CS586 - Introduction to Database Systems Project Domain Description and Data Source

Submitted by: - Parth Parashar (923928157)

Topic: - Indian Premier League (IPL)

Introduction

Cricket has long been a part of the Indian society for decades now and with the evolution of cricket from the traditional 5-day long test matches to the modern 3-hour T-20 matches, the amount of data that is being generated is huge.
One such iteration of the modern form of the game is Indian Premier league which is the biggest and the most popular version across the globe. Hundreds of players from across the globe come and participate in the 8-16-team tournament.

A Cricket match is generally played with 2 teams each consisting of 11 players. One team is the bowling team, and the other is the batting team. Each team takes turns for batting and bowling and whosoever scores more "runs" at the end of "20 overs" wins the game.
Since there are a lot of players involved in a single match along with different combinations of batsman and bowlers possible, making a one-to-one player strategy can become cumbersome and this is where the power of database can come to rescue.
Thus, it becomes important to categorize the data in the form of tables and then deduct meaningful insights and conclusions from the vast amount of data.

Dataset utilized for the project -
https://www.kaggle.com/datasets/ramjidoolla/ipl-data-set

Objectives

IPL started in the year 2008 and takes place every year from the month of March to May.

There are 60-70 matches played every season. This generates a lot of statistics and data in each season. Therefore, the main aim of this project is to design a database that contains all the statistics for IPL from the past seasons in a way that meaningful insights and conclusions can be drawn from those raw statistics. This will benefit a lot of people including the team management to normal cricket fans out there as they would be able to easily categorize the statistics and find out whatever statistics they are seeking.

Users and Requirements

There would be a lot of users for this database. The users could range from the professional managements of the teams for strategizing their next matches or making plans against players to everyday cricket fans, pundits and experts who want to draw meaningful insights about the game and the players.

There is one requirement for retrieving information from the database and that is: -these users should have knowledge of writing queries to retrieve the information from the database.

Source of Data and data ingestion methodology from the source

I will be using the below mentioned dataset from Kaggle.com
(*https://www.kaggle.com/datasets/ramjidoolla/ipl-data-set*)

This dataset contains CSV files which contains a lot of data in the form of

1) List of players who played in the IPL from 2008-2019

2) List of teams who play IPL

3) List of all the deliveries ever bowled in the IPL (contains ~1,80,000) entries

4) List of runs, average and strike rates of players

5) List of victories for all the teams along with bifurcation in the form of

home and away wins

6) List of all the matches ever played in the IPL

7) List of players who became the man-of-match for each match along with the umpires officiating these matches.

8) List of how the players got out and who got the wickets for a particular batsman

Since I will be using data from the CSV files, I will be directly using the copy command for copying the data from CSV files to the database.
I might have to make some changes along with adding ids for some tables which will help in creating key constraints. This will be taken up later as the project progresses.

List of relevant questions from the dataset

1) Finding out the batsman with the highest average among all the teams.
2) Finding out the batsman who were bowled on the first delivery of a match
3) Finding out the bowler, his team name, his dominant bowling hand and strike rate along with the IPL season where he has bowled no more than 100 deliveries and has picked up more than 5 wickets in that season.
4) Finding out the team with second highest away win percentage among all the teams to participate.
5) Finding out the all the players who are "Right-handed batsman and Right-handed bowlers" and who do not belong to "India"
6) Finding out all the details of the matches which ended in a tie played at the "Wankhede stadium".
7) Finding out the DOB, team name , batting average and strike-rate of "V Kohli" for IPL season 2008-2009?
8) Finding out the details of all the matches where man-of-match was "R G Sharma" when batting second and his team won by a margin of at least 1 wicket.
9) Finding out all the matches where player "S Dhawan" played the first ball in first inning and team won the match by a margin of more than 1 run.
10) Finding the player(s) and their number of "players of match" awards who belong to "India" and play for "Mumbai Indians" in the IPL.
11) Finding out the list of team(s) who have away-win-percentage of more than 60% and have won at least one match at the "M Chinnaswamy

Stadium"

12) Finding right handed batsmen with average of more than 30 and has won 3 or more man-of-match awards.

13) Finding the bowler(s), his average, his strike rate and his team's name who has bowled the most and least number of deliveries to "V Kohli"?

14) Finding team(s) who won on "Feroz Shah Kotla" stadium while they decided to field first.

15) Which batsman has bowled 1st ball of the 1st inning and has taken wickets in the 2nd over as well.

16) Finding out all the matches where "Royal challengers Bangalore" lost in the second innings and the losing margin was more than 10 runs.

17) Finding out the team and its players of winner of IPL- season 2017.

18) Finding the team(s) which has won the highest and second highest number of matches from IPL 2008-2012 while fielding first.

19) Finding out the player(s) who played for "Sunrisers Hyderabad" and has the lowest strike rate while bowling "Right handed".

20) Finding out the list of player(s), their batting average and dominant batting hand who were bowled by "TS Mills" from IPL season 2014-2019.

Since the dataset was quite big, I inserted it to the database to check whether it is inserted properly or not and it does not pose any problems as can be checked in the screenshot given below: -

```
spr2022adb35=> select count(*) from deliveries;
 count
--------
 179078
(1 row)

spr2022adb35=> select count(*) from players_ipl;
 count
-------
   566
(1 row)

spr2022adb35=> select count(*) from matches;
 count
-------
   756
(1 row)

spr2022adb35=> select count(*) from teams;
 count
-------
    15
(1 row)
```