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

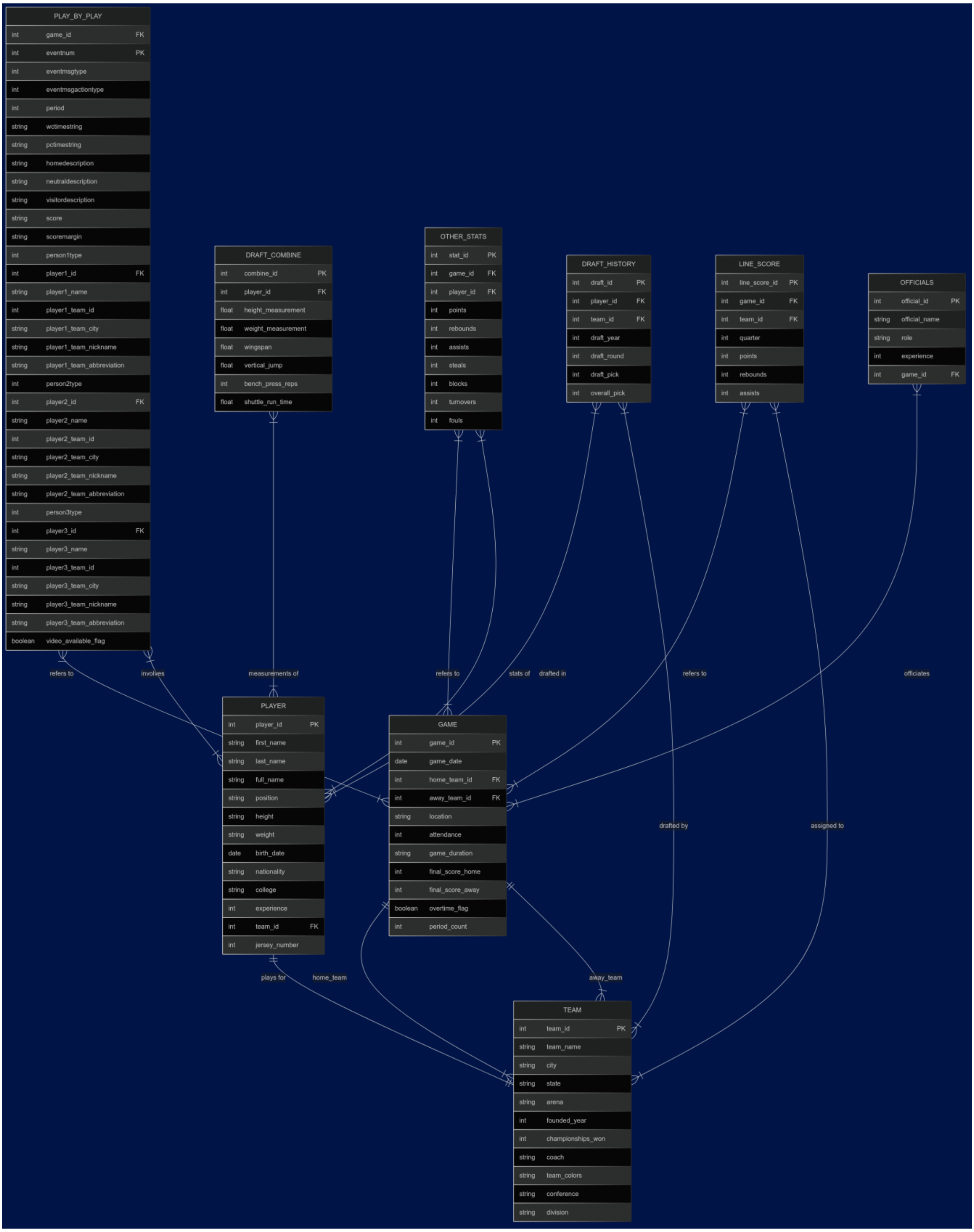
- **Additional Data:**
    - line_score.csv
    - officials.csv
    - other_stats.csv
    - draft_combine_stats.csv
    - draft_history.csv

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.



PLAY_BY_PLAY

int	game_id	FK
int	eventnum	PK
int	eventmsgtype	
int	eventmsgactiontype	
int	period	
string	wctimestring	
string	pctimestring	
string	homedescription	
string	neutraldescription	
string	visitordescription	
string	score	
string	scoremargin	
int	person1type	
int	player1_id	FK
string	player1_name	
int	player1_team_id	
string	player1_team_city	
string	player1_team_nickname	
string	player1_team_abbreviation	
int	person2type	
int	player2_id	FK
string	player2_name	
int	player2_team_id	
string	player2_team_city	
string	player2_team_nickname	
string	player2_team_abbreviation	
int	person3type	
int	player3_id	FK
string	player3_name	
int	player3_team_id	
string	player3_team_city	
string	player3_team_nickname	
string	player3_team_abbreviation	
boolean	video_available_flag	

DRAFT_COMBINE

int	combine_id	PK
int	player_id	FK
float	height_measurement	
float	weight_measurement	
float	wingspan	
float	vertical_jump	
int	bench_press_reps	
float	shuttle_run_time	

OTHER_STATS

int	stat_id	PK
int	game_id	FK
int	player_id	FK
int	points	
int	rebounds	
int	assists	
int	steals	
int	blocks	
int	turnovers	
int	fouls	

DRAFT_HISTORY

int	draft_id	PK
int	player_id	FK
int	team_id	FK
int	draft_year	
int	draft_round	
int	draft_pick	
int	overall_pick	

LINE_SCORE

int	line_score_id	PK
int	game_id	FK
int	team_id	FK
int	quarter	
int	points	
int	rebounds	
int	assists	

OFFICIALS

int	official_id	PK
string	official_name	
string	role	
int	experience	
int	game_id	FK

PLAYER

int	player_id	PK
string	first_name	
string	last_name	
string	full_name	
string	position	
string	height	
string	weight	
date	birth_date	
string	nationality	
string	college	
int	experience	
int	team_id	FK
int	jersey_number	

GAME

int	game_id	PK
date	game_date	
int	home_team_id	FK
int	away_team_id	FK
string	location	
int	attendance	
string	game_duration	
int	final_score_home	
int	final_score_away	
boolean	overtime_flag	
int	period_count	

TEAM

int	team_id	PK
string	team_name	
string	city	
string	state	
string	arena	
int	founded_year	
int	championships_won	
string	coach	
string	team_colors	
string	conference	
string	division	