Report on Data Normalization

Mahir Hashimov
12695120
Unit 7
Deciphering Big Data
University of Essex Online
21 June 2024

Contents

Report on Data Normalization 3
Introduction 3
Steps of Normalization 3
Challenges Faced 4
Critical Analysis 5
Conclusion 5
Output 6
References 7

Report on Data Normalization

Introduction

Data normalization is a critical process in database design to ensure the integrity and efficiency of data storage (Diène et al., 2020). The goal of normalization is to eliminate redundant data and ensure logical data dependencies (Tchernykh et al., 2020 and Singh & Singh, 2020) . This report details the process of normalizing a dataset from an unnormalized form to the Third Normal Form (3NF), highlighting the steps, challenges faced, and a critical analysis of the process.

Steps of Normalization

1. Unnormalized Data (UNF)

The initial dataset was provided in an unnormalized form with various anomalies such as missing values and duplicated rows. The dataset included student details, exam scores, course names, exam boards, and teacher names.

2. First Normal Form (1NF)
    - Objective: Ensure that each column contains atomic (indivisible) values and each record is unique.
    - Process:
        - Remove any completely empty rows.
        - Forward-fill missing values to ensure each row contains a complete record.
    - Outcome: The data was structured such that each field contained only one value, and records were made unique by filling in missing information from adjacent rows.
3. Second Normal Form (2NF)
    - Objective: Remove partial dependencies; ensure that all non-key attributes are fully functionally dependent on the primary key.
    - Process:
    - Identify the primary entities: Students, Courses, and Teachers.
    - Create separate tables for each entity to remove partial dependencies:
    - Students Table with Student Number, Student Name, and Date of Birth.

    - Courses Table with Course Name and Exam Board.
    - Teachers Table with Teacher Name.
- Merge these tables back to form a composite table that includes Student Number, Course Name, Teacher Name, Exam Score, and Support.
- Outcome: Partial dependencies were removed, ensuring that each non-key attribute was fully dependent on the primary key.

4. Third Normal Form (3NF)
    - Objective: Remove transitive dependencies; ensure that non-key attributes are not dependent on other non-key attributes.
    - Process:
    - Identify transitive dependencies in the merged table.
    - Create a separate Exam Scores Table with Student Number, Course Name, Exam Score, and Support to eliminate transitive dependencies.
    - Outcome: The dataset was split into well-structured tables where each non-key attribute was only dependent on the primary key.

Challenges Faced

1. Handling Missing Data:
    - Initially, the dataset contained rows with entirely missing values, which had to be carefully removed.
    - Forward-filling missing values required careful consideration to avoid incorrect data propagation.
2. Identifying Dependencies:
    - Determining partial and transitive dependencies required a thorough understanding of the relationships between different data attributes.
3. Column Name Anomalies:
    - The dataset had inconsistencies in column names, such as extra spaces and non-printable characters, which had to be cleaned for accurate data processing.
4. Ensuring Data Integrity:
    - Splitting the data into multiple normalized tables while maintaining the integrity and consistency of the data was crucial.

Critical Analysis

Normalization is a systematic approach to organizing data in a database to reduce redundancy and improve data integrity (Sharma & Kazim, 2021). The process of converting data to 3NF involved several critical steps:

1. Data Cleaning:
    - Cleaning the data and ensuring it was in 1NF was the foundation for further normalization. This step was essential to eliminate duplicated and incomplete records, setting the stage for more advanced normalization.
2. Entity Identification:
    - Identifying key entities such as Students, Courses, and Teachers was crucial. This step ensured that the database was structured around logical entities, which is a core principle of database normalization.
3. Dependency Removal:
    - Removing partial and transitive dependencies is vital for ensuring that the database supports efficient querying and updates. This step significantly reduced data anomalies and redundancy, leading to a more robust database design.
4. Practical Considerations:
    - While normalization improves database design, it can lead to increased complexity in database queries due to the need to join multiple tables. Therefore, a balance between normalization and query performance must be considered.

Conclusion

The normalization process transformed the unnormalized dataset into a well-structured set of tables in 3NF, eliminating redundancy and ensuring data integrity. Despite the challenges faced, the process highlighted the importance of systematic data cleaning, entity identification, and dependency removal. This structured approach not only enhances data storage efficiency but also facilitates more accurate and efficient data retrieval, which is crucial for any data-driven application.

Output

```
Students Table:
3           1002.0  Sally Davies    1999-10-02
6           1003.0  Mark Hanmill    1995-06-05
9           1004.0      Anas Ali    1980-08-03
12          1005.0     Cheuk Yin    2002-05-01

Courses Table:
        Course Name Exam Board
0  Computer Science        BCS
1             Maths    EdExcel
2           Physics        OCR
3             Maths        AQA
4           Biology       WJEC

Teachers Table:
  Teacher Name
0     Mr Jones
1    Ms Parker
2    Mr Peters
4    Mrs Patel
5   Ms Daniels

Exam Scores Table:
   Student Number       Course Name  Exam Score Support
0          1001.0  Computer Science        78.0      No
1          1001.0             Maths        78.0      No
3          1001.0           Physics        78.0      No
4          1002.0             Maths        55.0     Yes
6          1002.0           Biology        55.0     Yes
```

References

Diène, B., Rodrigues, J.J.P.C., Diallo, O., Ndoye, E.H.M., and Korotaev, V.V., 2020. Data management techniques for Internet of Things. Mechanical Systems and Signal Processing, 138, 106564. Available at: https://doi.org/10.1016/j.ymssp.2019.106564.

Tchernykh, A., Babenko, M., Chervyakov, N., Miranda-López, V., Avetisyan, A., and Drozdov, A.Y., 2020. Scalable data storage design for nonstationary IoT environment with adaptive security and reliability. IEEE Internet of Things Journal, 7(10), pp.8763-8777. Available at: https://ieeexplore.ieee.org/document/9037363.

Singh, D. and Singh, B., 2020. Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97, 105524. Available at: https://doi.org/10.1016/j.asoc.2019.105524.

Sharma, R. and Kazim, A., 2021. A low end case tool synthesizing 3NF relations using Bernstein's algorithm. In Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021. Available at: http://dx.doi.org/10.2139/ssrn.3884346.