**Project Title:**  
Basketball Analytics – Integration and Big Data Analysis of Basketball Data

**Project Objective:**  
The goal of this project is to develop an application that integrates heterogeneous basketball data and performs Big Data analytics using PySpark. The project will create a global schema that unifies data on players, teams, games, play-by-play events, additional game statistics, and draft information. This unified view will enable advanced querying and analysis, such as evaluating player performance, comparing team statistics, and predicting game outcomes.

**Datasets and Sources:**  
The project is based on a collection of CSV files containing detailed basketball information:

**Link to the dataset:** https://www.kaggle.com/datasets/wyattowalsh/basketball/data

* **Players:**
  + player.csv
  + inactive\_players.csv
  + common\_player\_info.csv
* **Teams:**
  + team.csv
  + team\_details.csv
  + team\_history.csv
  + team\_info\_common.csv
* **Games:**
  + game.csv
  + game\_info.csv
  + game\_summary.csv
* **Play-by-Play Events:**
  + play-by-play.csv
* **Additional Data:**
  + line\_score.csv
  + officials.csv
  + other\_stats.csv
  + draft\_combine\_stats.csv
  + draft\_history.csv

**Integration Method – Global Schema using GAV:**  
I will adopt the **Global-As-View (GAV)** approach, where the global schema is defined as a set of views directly mapped onto the local data sources. The process involves:

1. **Analyzing the Local Schemas:**  
   Examining each CSV file to understand its structure, identifying key attributes (e.g., player\_id, team\_id, game\_id), and understanding the domain.
2. **Identifying Common Entities and Attributes:**  
   Grouping related data to define global entities such as GLOBAL\_PLAYER, GLOBAL\_TEAM, GLOBAL\_GAME, GLOBAL\_PLAY\_BY\_PLAY, etc.
3. **Mapping and Relationships:**  
   For each global entity, defining how its attributes map to the corresponding columns in the local sources. For instance, GLOBAL\_PLAYER will combine data from the player files using player\_id as the key.
4. **Creating a Conceptual Model:**  
   Developing an ER diagram that shows the global entities and the relationships between them.

**Project Phases:**

1. **Global Schema Design:**
   * Create a detailed ER diagram to unify all the tables and highlight the common keys and relationships.
   * Define the global views (using the GAV approach) that map the local data into entities like GLOBAL\_PLAYER, GLOBAL\_TEAM, GLOBAL\_GAME, GLOBAL\_PLAY\_BY\_PLAY, GLOBAL\_LINE\_SCORE, GLOBAL\_OFFICIALS, GLOBAL\_OTHER\_STATS, GLOBAL\_DRAFT\_COMBINE, and GLOBAL\_DRAFT\_HISTORY.
2. **ETL Pipeline Implementation:**
   * Extract data from the CSV files and transform it according to the global schema mappings.
   * Use tools such as Pentaho or Python scripts to build the ETL process.
3. **Big Data Analytics with PySpark:**
   * Load the integrated dataset into PySpark for advanced querying and analysis.
   * Develop models for player performance analysis and game outcome prediction using simple machine learning algorithms.
4. **Application and Dashboard Development:**
   * Build a small application and interactive dashboard to display the integrated data and analysis results.

