Dr. SNS RAJALAKSHMI COLLEGE OF ARTS & SCIENCE
(Autonomous)
Coimbatore -641049

Accredited by NAAC(Cycle–III) with ‘A+’ Grade
(Recognized by UGC, Approved by AICTE, New Delhi and
Affiliated to Bharathiar University, Coimbatore)

DEPARTMENT OF COMMERCE (IT)

COURSE NAME : 21UCI508 - Business Intelligence

III YEAR / V SEMESTER

Unit IV

Data Cleansing

Meaning of Data Cleansing:

Data cleansing refers to the process of systematically identifying and correcting errors or inconsistencies in data to improve its quality. It involves the removal of duplicate entries, correction of formatting issues, handling of missing values, and validation of data against defined standards. The primary goal is to make the data accurate, complete, consistent, and ready for meaningful analysis and business use.

Definition of Data Cleansing:

"Data Cleansing is the process of detecting, correcting, or removing corrupt, inaccurate, irrelevant, incomplete, or improperly formatted data from a dataset to improve its quality and reliability for decision-making and analysis."

Steps in the Data Cleansing Process

1. Data Profiling

The first step is to examine the existing data to understand its structure, content, and quality. This involves analyzing data to detect patterns, anomalies, duplicates, missing values, and inconsistencies. Data profiling provides insights into the types of problems that need to be fixed during cleansing.

2. Remove Duplicate Records

Duplicate records can distort analysis and cause operational inefficiencies. This step involves identifying and eliminating repeated or redundant data entries using techniques like fuzzy matching, unique keys, or comparison logic to ensure each record is unique.

3. Handle Missing or Incomplete Data

This step addresses gaps or null values in the dataset. Missing data can be managed by filling in default values, estimating values through statistical methods, or removing records entirely if the missing information is critical and cannot be recovered.

4. Correct Data Errors

Errors such as spelling mistakes, incorrect entries, or invalid values are corrected in this step. For example, fixing city names like “Mumbay” to “Mumbai” or replacing “NA” with actual values based on logic or reference tables.

5. Standardize Data Formats

In this step, data is transformed into a consistent and uniform format. For example, dates can be standardized to “DD-MM-YYYY,” names can be converted to proper case, and phone numbers can be formatted uniformly. This enhances consistency across datasets.

6. Validate Data

Validation ensures that data adheres to predefined rules and constraints. For instance, email addresses should follow the correct syntax, numeric fields should not contain letters, and age fields should not be negative. Validation helps maintain logical accuracy.

Techniques Used in Data Cleansing:

1. Removing Duplicate Records

Duplicates can result from multiple entries or merging databases and may lead to incorrect results.

2. Standardization of Data:

Standardization ensures that all data follows a consistent format and style, which is crucial for data integration.

3. Handling Missing or Incomplete Data

Missing data can cause analytical distortions and needs appropriate handling.

4. Data Validation

This ensures data meets business and logical rules, maintaining integrity and reliability.

5. Data Transformation

Involves changing the structure or representation of data to meet analysis or business needs.

6. Error Correction

7. Lookup and Reference Table Usage

Reference tables help in validating and cleaning data efficiently.

8. Parsing and Reformatting

Helps in cleaning and organizing complex fields.

a. Field Parsing:

9. Outlier Detection and Handling

Outliers may indicate errors or special cases and need careful treatment.

10. Automation and Tool-Based Cleansing

Cleansing becomes faster and more accurate with tools and scripts.


Thank
you