CHAPTER - 6: DATA EXPLORATION

No Questions were asked from this Chapter in May 16 and Dec 16 Paper of Mumbai University.

CHAPTER - 7: DATA PREPROCESSING

Q1] DISCUSS DIFFERENT STEPS INVOLVED IN DATA PREPROCESSING. **[10M – MAY16]**

ANS:

DATA PREPROCESSING:

1. Data pre-processing is an important step in the data mining process.
2. Data preprocessing involves **transforming** raw data into an understandable format.
3. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors.
4. Data preprocessing is a proven method of resolving such issues.
5. Data preprocessing prepares raw data for further processing.

STEPS INVOLVED IN DATA PREPROCESSING:

I) Data Cleaning:

- Data Cleaning is also known as **Scrubbing**.
- It is a technique that is applied to **remove the noisy data** and **correct the inconsistencies** in data.
- It involves filling missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- Steps in data cleansing:
  - **Parsing:** Parsing is the process in which individual data elements are located and identified in the source systems and then these elements are isolated in the target files.
  - **Correcting:** In this step, using data algorithm the individual data elements are corrected.
  - **Standardizing:** In standardizing process, conversion routines are used to transform data into a consistent format using both standard and custom business rules.
  - **Matching:** Matching process involves eliminating duplications by searching and matching records.
  - **Consolidating:** Consolidating process involves merging the records into one representation by analyzing and identifying relationship between matched records.

- Figure 7.1 shows example of data cleaning process.

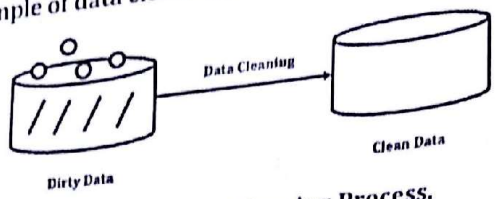

Figure 7.1: Data Cleaning Process.

II) Data Integration:

- Data integration involves combining data residing in different sources and providing users with a unified view of these data.
- Sources may include multiple databases, data cubes or data files.
- Data Integration removes the duplicate and redundant data.
- Figure 7.2 shows example of data integration process.

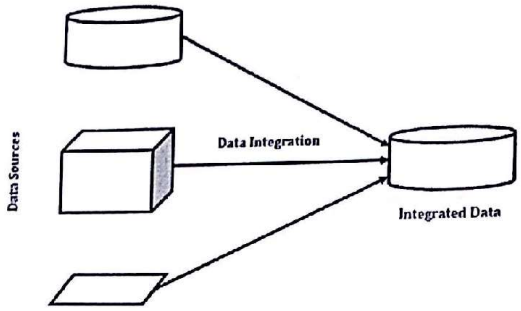

Figure 7.2: Data Integration Process.

III) Data Transformation:

- In Data Transformation, data are transformed or consolidated into forms appropriate for mining.
- Data transformation involves:
  - Smoothing.
  - Aggregation.
  - Generalization.
  - Normalization.
- Figure 7.3 shows the example of data transformation process.

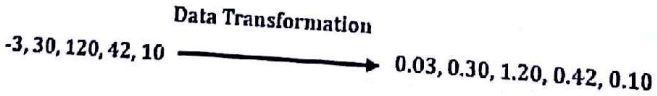

Figure 7.3: Data Transformation Process.

IV) Data Reduction:

- Data Reduction is used to obtain a reduced representation of the data set that is much smaller in volume.

- Strategies for data reduction includes:
  - **Data Cube Aggregation:** In Data Cube Aggregation, aggregation operations are applied to the data in the construction of a data cube.
  - **Attribute subset selection:** This process is used to detect and remove irrelevant, weakly relevant, or redundant attributes or dimensions.
  - **Dimensionality Reduction:** In this process encoding mechanisms are used to reduce the data set size.
  - **Numerosity Reduction:** In this process, the data are replaced by alternative, smaller data representations such as parametric models and non-parametric models like clustering.
- Figure 7.4 shows data reduction process example.

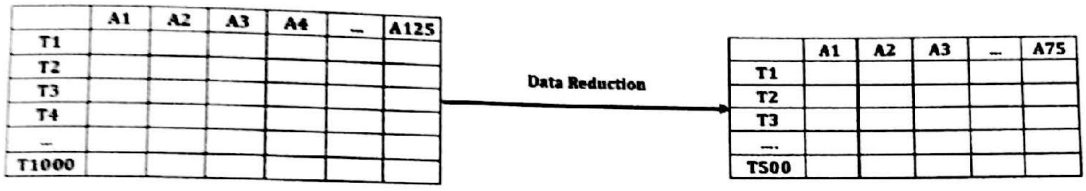

Figure 7.4: Data Reduction Process.

Data Discretization:

In Data Discretization, the range of a continuous attribute is divided into intervals.

By discretization the size of the data is reduced.

In this process, the data is prepared for further analysis.

Discretization process is applied recursively on an attribute.

Three types of attributes:

- **Nominal:** Values from an unordered set.
- **Ordinal:** Values from an ordered set.
- **Continuous:** Real numbers.