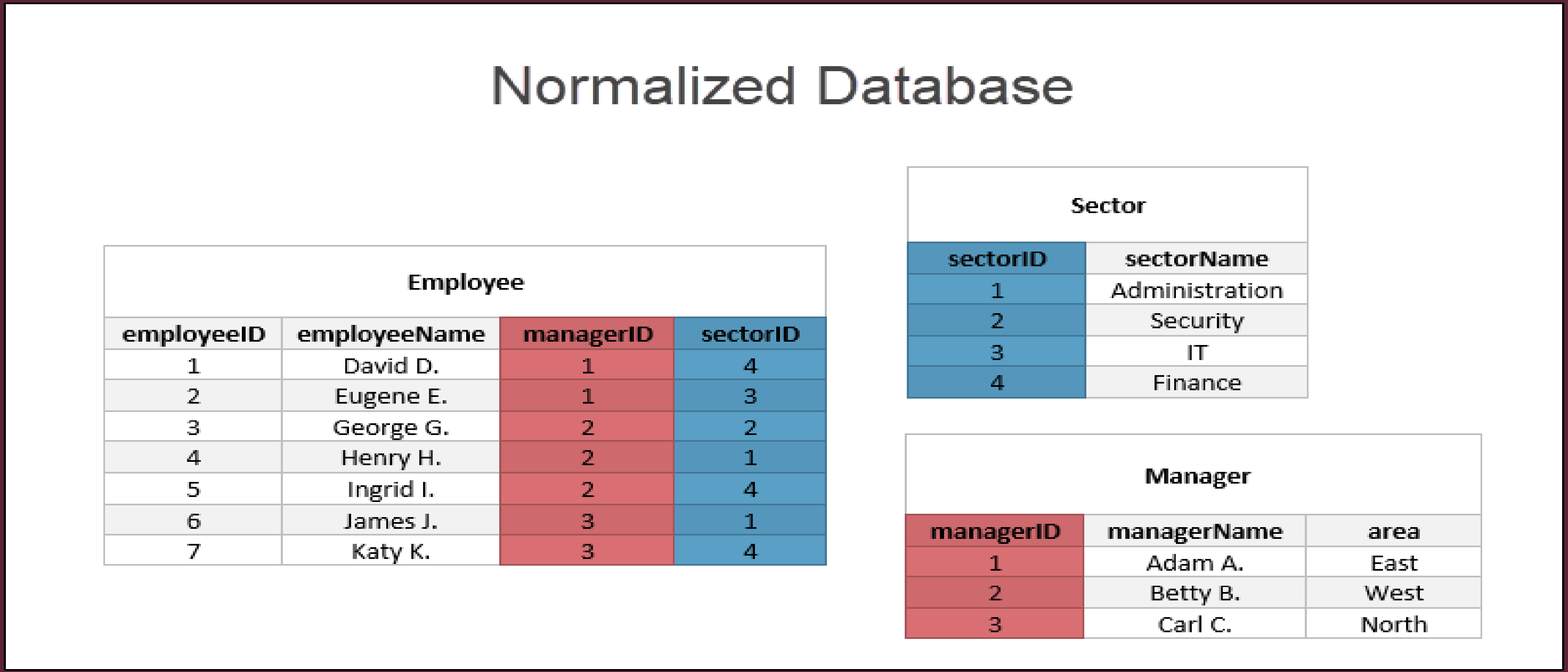
We calculate now an aggregated measure for each group

Aggregation Methods

- Basic measures
  - Count/percent of rows
  - Min/Max
  - …
- Numerical
  - Sum
  - Mean/median
  - Variance
  - Kurtosis
  - Skewedness
  - …
- Nominal
  - Concatenate [unique]
  - List [sorted]
  - …
- Date/Time
  - number of days
  - Min/Max
  - …

Data Normalization:

- Normalizing data involves scaling numerical values to a standard range, typically between 0 and 1.
- This technique helps to ensure that all variables have equal weight in the analysis and prevents bias towards variables with larger values.



Normalized Database

Employee

employeeID	employeeName	managerID	sectorID
1	David D.	1	4
2	Eugene E.	1	3
3	George G.	2	2
4	Henry H.	2	1
5	Ingrid I.	2	4
6	James J.	3	1
7	Katy K.	3	4

Sector

sectorID	sectorName
1	Administration
2	Security
3	IT
4	Finance

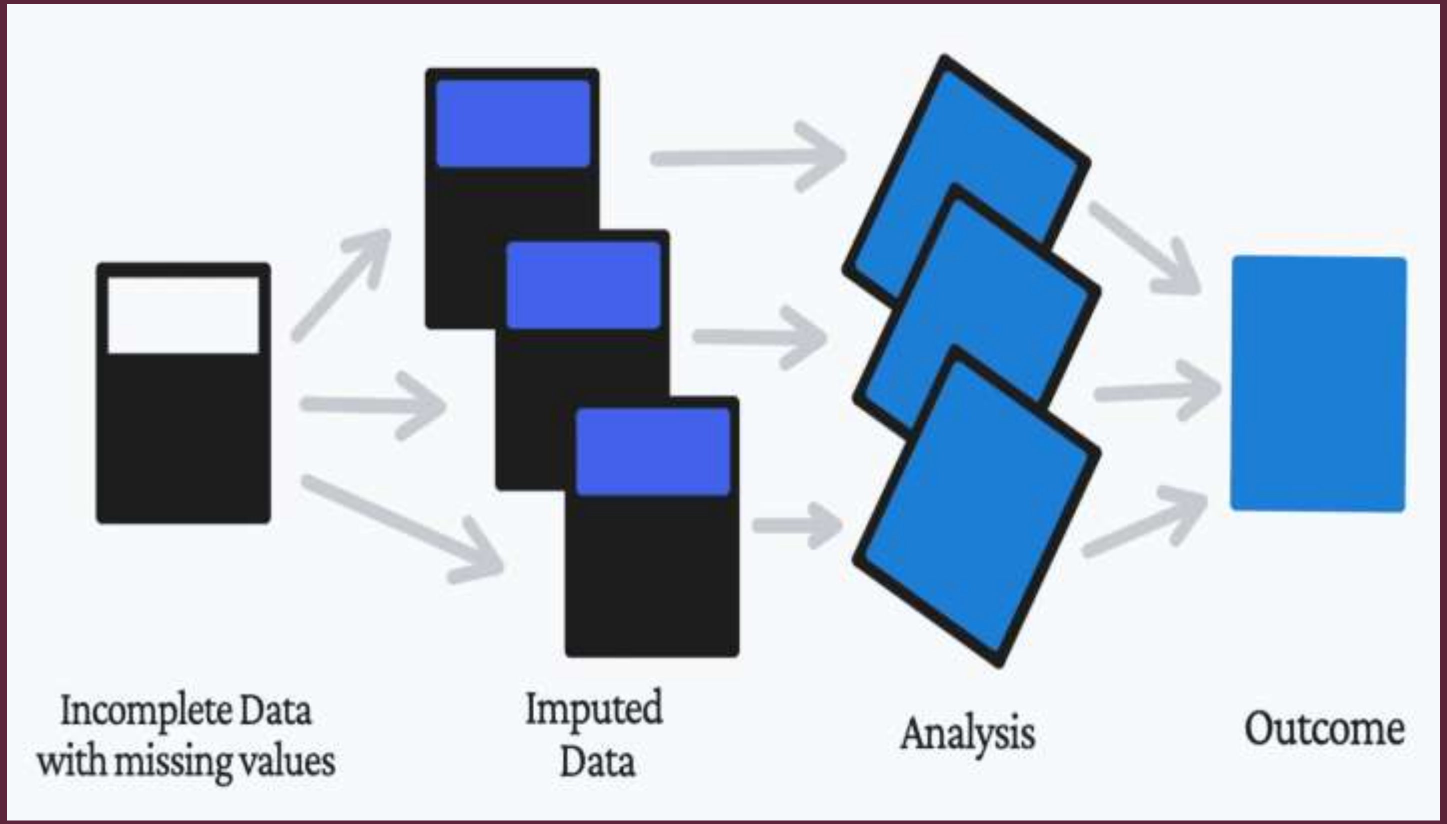
Manager

managerID	managerName	area
1	Adam A.	East
2	Betty B.	West
3	Carl C.	North

Data Encoding:

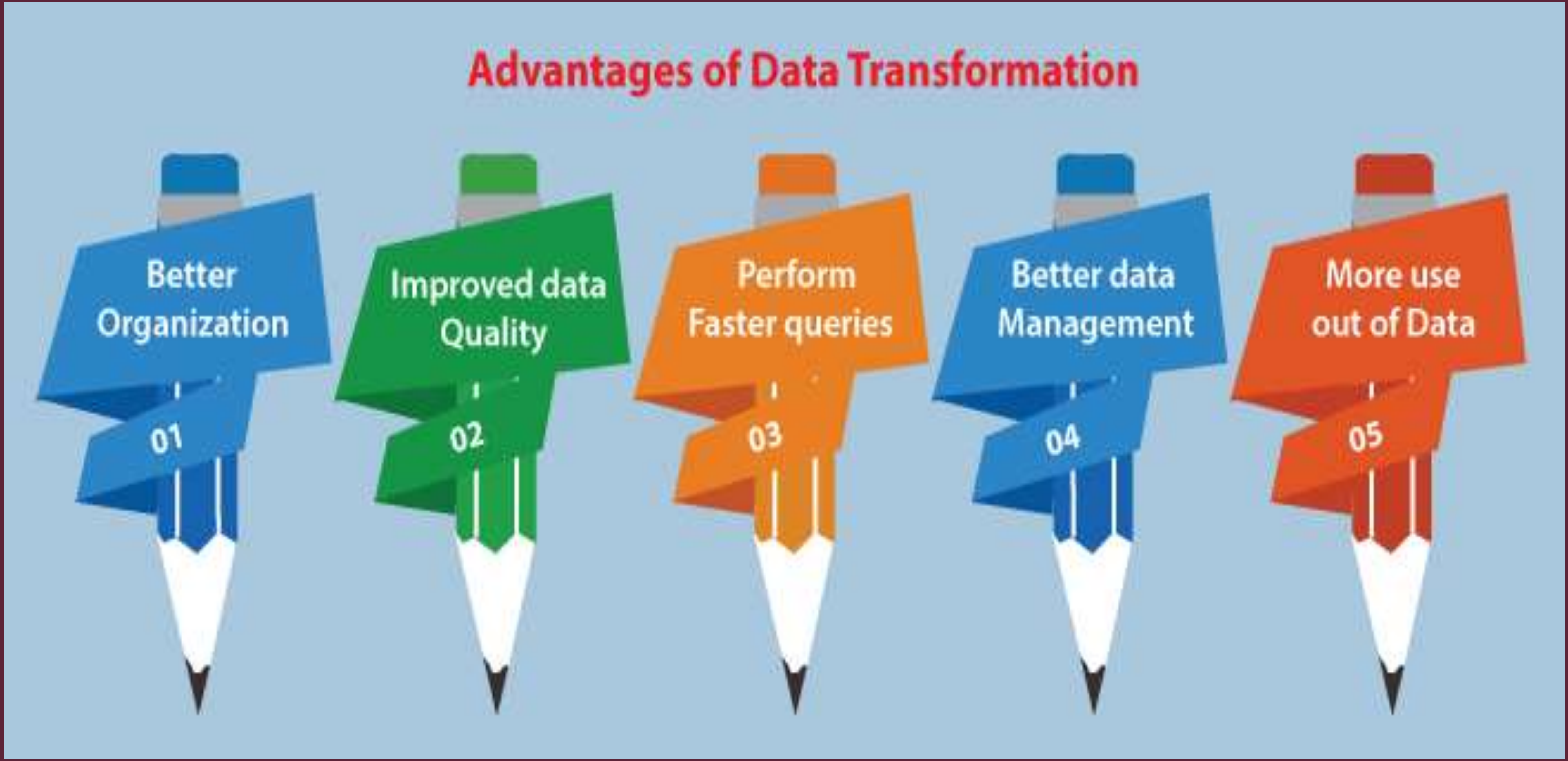
Encoding categorical variables involves converting text-based categories into numerical values that can be used in statistical models. This technique is essential for including categorical variables in machine learning algorithms.

Data Imputation:



- Computing missing values involves filling in missing data points with estimated values based on other observations in the dataset.
- This technique helps to preserve the integrity of the dataset and maintain accuracy in the analysis.

Benefits of data transformation


Advantages of Data Transformation
Better Organization
01
Improved data Quality
02
Perform Faster queries
03
Better data Management
04
More use out of Data
05