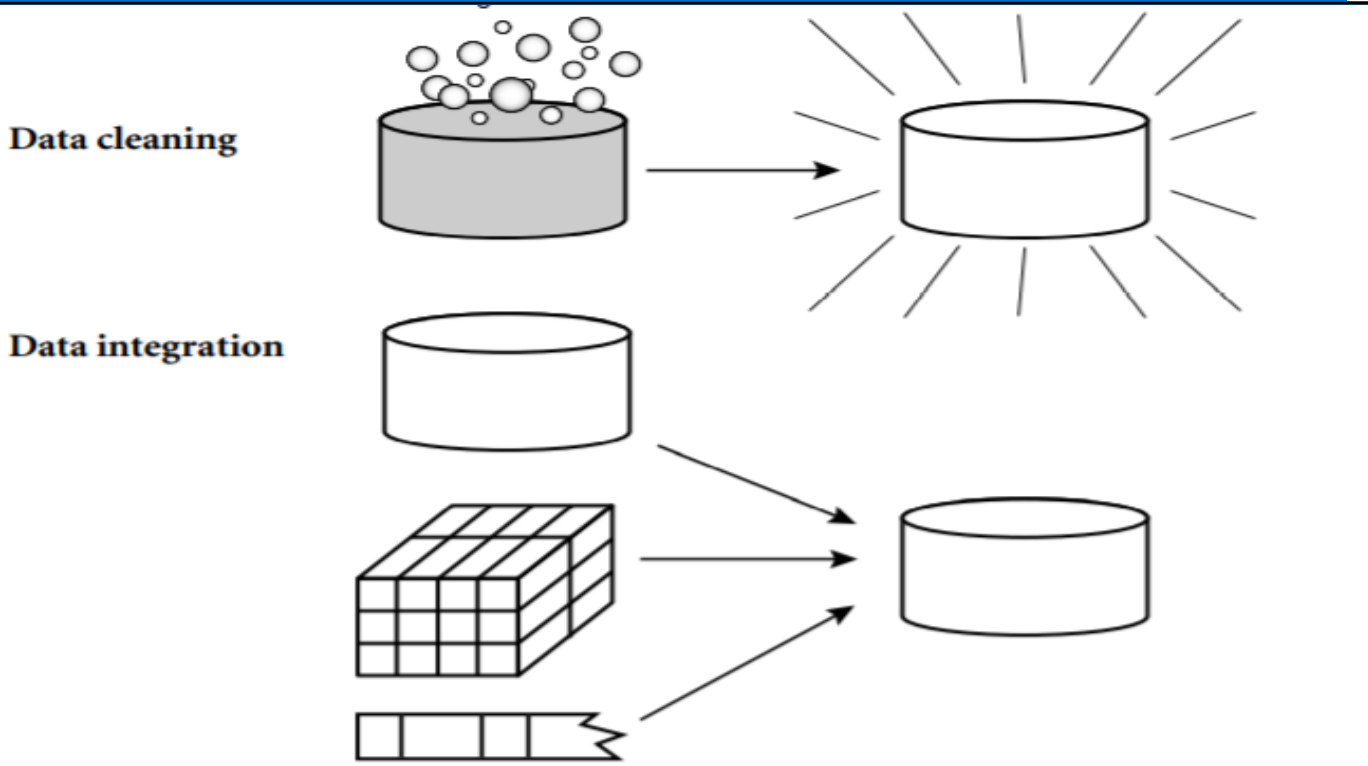
UNIT-2 : DATA PRE-PROCESSING

NEED FOR DATA PRE-PROCESSING

DATA PRE-PROCESSING

INCOMPLETE DATA
NOISY DATA
INCONSISTATNT DATA

FORMS OF DATA PRE-PROCESSING



Descriptive Data Summarization

Central Tendency of the data
Dispersion of the data

Descriptive Data Summarization

Central Tendency of the data

- Mean
- Median
- Mode
- Midrange

Descriptive Data Summarization

(ESTD-1995)

Mean

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

Descriptive Data Summarization

Median

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$

Descriptive Data Summarization

Mode

The mode for a set of data is the value that occurs most frequently in the set

$$mean - mode = 3 \times (mean - median)$$

Descriptive Data Summarization

Midrange

The midrange can also be used to assess the central tendency of a dataset.

Descriptive Data Summarization

Example:

Suppose that the data for analysis includes the attribute grade. The grade values for the data tuples are:

4, 5, 9, 11, 12, 13, 13, 13, 13, 14, 15, 15, 16, 17, 18, 18, 19, 20

Descriptive Data Summarization

Example:

4, 5, 9, 11, 12, 13, 13, 13, 13, 14, 15, 15, 16, 17, 18, 18, 19, 20
N=18 (EVEN)

the mean = 13.61

The median = (13+14)/2 = 13.5

The mode (value occurring with the greatest frequency) of the data is 13, the mode is only one value so it's called unimodal.

The midrange (average of the largest and smallest values in the data set) of the data is:
(20+ 4) / 2 = 12

Descriptive Data Summarization

Range, Quartiles, Outliers, and Boxplots

Let x1,x2,...,xN be a set of observations for some attribute

The range of the set is the difference between the largest (max()) and smallest (min()) values

The kth percentile of a set of data in numerical order is the value xi having the property that k percent of the data entries lie at or below xi . The median (discussed in the previous subsection) is the 50th percentile

Descriptive Data Summarization

Range, Quartiles, Outliers, and Boxplots

The most commonly used percentiles other than the median are quartiles. The first quartile, denoted by Q1, is the 25th percentile; the third quartile, denoted by Q3, is the 75th percentile

This distance is called the interquartile range (IQR) and is defined as IQR = Q3 –Q1

A common rule of thumb for identifying suspected outliers is to single out values falling at least 1.5×IQR above the third quartile or below the first quartile.

Descriptive Data Summarization

Range, Quartiles, Outliers, and Boxplots

The most commonly used percentiles other than the median are quartiles. The first quartile, denoted by Q1, is the 25th percentile; the third quartile, denoted by Q3, is the 75th percentile

This distance is called the interquartile range (IQR) and is defined as IQR = Q3 –Q1

A common rule of thumb for identifying suspected outliers is to single out values falling at least 1.5×IQR above the third quartile or below the first quartile.

Descriptive Data Summarization

Five-number summary of a distribution consists of the median, the quartiles Q1 and Q3, and the smallest and largest individual observations, written in the order Minimum, Q1, Median, Q3, Maximum.

Boxplots are a popular way of visualizing a distribution.

Typically, the ends of the box are at the quartiles, so that the box length is the interquartile range, IQR. The median is marked by a line within the box. Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations

Descriptive Data Summarization