IPL Data Analyst Project - SQL Data Preparation & Table Documentation

Section 1: Project Overview

Objective

This project analyzes 15 seasons of IPL data (2008-2022) to uncover performance patterns, team strategies, and match dynamics through SQL analysis and Power BI visualization.

Data Pipeline

1. **Data Acquisition**: CSV files containing match and ball-by-ball records
2. **SQL Preparation**: Table creation and data import
3. **Analysis**: SQL queries for key metrics
4. **Visualization**: Power BI dashboard development

Section 2: Dataset Specifications

| **Dataset** | **Records** | **Time Period** | **Columns** | **Size** |
| --- | --- | --- | --- | --- |
| ipl\_matches\_2008\_2022.csv | 950+ matches | 2008-2022 | 18 | ~200KB |
| ipl\_ball\_by\_ball\_2008\_2022.csv | 225,000+ deliveries | 2008-2022 | 17 | ~25MB |

**Key Entities:**

* Teams
* Players (batters, bowlers)
* Venues
* Match officials
* Section 3: Database Schema Documentation
* Table 1: ipl\_matches\_2008\_2022

| **Column** | **Data Type** | **Description** | **Constraints** |
| --- | --- | --- | --- |
| id | BIGINT | Unique match identifier | PRIMARY KEY |
| city | VARCHAR(50) | Host city |  |
| match\_date | DATE | Date of match (YYYY-MM-DD) |  |
| season | VARCHAR(50) | Tournament year (e.g., "2022") |  |
| match\_number | VARCHAR(50) | Match identifier (e.g., "Final") |  |
| team1/team2 | VARCHAR(50) | Competing teams |  |
| venue | VARCHAR(100) | Stadium name |  |
| toss\_winner | VARCHAR(50) | Team winning toss |  |
| toss\_decision | VARCHAR(50) | 'field' or 'bat' |  |
| superover | VARCHAR(50) | Indicates superover match |  |
| winning\_team | VARCHAR(50) | Match winner |  |
| won\_by | VARCHAR(50) | 'runs' or 'wickets' |  |
| margin | VARCHAR(50) | Victory margin |  |
| method | VARCHAR(50) | DLS method indicator |  |
| player\_of\_match | VARCHAR(50) | Award recipient |  |
| umpire1/umpire2 | VARCHAR(50) | Match officials |  |

Table 2: ipl\_ball\_by\_ball\_2008\_2022

| **Column** | **Data Type** | **Description** | **Relations** |
| --- | --- | --- | --- |
| id | BIGINT | Match identifier | FOREIGN KEY → matches.id |
| innings | BIGINT | Match innings (1-4) |  |
| overs | BIGINT | Over number (1-20) |  |
| ball\_number | BIGINT | Ball in over (1-6) |  |
| batter | VARCHAR(50) | Striker name |  |
| bowler | VARCHAR(50) | Bowler name |  |
| non\_striker | VARCHAR(50) | Non-striker name |  |
| extra\_type | VARCHAR(50) | Type of extra (wide/noball/etc) |  |
| batsman\_run | BIGINT | Runs scored off bat |  |
| extras\_run | BIGINT | Extra runs conceded |  |
| total\_run | BIGINT | Total runs from delivery |  |
| non\_boundry | BIGINT | Non-boundary indicator |  |
| iswicket\_delivery | BIGINT | Wicket indicator (0/1) |  |
| player\_out | VARCHAR(50) | Dismissed batter |  |
| dismisal\_kind | VARCHAR(50) | Dismissal type |  |
| fielders\_involved | VARCHAR(50) | Fielders in dismissal |  |
| batting\_team | VARCHAR(50) | Team currently batting |  |

Section 4: Data Import Protocol

PostgreSQL Import Commands

sql

Copy

Download

*-- Match data import*

COPY ipl\_matches\_2008\_2022

FROM '/path/to/ipl\_matches\_2008\_2022.csv'

DELIMITER ',' CSV HEADER;

*-- Ball-by-ball data import*

COPY ipl\_ball\_by\_ball\_2008\_2022

FROM '/path/to/ipl\_ball\_by\_ball\_2008\_2022.csv'

DELIMITER ',' CSV HEADER;

**Verification Queries:**

sql

Copy

Download

*-- Verify match count (expected: 950+)*

SELECT COUNT(\*) FROM ipl\_matches\_2008\_2022;

*-- Verify delivery count (expected: 225,000+)*

SELECT COUNT(\*) FROM ipl\_ball\_by\_ball\_2008\_2022;

*-- Sample data validation*

SELECT \* FROM ipl\_matches\_2008\_2022 LIMIT 5;

Section 5: Project Directory Structure

text

Copy

Download

ipl-analytics/

├── data/ # Raw data files

│ ├── ipl\_matches\_2008\_2022.csv

│ └── ipl\_ball\_by\_ball\_2008\_2022.csv

│

├── sql/

│ ├── 01\_schema\_setup.sql # Table creation

│ ├── 02\_data\_import.sql # COPY commands

│ └── 03\_analysis\_queries/ # KPI queries

│ ├── batting\_metrics.sql

│ ├── bowling\_metrics.sql

│ └── team\_performance.sql

│

├── documentation/

│ ├── 01\_data\_dictionary.md

│ └── 02\_etl\_process.md

│

├── powerbi/

│ ├── ipl\_dashboard.pbix

│ └── data\_model.docx

│

└── README.md # Project overview

Section 6: Data Quality Considerations

1. **Validation Rules**:
   * All matches should have two valid teams
   * Ball records must reference existing matches
   * Date range constrained to 2008-2022
   * Toss decisions limited to 'field' or 'bat'
2. **Cleaning Recommendations**:

sql

Copy

Download

*-- Standardize team names*

UPDATE ipl\_matches\_2008\_2022

SET team1 = 'Punjab Kings'

WHERE team1 = 'Kings XI Punjab';

1. **Performance Optimization**:

sql

Copy

Download

CREATE INDEX idx\_ball\_match\_id ON ipl\_ball\_by\_ball\_2008\_2022(id);

CREATE INDEX idx\_batter ON ipl\_ball\_by\_ball\_2008\_2022(batter);

Next Steps

Phase 2: Analytical KPIs

**Batting Analysis:**

* Top run scorers (career/seasonal)
* Strike rates by match phase
* Boundary percentages

**Bowling Analysis:**

* Economy rates by bowler
* Death over specialists
* Wicket-taking patterns

**Team Performance:**

* Win/loss ratios by venue
* Toss decision impact
* Head-to-head records

Phase 3: Power BI Development

* Data model relationships
* Interactive filters (season/team/player)
* Performance benchmarking visuals