

Project Title:

**Complex Engineering Activity (CEA)**

Submitted by:

**Samreen Razzaq 2021-CE-13**

**Urwah Imran 2021-CE-15**

Submitted to:

**Mam Darakhshan**

Course:

**CMPE-332: Database Systems Lab**

Semester:

**5th**

Date of Submission:

**26th November 2023**

**Department of Computer Engineering**

**University of Engineering and Technology, Lahore**

Table of Contents

[Dataset 2](#_Toc149923044)

[1. Dataset Overview 3](#_Toc149923045)

[2. Dataset Information 3](#_Toc149923046)

[3. Dataset Structure 3](#_Toc149923047)

[4. Tables Features 3](#_Toc149923048)

[5. Dataset Questions 3](#_Toc149923049)

[References 3](#_Toc149923050)

# 

# Dataset

**Baseball (Bigquery-public-data.baseball)**

# Dataset Overview:

This public data includes pitch-by-pitch data for Major League Baseball (MLB) games in 2016. This dataset contains the following tables: games\_wide (every pitch, steal, or lineup event for each at bat in the 2016 regular season), games\_post\_wide(every pitch, steal, or lineup event for each at-bat in the 2016 post season), and schedules ( the schedule for every team in the regular season). The schemas for the games\_wide and games\_post\_wide tables are identical. With this data we can effectively replay a game and rebuild basic statistics for players and teams.

# Dataset Information:

* **Dataset ID:** bigquery-public-data.baseball
* **Created:** Oct 18, 2016, 10:08:24 PM UTC+5
* **Default table expiration:** Never
* **Last modified:** Sep 20, 2022, 12:42:20 PM UTC+5
* **Data location:** US
* **Update frequency**: Historic (none)
* **Dataset source:** SportRadar
* **Default rounding mode**: ROUNDING\_MODE\_UNSPECIFIED
* **Case insensitive**: false

# Dataset Structure:

This dataset has organized into three tables:

* + games\_post\_wide.
  + games\_wide.
  + schedules.

**Tables Features:**

1. **Games\_post\_wide:**

* **Table ID:** bigquery-public-data.baseball.games\_post\_wide
* **Primary Key:** gameid
* **Foreign Key:** gameid+seasonid
* **Number of rows:** 8,676
* **Total logical bytes:** 20.53 MB
* **Total physical bytes**: 338.76 KB

1. **Games\_wide:**
   * + **Table ID**: bigquery-public-data.baseball.games\_wide
     + **Primary key:** gameid
     + **Foreign key**: gameid+seasoned
     + **Number of rows:** 761,618
     + **Total logical bytes:** 1.76 GB
     + **Total physical bytes:** 32.22 MB
2. **Schedules:**
   * **Table ID** : bigquery-public-data.baseball.schedules
   * **Primary key:** gameid
   * **Foreign key:** homeTeamid+awayTeamid
   * **Number of rows:** 2,431
   * **Total logical bytes**: 569.15 KB
   * **Total physical bytes:** 80 KB

**Questions Regarding BigQuery Dataset:**

Exploring a baseball dataset in BigQuery and applying machine learning involves asking questions that uncover insights about player performance, team dynamics, and game outcomes.

**Player Performance:**

* Who were the top three players with the highest on-base percentage in the 2016 postseason?
* Can you identify any correlation between the number of steals by a player and their batting average during the regular season?
* Which player had the most consistent performance across both the regular season and postseason in terms of home runs?

**Team Dynamics:**

* What is the average win-loss record for teams that had the highest total physical bytes in the games\_wide table?
* Can you identify any patterns in team performance based on the scheduling data? For example, do teams perform differently on certain days of the week?
* Is there a correlation between the total logical bytes of the games\_wide table and the average runs scored by a team in the regular season?

**Game Outcomes:**

* Are there specific game situations, such as a high number of steals or particular lineup events, that correlate with a team's success in the postseason?
* Analyze the duration of games in the postseason compared to the regular season. Are there notable differences in game lengths during these two phases?
* Explore whether the total number of games played by a team in the regular season has any impact on their postseason success.

**Advanced Analytics:**

* Implement a machine learning model to predict player performance metrics (e.g., batting average, home runs) based on the pitch-by-pitch data.
* Can you identify any outliers in team performance based on the provided schema information? What factors might contribute to these outliers?
* Explore the correlation between the homeTeamid and awayTeamid in the schedules table and identify any insights into home-field advantage.

# References

1. https://console.cloud.google.com/bigquery?p=bigquery-public-data&d=baseball&page=dataset&project=invertible-now-398616&ws=!1m13!1m3!3m2!1sbigquery-public-data!2sbaseball!1m4!4m3!1sbigquery-public-data!2sbaseball!3sgames\_post\_wide!1m3!3m2!1sbigquery-public-data!2saustin\_bikeshare