All T20 Internationals Dataset (2005 - 2023) | Normalization Justification for ERD

Avinash Bunga

Information Systems and Business Analytics, Park University

CIS622DLAF2P2023 Data Architecture for Business Analytics

Professor: Gulnoza Khakimova

Nov 05, 2023

All T20 Internationals Dataset (2005 - 2023) | Normalization Justification for ERD

Introduction:

Embarking on the meticulous task of designing an Entity-Relationship Diagram (ERD) for the "All T20 Internationals Dataset (2005 - 2023)" presented a unique challenge: to create a data structure that is not only reflective of the complexities of cricket data but also adheres to the strict principles of 3rd Normal Form (3NF). The goal was to design a database schema that ensures data integrity, eliminates redundancy, and provides flexibility for future updates and queries. By carefully deconstructing and analyzing the dataset, we developed a comprehensive normalization strategy that transforms raw data into a well-organized, relational database.

Normalization Justification Summary

The comprehensive normalization of the "All T20 Internationals Dataset (2005 - 2023)" into 12 distinct tables represents a strategic approach to database design that carefully balances the granularity of data storage with operational efficiency.

Strategic Table Segregation for Enhanced Normalization

The decision to create individual tables for Referee, Umpire, Series, and Venue stems from a fundamental principle of normalization: the singularity of purpose. By isolating these entities, we eliminate redundant data across multiple records, thus ensuring that updates to a particular referee or venue do not necessitate widespread changes throughout the database. This approach not only simplifies maintenance but also enhances data retrieval speed for queries specific to these attributes.

Team Table for Centralized Updates

Creating a Team table extracted from the t20i_Matches_Data allows us to centralize team names in a single repository. This is particularly beneficial when team names undergo changes; we can update one record in the Team table without the need to cascade changes through the entire Match table. It reduces the risk of data anomalies and ensures consistency across historical and current match records.

MatchPlayers Table for Simplified Record Management

With the MatchPlayers table, we can cleanly store the roster of players for each match. This not only streamlines the match data but also prevents the replication of player information across different matches. The efficiency of data entry is significantly improved, and the integrity of player records is upheld.

Players Table as a Single Source of Truth

The Players table serves as the single source of truth for all player-related data. By maintaining player names in one centralized location, we avoid duplication and facilitate easier updates. It allows for a more straightforward tracking of player statistics over time, regardless of the number of matches they participate in.

TeamCaptain Table for Role-Specific Data

The TeamCaptain table is instrumental in associating team captains with their respective teams and matches. Since the role of the captain can change from match to match, having a dedicated table allows us to record these shifts without redundancies. This design choice ensures that historical data remains accurate even during leadership changes.

Wicket Table for Discrete Event Recording

Finally, the Wicket table provides a discrete record of wicket events, avoiding the repetition of wicket types within the BattingCard table. This ensures that our database does not inflate with repetitive text strings, and it allows for a more complex analysis of wicket events over the dataset's lifespan.

Conclusion:

The careful curation of our ERD into 12 distinct entities epitomizes the essence of 3NF normalization. Each table has been crafted to ensure that every non-primary key attribute is fully functionally dependent on the primary key alone, thus eliminating transitive dependency and ensuring no redundancy. As a result, we have achieved a level of normalization that not only systematically organizes the data for optimal utility and scalability but also fortifies the structural integrity of the database against anomalies. This particular approach lays a solid foundation for a database system that is robust, efficient, and capable of handling complex cricketing data with precision, thereby setting a benchmark for data management within the domain of sports analytics.

Reference

PRASAD, B. (2023, October 23). *All T20 Internationals Dataset (2005 - 2023).* Kaggle. https://www.kaggle.com/datasets/bhuvaneshprasad/all-t20-internationals-dataset-2005-to-2023?select=t20i_Matches_Data.csv