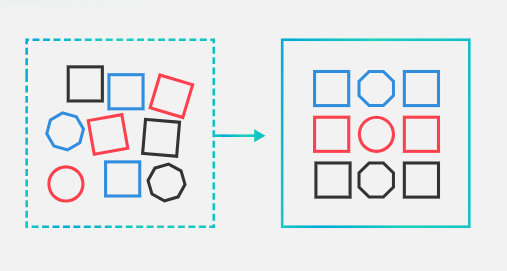
**Day 08 Assignment 2**

**What is data Transformation?**

Data transformation is the process of converting, [cleansing](https://www.tibco.com/glossary/what-is-data-cleansing), and structuring data into a usable format that can be analysed to support decision making processes, and to propel the growth of an organization.



1. Data transformation process converts raw data into a usable format by removing duplicates, converting data types and enriching the dataset.
2. This data transformation process involves defining the structure, mapping the data, extracting the data from the source system, performing the transformations, and then storing the transformed data in the appropriate dataset.
3. Data then becomes accessible, secure and more usable, allowing for use in a multitude of ways.

**Types of data transformation:**

**1. Constructive Transformation:**

Constructive transformations create new data attributes or features within the dataset. They enhance existing features or generate new ones to improve the quality and effectiveness of data analysis or machine learning models.

Example:

Imagine you have a dataset of customer orders, including the following information:

* **Order Date:** The date the order was placed (e.g., 2024-03-15)
* **Product ID:** A unique identifier for each product
* **Quantity:** The number of units ordered
* **Price per Unit:** The price of each unit

**Constructive Transformation:**

You want to calculate the total revenue for each order. This isn't directly available in your dataset, so you create a new column called "Total Revenue" by multiplying the "Quantity" and "Price per Unit" columns.

**Example:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Order Date** | **Product ID** | **Quantity** | **Price per Unit** | **Total Revenue** |
| 2024-03-15 | A123 | 2 | $10 |  |
| 2024-03-15 | B456 | 1 | $25 |  |
| 2024-03-16 | A123 | 3 | $10 |  |

**After Transformation:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Order Date** | **Product ID** | **Quantity** | **Price per Unit** | **Total Revenue** |
| 2024-03-15 | A123 | 2 | $10 | $20 |
| 2024-03-15 | B456 | 1 | $25 | $25 |
| 2024-03-16 | A123 | 3 | $10 | $30 |

**2.** **Destructive Transformations:**

* Destructive transformations remove unnecessary or irrelevant data from the dataset to make it more focused and efficient.
* Destructive data transformation types include data cleaning (removing duplicates or correcting errors), dealing with missing values (imputation or deletion).

Examples:

Lets say we are having data about customers address

* **Customer ID:** A unique identifier for each customer
* **Street Address:** The customer's street address
* **City:** The customer's city
* **State:** The customer's state
* **Zip Code:** The customer's zip code

**Scenario:** Due to a data entry error, some customers have duplicate address entries in the database. You want to clean up this data.

**Before Transformation (with duplicates):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Customer ID** | **Street Address** | **City** | **State** | **Zip Code** |
| 1 | 123 Main St | Anytown | CA | 91234 |
| 1 | 123 Main St | Anytown | CA | 91234 |
| 2 | 456 Oak Ave | Springfield | IL | 62701 |
| 3 | 789 Pine Ln | Anytown | CA | 91234 |
| 4 | 101 Elm St | Riverdale | NY | 10471 |
| 4 | 101 Elm St | Riverdale | NY | 10471 |

**After Transformation (duplicates removed):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Customer ID** | **Street Address** | **City** | **State** | **Zip Code** |
| 1 | 123 Main St | Anytown | CA | 91234 |
| 2 | 456 Oak Ave | Springfield | IL | 62701 |
| 3 | 789 Pine Ln | Anytown | CA | 91234 |
| 4 | 101 Elm St | Riverdale | NY | 10471 |

Its example of destructive transformation which is dealing with duplicate rows making data more organized and efficient.

**3. Formatting transformations:**

Formatting transformations deal with the presentation and organization of data, ensuring it adheres to a common format. These transformations include data standardization (converting data to a common format), sorting, and formatting.

Example:

Suppose I have a dataset of dates stored in various formats:

* **Date:** The date of an event

**Scenario:** You want to standardize the date format to YYYY-MM-DD for consistency and easier analysis.

**Before Transformation (various formats):**

|  |
| --- |
| **Date** |
| 03/15/2024 |
| March 15, 2024 |
| 2024-03-15 |

**After Transformation (YYYY-MM-DD format):**

|  |
| --- |
| **Date** |
| 2024-03-15 |
| 2024-03-15 |
| 2024-03-15 |

**4.** **Structural Transformations:**

Structural transformations involve modifying the overall structure and organization of the dataset, making it more suitable for analysis or machine learning models. This includes reshaping data, normalizing or denormalizing databases.

Example:

I have a dataset of student grades, but it's organized by class:

**Before Transformation (by class):**

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Student 1** | **Student 2** | **Student 3** |
| Math | 85 | 92 | 78 |
| Science | 90 | 88 | 95 |
| English | 75 | 80 | 85 |

**Scenario:** You want to reorganize this data so that each row represents a student, and the columns represent the subjects. This is a common task when you want to analyze student performance across different classes.

**After Transformation (by student):**

|  |  |  |  |
| --- | --- | --- | --- |
| **Student** | **Math** | **Science** | **English** |
| Student 1 | 85 | 90 | 75 |
| Student 2 | 92 | 88 | 80 |
| Student 3 | 78 | 95 | 85 |

In this type no data loss occurs only the restructuring of data is done.

Before it was organised as a class now for the better understanding and analyzing it is restructured by student.

**Data Transformation Techniques:**

Let’s see about each Technique of Data Transformation in detail.

**1. Data Smoothing:**

Smoothing is a technique where you apply an algorithm in order to remove noise from your dataset when trying to identify a trend. Noise can have a bad effect on your data and by eliminating or reducing it you can extract better insights or identify patterns that you wouldn’t see otherwise.

There are 3 algorithm types that help with data smoothing:

* Clustering: Where you can group similar values together to form a cluster while labelling any value out of the cluster as an outlier.
* Binning: Using an algorithm for binning will help you split the data into bins and smooth the data value within each bin.
* Regression: Regression algorithms are used to identify the relation between two dependent attributes and help you predict an attribute based on the value of the other.

**Example:**

K- means algorithm of clustering. (groups the data according to similarities)

**2. Attribute Construction:**

Attribution construction is one of the most common techniques in data transformation pipelines. Attribution construction or feature construction is the process of creating new features from a set of existing features/attributes in the dataset.

Example:

A simple example related to customer behavior.

|  |  |  |
| --- | --- | --- |
| **Customer ID** | **Total Purchases** | **Total Amount Spent** |
| 1 | 5 | $200 |
| 2 | 1 | $50 |
| 3 | 10 | $500 |
| 4 | 2 | $75 |
| 5 | 7 | $300 |

We can create another attribute from the above table.

The attribute showing the Average purchase amount calculated by using formula

**Average purchase amount = Total Amount Spent/ Total Purchases**

|  |  |  |  |
| --- | --- | --- | --- |
| **Customer ID** | **Total Purchases** | **Total Amount Spent** | **Average Purchase Value** |
| 1 | 5 | $200 | $40 |
| 2 | 1 | $50 | $50 |
| 3 | 10 | $500 | $50 |
| 4 | 2 | $75 | $37.50 |
| 5 | 7 | $300 | $42.86 |

**3.** **Data Aggregation:**

Data aggregation is the process where raw data is gathered and expressed in a summary form for statistical analysis.

For example, raw data can be aggregated over a given time period to provide statistics such as average, minimum, maximum, sum, and count.

Finding the average age of customer buying a particular product which can help in finding out the targeted age group for that particular product.

**Common Aggregation Functions:**

* SUM: Calculates the sum of values.
* COUNT: Counts the number of values.
* AVERAGE or MEAN: Calculates the average.
* MIN: Finds the minimum value.
* MAX: Finds the maximum value.
* MEDIAN: Finds the middle value.

**4. Data Generalization:**

Data generalization is a type of dynamic data masking where a particular data value is replaced with a less accurate value.

OR

Data generalization is a technique used to make data less specific and more general. This is often done to protect privacy or to make data easier to analyze.

Example:

Imagine you have a list of ages:

* 25
* 32
* 48
* 28
* 55

You could generalize this data by grouping the ages into ranges:

* 20-29
* 30-39
* 40-49
* 50-59

**5. Data Normalization:**

Data normalization is the process of scaling the data to a much smaller range, without losing information to help minimize or exclude duplicated data and improve algorithm efficiency and data extraction performance.

There are three methods to normalize an attribute:

* **Min-max normalization:**Where you perform a linear transformation on the original data.
* **Z-score normalization:**In z-score normalization (or zero-mean normalization) you are normalizing the value for attribute*A* using the mean and standard deviation.
* **Decimal scaling:**Where you can normalize the value of attribute *A* by moving the decimal point in the value.

**6. Data Decentralization:**

Data discretization refers to the process of transforming continuous data into a set of data intervals. This is an especially useful technique that can help you make the data easier to study and analyze and improve the efficiency of any applied algorithm.

Example:

Imagine having tens of thousands of rows representing people in a survey providing their first name, last name, age, and gender. Age is a numerical attribute that can have a lot of different values. To make our life easier we can divide the range of this continuous attribute into intervals.

Mapping this attribute to a higher-level concept, like youth, middle-aged, and senior, can help a lot with the efficiency of the task and improve the speed of the algorithms applied.

**7. Data Integration:**

Data integration is not a data transformation technique but rather a critical step during the pre-processing phase.

Data integration is the process of bringing together information from different sources to create a unified view of the data. These sources can be:

* Traditional databases
* Data warehouses
* Simple CSV or Excel files

**8. Data Blending:**

Data blending is the process of combining data from multiple sources. Sometimes this process is also referred to as data join.

With data blending, you can expand your reach to multiple data sources.

Example: In SQL we performs joins on the tables to blend the desired tables.

* (INNER) JOIN: Returns records that have matching values in both tables
* LEFT (OUTER) JOIN: Returns all records from the left table, and the matched records from the right table
* RIGHT (OUTER) JOIN: Returns all records from the right table, and the matched records from the left table
* FULL (OUTER) JOIN: Returns all records when there is a match in either left or right table

**9. Data Manipulation:**

Data manipulation refers to the process of making your data more readable and organized. This can be achieved by changing or altering your raw datasets.

Data manipulation tools can help you identify patterns in your data and apply any data transformation technique (e.g. attribution creation, normalization, or aggregation) in an efficient and easy way.

Data manipulation tools are Excel, PowerBI.