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 Profiling , Data Profiling activities

Unit 4
Data Profiling

Data Profiling refers to the systematic analysis and assessment of data to extract useful information about its condition. It is a preliminary step to explore data, verify its quality, and prepare it for further processing, such as cleansing, transformation, or integration.

Definition of Data Profiling

Data Profiling can be defined as “The process of examining data from an existing source and collecting statistics or informative summaries about that data to understand its structure, content, relationships, and quality.”

Purpose of data profiling

Data profiling serves several essential purposes in modern data management, each contributing to the overall reliability and usability of organizational data:

Assess Data Quality: Data profiling evaluates data for completeness, accuracy, uniqueness, consistency, and validity, helping to identify and address issues such as missing values, duplicates, and anomalies. This assessment ensures that decisions are based on high-quality, credible data and supports compliance with regulations and standards.

Understand Data Structure: Profiling uncovers the structure of data by identifying data types, formats, lengths, value distributions, and patterns. This structural understanding is crucial for effective data management, integration, and migration.

Discover Relationships: The process reveals key relationships between columns and tables, such as primary keys, foreign keys, and functional dependencies. Recognizing these relationships is vital for optimizing databases and ensuring data consistency across systems.

Support Data Cleansing: Data profiling pinpoints errors and inconsistencies, enabling targeted data cleansing efforts to correct or remove problematic records. By detecting issues early, organizations can prevent small mistakes from escalating into larger problems.

Aid Data Integration: Profiling ensures that data from different sources can be reliably combined by checking for compatibility and identifying potential integration challenges. This step is especially important in data warehousing, business intelligence, and big data projects.

Help with Data Governance: The insights from profiling support data governance by documenting data quality metrics, monitoring compliance, and building robust data policies and rules. Data profiling is a foundational activity in comprehensive data governance strategies.

Data Profiling Activities:

1. Structure Analysis (Column Profiling)

Structure analysis, also known as column profiling, examines each column in a dataset to understand its fundamental properties and characteristics. This process involves analyzing data types, formats, minimum and maximum lengths, presence of null values, and identifying patterns within the data. The main goal is to validate that the data conforms to expected formats and to detect inconsistencies or anomalies.

2. Completeness Analysis

Completeness analysis evaluates the number and percentage of missing (null) values in each field. It measures how much of the required data is present and identifies fields with insufficient data, which may impact reporting or analytics.

3. Uniqueness Analysis

Uniqueness analysis checks whether values in a field are unique or duplicated. It helps identify potential primary keys or candidate keys, which are critical for data normalization and maintaining relational integrity.

4. Value Frequency Analysis

This analysis determines how often each value appears in a field. It helps spot outliers, default values (such as 9999), and dominant patterns, which can indicate issues or business rules.

5. Range Analysis

Range analysis checks the minimum and maximum values in numeric or date fields. It detects anomalies or invalid data entries, such as dates set in the future or values outside acceptable business ranges.

6. Pattern Analysis

Pattern analysis examines the patterns within text or alphanumeric data to detect inconsistencies in data formats (e.g., phone numbers, email addresses).

Use: Supports standardization and format validation, ensuring consistency across data entries.

7. Dependency Analysis (Cross-Column Profiling)

Dependency analysis examines relationships between different fields, such as ensuring that "City" matches "Pincode" or "Country." It helps uncover hidden business rules or data inconsistencies.

Use: Identifies hidden business rules and data inconsistencies, supporting data quality and integrity.

8. Redundancy Analysis

Redundancy analysis identifies repeated or duplicated data entries across datasets or records. Reducing redundancy improves data efficiency and reduces storage requirements.

Use: Enhances data efficiency and storage optimization by eliminating unnecessary duplicates.

9. Drill-Down Profiling

Drill-down profiling enables in-depth exploration of suspect values by examining subgroups or sub-dimensions of data. This helps trace the root causes of data quality issues.

Use: Facilitates root cause analysis for data quality problems by allowing focused investigation.

10. Metadata Validation

Metadata validation compares the actual data structure and values against the defined metadata (schema). It ensures that the database structure matches what is documented, which is crucial for system integration and ETL processes.


Thank
you