Data Integration and Transformation

Data Integration:

It combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files. The data integration systems are formally defined as triple. There are a number of issues to consider during the integration

- Schema Integration
- Redundancy
- Detection and Resolution of data value conflicts.

1. Entity Identification Problem

Schema Integration

- The real world entities from multiple sources be matched is referred to as the entity identification problem.
- The same attribute or object may have different names in different databases. For example, **customer_id** in one database and **cust_number** in another.

Example :

DS-1

E.ID	E.Name	Salary	Location
1	AA	40000	TN
2	BB	32000	TN
3	CC	50000	TN
4	DD	25000	TN
5	EE	15000	TN
6	FF	20000	TN

DS-2

Emp.ID	Emp.Name	Salary	Location
1	PAA	50000	TN
2	KBB	25000	TN
3	CTC	50000	TN
4	SDD	25000	TN
5	OEE	75000	TN
6	FPF	28000	TN

2. Redundancy

- It is another important issue. An attribute may be redundant if it can be derived from another table, such as age, annual revenue.
- Object Identification: The same attribute may have different names in different database.
- Derived data: one attribute may be derived from another attribute.

Example

Reg.No	S.Name	Course
1	AA	CSE
2	BB	ECE
3	CC	EEE
4	DD	CSE
5	EE	MECH

Reg.No	S.Name	Department
1	AA	CSE
2	BB	ECE
3	CC	EEE
4	DD	CSE
5	EE	MECH

Reg.No	S.Name	Department	Age	DOB
1	AA	CSE	19	02-01-2001
2	BB	ECE	18	22-04-2002
3	CC	EEE	20	12-11-2000
4	DD	CSE	18	05-01-2002
5	EE	MECH	19	08-02-2001

3. Detection And Resolution Of Data value Conflicts

- The same real world entity, attribute value from different sources. This may be due to differences in representation, scaling or encoding.
- An attribute in one system may be recorded as a lower level of abstraction than the

same attribute in another.

- For example, the total sales in one database may refer to another branch of all electronics , an attribute of the same in another database may refer to the total sales for electronics stores in a given region.

Product	Quantity	Price $
Hats	200	18
Hamburgers	1800	1
Tablets	40	200

Delhi Branch

Product	Quantity	Price ₹
Hats	205	1318
Hamburgers	1900	74
Tablets	42	14650

All Branch

Products Category	Actual Sales 2019
Product A	500000
Product B	100000
Product C	600000
Product D	50000
Product E	700000
Product F	750000

Products Category	Actual Sales 2019
Product A	10000000
Product B	2100000
Product C	3000000
Product D	42000000
Product E	53400000
Product F	74300000

Data Transformation

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Smoothing**, which works to remove noise from the data. Such techniques include binning, regression, and clustering
2. **Normalization**: It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
3. Attribute Selection: In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

4. **Discretization**:

   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

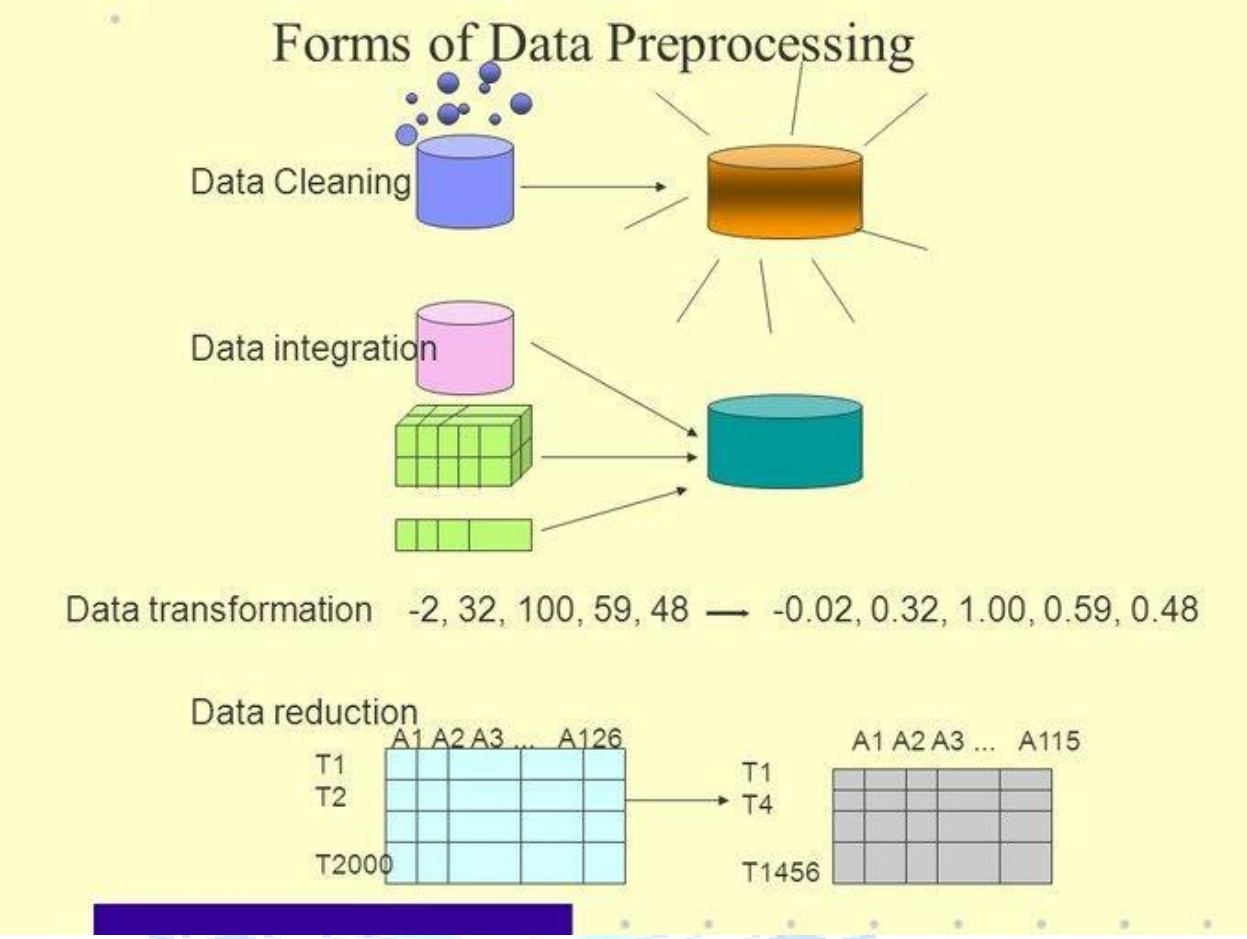
5. **Concept Hierarchy Generation**:

   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

Data Reduction

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs. The various steps to data reduction are:

1. **Data Cube Aggregation**:
   Aggregation operation is applied to data for the construction of the data cube.
2. **Attribute Subset Selection**:
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute. The attribute having p-value greater than significance level can be discarded.
3. **Numerosity Reduction**:
   This enable to store the model of data instead of whole data, for example: Regression Models.
4. **Dimensionality Reduction**:
   This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: Wavelet transforms and PCA (Principal Component Analysis).


Forms of Data Preprocessing
Data Cleaning
Data integration
Data transformation -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48
Data reduction
A1 A2 A3 ... A126
T1
T2
T2000
A1 A2 A3 ... A115
T1
T4
T1456

Data Discretization and Concept Hierarchy Generation

Discretization in data mining

- Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy.
- In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss.
- There are two forms of data discretization first is supervised discretization, and the second is unsupervised discretization.
- Supervised discretization refers to a method in which the class data is used.
- Unsupervised discretization refers to a method depending upon the way which operation proceeds. It means it works on the top-down splitting strategy and bottom-up merging strategy.

Now, we can understand this concept with the help of an example Suppose we have an attribute of Age with the given values

Age	1, 5, 9, 4, 7, 11, 14, 17, 13, 18, 19, 31, 33, 36, 42, 44, 46, 70, 74, 78, 77

Table Before & After Discretization

Attribute	Age	Age	Age	Age
Before Discretization	1,5,4,9,7	11,14,17,13,18,19	31,33,36,42,44,46	70,74,77,78
After Discretization	Child	Young	Mature	Old

Another example is analytics, where we gather the static data of website visitors. For example, all visitors who visit the site with the IP address of India are shown under country level. Some Famous techniques of data discretization

Histogram analysis

Histogram refers to a plot used to represent the underlying frequency distribution of a continuous data set. Histogram assists the data inspection for data distribution. For example, Outliers, skewness representation, normal distribution representation, etc.

Binning

Binning refers to a data smoothing technique that helps to group a huge number of continuous values into smaller values. For data discretization and the development of idea hierarchy, this technique can also be used.

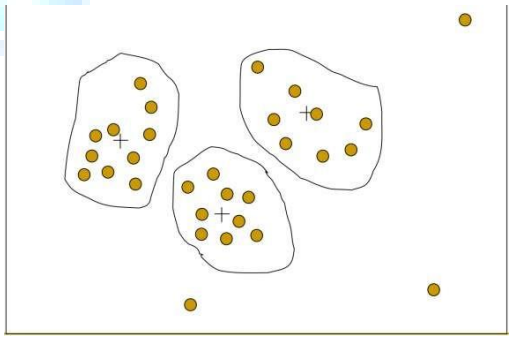
Binning methods for smoothing

Sorted data for price (in dollars) : 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

- Partition into equal frequency (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smoothing by Bin means:
  - Bin 1: 9, 9 ,9, 9
  - Bin 2: 23, 23 ,23, 23
  - Bin 3: 29, 29, 29, 29
- Smoothing by Bin Boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Cluster Analysis

Cluster analysis is a form of data discretization. A clustering algorithm is executed by dividing the values of x numbers into clusters to isolate a computational feature of x.



Data discretization using decision tree analysis

Data discretization refers to a decision tree analysis in which a top-down slicing technique is used. It is done through a supervised procedure. In a numeric attribute discretization, first, you need to select the attribute that has the least entropy, and then you need to run it with the help of a recursive process. The recursive process divides it into various discretized disjoint intervals, from top to bottom, using the same splitting criterion.

Data discretization using correlation analysis

Discretizing data by linear regression technique, you can get the best neighboring interval, and then the large intervals are combined to develop a larger overlap to form the final 20 overlapping intervals. It is a supervised procedure.

Concept Hierarchy Generation

Define Concept hierarchy

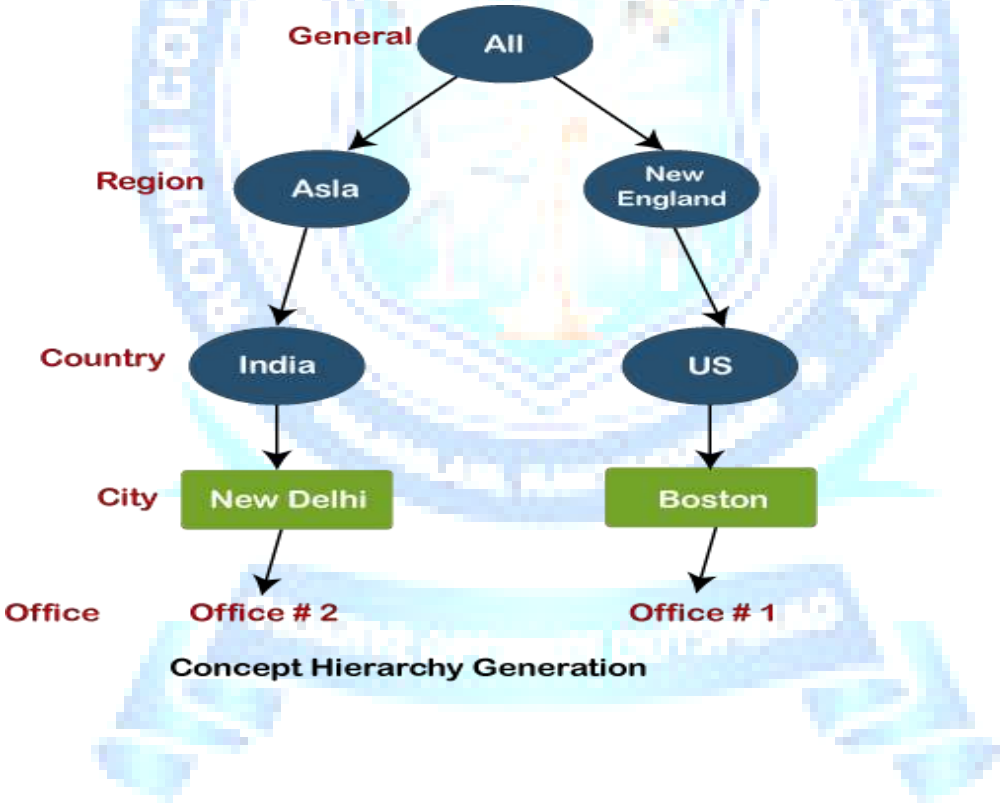
- It reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).
- The term hierarchy represents an organizational structure or mapping in which items are ranked according to their levels of importance.
- In other words, we can say that a hierarchy concept refers to a sequence of mappings with a set of more general concepts to complex concepts.
- It means mapping is done from low-level concepts to high-level concepts. For example, in computer science, there are different types of hierarchical systems.
- A document is placed in a folder in windows at a specific place in the tree structure is the best example of a computer hierarchical tree model.
- There are two types of hierarchy: top-down mapping and the second one is bottom-up mapping.
- Let's understand this concept hierarchy for the dimension location with the help of an example.
- A particular city can map with the belonging country. For example, New Delhi can be mapped to India, and India can be mapped to Asia.

Top-down mapping

Top-down mapping generally starts with the top with some general information and ends with the bottom to the specialized information.

Bottom-up mapping

Bottom-up mapping generally starts with the bottom with some specialized information and ends with the top to the generalized information.



Data discretization and binarization

- Data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss.
- In contrast, data binarization is used to transform the continuous and discrete attributes into binary attributes

Why is Discretization important?

As we know, an infinite of degrees of freedom mathematical problem poses with the continuous data. For many purposes, data scientists need the implementation of discretization. It is also used to improve signal noise ratio.