**NBA Data Integration for Sports Analytics**

Group 15:

## 1) Sai Vinay Teja Jakku (jakku.s@northeastern.edu)

2) Saurabh Shukla (shukla.sa@notheastern.edu)

Objective:

The NBA is one of the most watched sports in the United States with many teams and notable players participating in it. Each year, the league conducts many games in which they compete against

each other to win a championship. All the games produce data that can be used to analyze to improve the games, player and team performances. Some of the interesting questions we intend to follow up is

1) Player Performance Analysis: Analyze the performance of individual players over time to identify trends, strengths, and weaknesses, and make predictions about future performance.

2) Team Performance Analysis: Evaluate the performance of teams in the league over time and identify which teams are likely to perform well in the future.

3) Player Comparison: Compare the performance of individual players across different teams and seasons to identify who the best players in the league are.

4) Game Result Prediction: Predict the outcome of games based on various factors such as player performance, team performance, and historical data.

# Data Source:

We decided to opt for the following datasets shown below<https://www.kaggle.com/datasets/wyattowalsh/basketball>

The updated dataset has data for all the games played in NBA history from 1946 and has 64,000 games data with 4800 players and 30 teams. It has different csv files attributed to different dimensions we intend to create

● We will first design an Extended Entity relationship model and a Relational data model to show a pictorial representation of the data warehouse designed. Once these databases have been visualized using EER and relational depictions, we will then create an

on-premise warehouse with the use of PostgreSQL and Talend.

● Our aim is to answer interesting logistic questions with the data and create reports using Tableau to help make data driven decisions

## The datasets mentioned above comprises of the following dimensions along with the mentioned attributes:

| **Dimensions** | **Attributes** |
| --- | --- |
| Player | Player ID, Player city, Player state, Player zip code. |
| Team | Team ID, Team description, No. of games played, team city |
| Games | Game ID, City, State, Season ID, |
| Draft History | Draft Id, Season ID, team city, player ID, overall pick |
| Player game log | player ID, Game ID, Team ID, No of goals, season ID |
| Season | Season ID, Game ID, City, State, Zip Code. |

# **Extended Entity Relationship Diagram:**

|  |  |
| --- | --- |

**Player**

player\_id, full\_name, first name, last\_name, is\_active

# **Common\_player**

Player\_id, team\_id, first\_name, last\_name, last\_affiliation, season\_experience, roster\_status, team\_name, team\_abbreviation, player\_code, from\_year, to\_year, draft\_number etc.

# **Inactive Players**

First\_name, last\_name, jersey\_no, team\_city, team\_name, team\_abbreviation

# 

# **Game**

game\_ID, game\_date, team\_abbreviation\_home, team\_name\_home, matchup\_home, fga\_home, fg\_pct\_home, fta\_home, oreb\_home, dreb\_home, reb\_home, stl\_home, team\_abbreviation\_away, oreb\_away, dreb\_away, stl\_away etc.

# **Team**

team\_ID, full\_name, abbreviation, nickname, city, state, year, year\_founded etc.

## **Officials**

official\_ID, first\_name, last\_name, jersey\_num etc.

## **Play\_by\_play**

eventnum, eventmsgtype, eventmsgactiontype, period, wctimestring, pctimestring, homedescription, score, scoremargin, player1\_name, player2\_name, player3\_name etc

## **Draft\_history**

player\_name, season, round\_no, round\_pick, overall\_pick, draft\_type etc

# **Relational Database Schema:**

|  |  |
| --- | --- |
|  |  |

**Implementation:**

CREATE TABLE "Common\_Player" ( "Player\_ID" int,

"team\_ID" int, "First\_name" varchar(20), "last\_name" varchar(20), "Birthdate" datetime, "School" varchar(20), "Country" varchar(20),

"Last\_affiliation" varchar(20), "height" varchar(10), "season\_exp" float, "season\_exp" int,

"position" int, "roster\_status" varchar(20),

"games\_played\_current\_season" varchar(1), "team\_name" varchar(20), "team\_abbreviation" varchar(10), "team\_code" varchar(20),

"team\_city" varchar(20), "player\_code" varchar(50), "from\_yr" date,

"to\_yr" date, "draft\_yr" date, "draft\_round" int, "draft\_number" int

);

CREATE INDEX "Fk" ON "Common\_Player" ("Player\_ID", "team\_ID"); CREATE TABLE "player" (

"player\_ID" int,

"full\_name" varchar(50), "first\_name" varchar(50), "last\_name" varchar(50), "is\_active" int

);

CREATE INDEX "Pk" ON "player" ("player\_ID");

CREATE TABLE "Inactive\_players" ( "player\_ID" int,

"team\_id" int, "game\_ID" int, "first\_name" varchar(20), "last\_name" varchar(20), "jersey\_num" int, "team\_city" varchar(20),

"team\_name" varchar(20), "team\_abbreviation" varchar(20)

);

CREATE INDEX "Fk" ON "Inactive\_players" ("player\_ID", "team\_id", "game\_ID");

CREATE TABLE "game" (

"game\_ID" int, "game\_date" datetime, "season\_id" int, "team\_id\_home" int,

"team\_abbreviation\_home" varchar(20), "team\_name\_home" varchar(50), "matchup\_home" varchar(20), "wl\_home" varchar(1),

"min\_home" float, "fgm\_home" float, "fga\_home" float, "fg\_pct\_home" float, "fg3m\_home" float, "fg3a\_home" float, "fg3\_pct\_home" float, "ftm\_home" float, "fta\_home" float, "ft\_pct\_home" float, "oreb\_home" float, "dreb\_home" float, "reb\_home" float, "ast\_home" float, "stl\_home" float, "blk\_home" float, "tov\_home" float, "pf\_home" float, "pts\_home" int, "plus\_minus\_home" int,

"video\_available\_home" int, "team\_ID\_away" int, "team\_abbreviation\_away" varchar(20), "team\_name\_away" varchar(50), "matchup\_away" varchar(20), "wl\_away" varchar(1),

"min\_away" float, "fgm\_away" float, "fga\_away" float, "fg\_pct\_away" float, "fg3m\_away" float, "fg3a\_away" float, "fg3\_pct\_away" float, "ftm\_away" float, "fta\_away" float, "ft\_pct\_away" float, "oreb\_away" float, "dreb\_away" float, "reb\_away" float, "ast\_away" float, "stl\_away" float, "blk\_away" float, "tov\_away" float, "pf\_away" float, "pts\_away" float, "plus\_minus\_away" float,

"video\_avaialable\_away" float

);

CREATE INDEX "Pk" ON "game" ("game\_ID");

CREATE INDEX "Fk" ON "game" ("team\_id\_home", "team\_ID\_away"); CREATE TABLE "Team" (

"team\_ID" int, "full\_name" varchar(50),

"abbreviation" varchar(20), "nickname" varchar(20), "city" varchar(20),

"state" varchar(20), "yearr\_founded" date

);

CREATE INDEX "Pk" ON "Team" ("team\_ID");

CREATE TABLE "Officials" ( "official\_id" int, "game\_id" int, "first\_name" varchar(20), "last\_name" varchar(20), "jersey\_num" int

);

CREATE INDEX "Pk" ON "Officials" ("official\_id"); CREATE INDEX "Fk" ON "Officials" ("game\_id");

CREATE TABLE "play\_by\_play (of only one game)" ( "game\_ID" int,

"eventnum" int, "eventmsgtype" int, "eventmsgactiontype" int, "period" int, "wctimestring" datetime, "pctimestring" datetime,

"homedescription" varchar(50), "score" varchar(20), "scoremargin" int, "person1type" int, "player1\_ID" int, "player1\_name" varchar(20), "player1\_team\_ID" int,

"player1\_team\_city" varchar(20), "player1\_team\_nickname" varchar(20), "player1\_team\_abbreviation" varchar(20), "player2type" int,

"player2\_ID" int, "player2\_name" varchar(20), "player2\_team\_ID" int, "player2\_team\_city" varchar(20),

"player2\_team\_nickname" varchar(20), "player2\_team\_abbreviation" varchar(20), "player3type" int,

"player3\_ID" int, "player3\_name" varchar(20), "player3\_team\_ID" int, "player3\_team\_city" varchar(20),

"player3\_team\_nickname" varchar(20),

"player3\_team\_abbreviation" varchar(20)

);

CREATE INDEX "Fk" ON "play\_by\_play (of only one game)" ("game\_ID", "player1\_ID", "player2\_ID", "player3\_ID");

CREATE TABLE "team\_details" ( "team\_id" int,

"abbreviation" varchar(20), "nickname" varchar(20), "year\_founded" date, "city" varchar(20),

"arena" varchar(50), "arena\_capacity" int, "owner" varchar(50),

"general\_manager" varchar(50), "headcoach" varchar(50), "dleagueaffiliation" varchar(50), "facebook" varchar, "instagram" varchar,

"twitter" varchar

);

CREATE INDEX "Fk" ON "team\_details" ("team\_id"); CREATE TABLE "team\_history" (

"team\_id" int, "city" varchar(20),

"nickname" varchar(20), "year\_founded" date, "year\_active\_till" date

);

CREATE INDEX "Fk" ON "team\_history" ("team\_id"); CREATE TABLE "team\_info" (

"team\_id" int,

"season" varchar(20), "team\_city" varchar(20), "team\_name" varchar(50), "team\_abbreviation" varchar(10), "team\_conference" varchar(10), "team\_division" varchar(20),

"team\_code" varchar(20), "team\_slug" varchar(20), "no\_of\_win" int, "no\_of\_loss" int,

"pct" float, "conf\_rank" int, "div\_rank" int, "min\_year" date, "max\_year" date, "league\_ID" int, "season\_ID" int, "pts\_rank" int, "pts\_pg" float, "reb\_rank" int, "reb\_pg" float, "ast\_rank" int, "ast\_pg" float, "opps\_pts\_rank" int, "opps\_pts\_pg" float

);

CREATE INDEX "Fk" ON "team\_info" ("team\_id"); CREATE TABLE "draft\_history" (

"player\_id" int,

"team\_id" int, "player\_name" varchar(20), "season" int,

"round\_no" int, "round\_pick" int, "overall\_pick" int, "Draft\_type" varchar(20), "team\_city" varchar(20), "team\_name" varchar(20),

"team\_abbreviation" varchar(20), "organization" varchar(20), "organization\_type" varchar(20)

);

CREATE INDEX "Fk" ON "draft\_history" ("player\_id", "team\_id"); CREATE TABLE "draft\_combine\_stats" (

"player\_ID" int, "season" date,

"first\_name" varchar(20), "last\_name" varchar(20), "player\_name" Type, "position" Type, "height\_woshoe\_in" float, "weight" float, "wingspan" float, "standing\_reach" float, "body\_fat\_perc" float, "hand\_length" float, "hand\_width" float,

"standing\_vertical\_leap" float, "max\_vertical\_leap" float, "lane\_agility\_time" float, "modified\_lane\_agility\_time" float, "three\_quarter\_sprint" float, "bench\_press" float

);

# **Warehouse Design Proposal:**

The **Multidimensional schema** proposed for our database can be following:

1. Player (Player ID, Player city, Player state, Player zip code)

2. Team (Team ID, Team description, No. of games played, team city)

3. Games (Game ID, City, State, Season ID….)

4. Draft History (Draft Id, Season ID, team city, player ID, overall pick)

5. Player game log (player ID, Game ID, Team ID, No of points, season ID)

6. Season (Season ID, Game ID, City, State, Zip Code)

## **Measures:**

1. Winning Team
2. Inactive Player Statistics
3. Officials information based on games
4. No. of teams per sub-division

## **Hierarchies:**

1. Player > Common Player > Inactive Players

2. Team > team details > team history

The **Multidimensional schema** proposed for our database can be following:

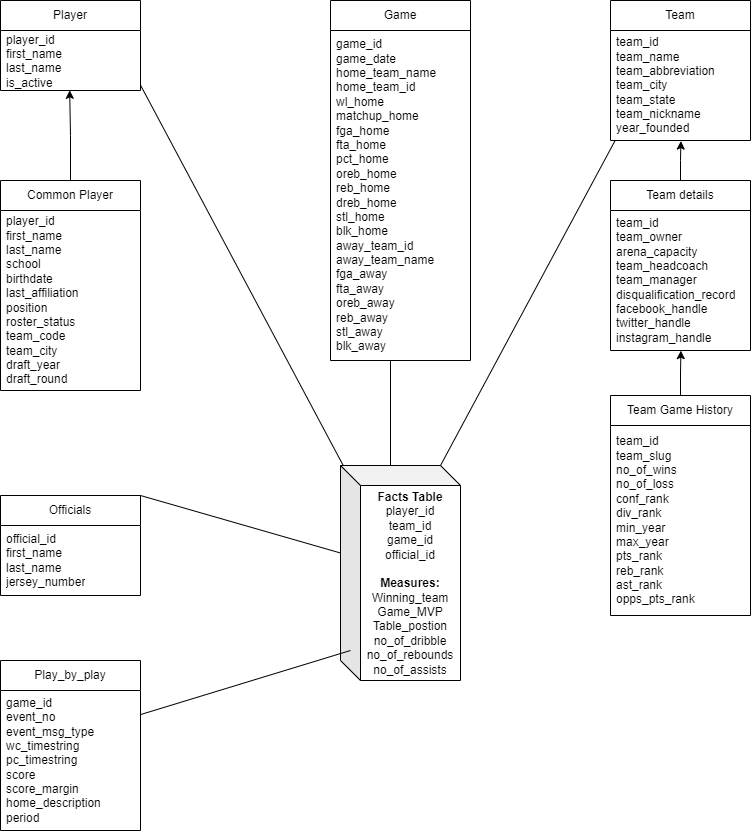
1. Player (player\_id, full\_name, first\_name, last\_name, is\_active)
2. Common Player (player\_id, first\_name, last\_name, birthdate, school, country, last\_affiliation, position, roster\_status, team\_code, team\_city, draft\_yr, draft\_round)
3. Game (game\_id, game\_date, home\_team\_name, home\_team\_id, wl\_home, matchup\_home, wl\_home, fga\_home, fta\_home, ft\_pct\_home, oreb\_home, dreb\_home, reb\_home, ast\_home, stl\_home, blk\_home, pts\_home, away\_team\_id, away\_team\_name, wl\_away, min\_away, fg\_pct\_away, ft\_pct\_away, oreb\_away, reb\_away etc.)
4. Team (team\_id, team\_name, team\_abbreviation, team\_nickname, team\_city, team\_state, year\_founded)
5. Team Details (team\_id, team\_owner, arena\_capacity, team\_manager, team\_headcoach, disqualification\_record, facebook\_handle, instagram\_handle, twitter\_handle)
6. Team Games History (team\_id, team\_slug, no\_of\_wins, no\_of\_loss, conf\_rank, div\_rank, min\_year, max\_year, pts\_rank, reb\_rank, ast\_rank, opps\_pts\_rank)
7. Officials (official\_id, first\_name, last\_name, jersey\_number)
8. Play\_by\_play (game\_id, event\_no, event\_msg\_type, wc\_timestring, pc\_timestring, score, score\_margin, home\_description, period)

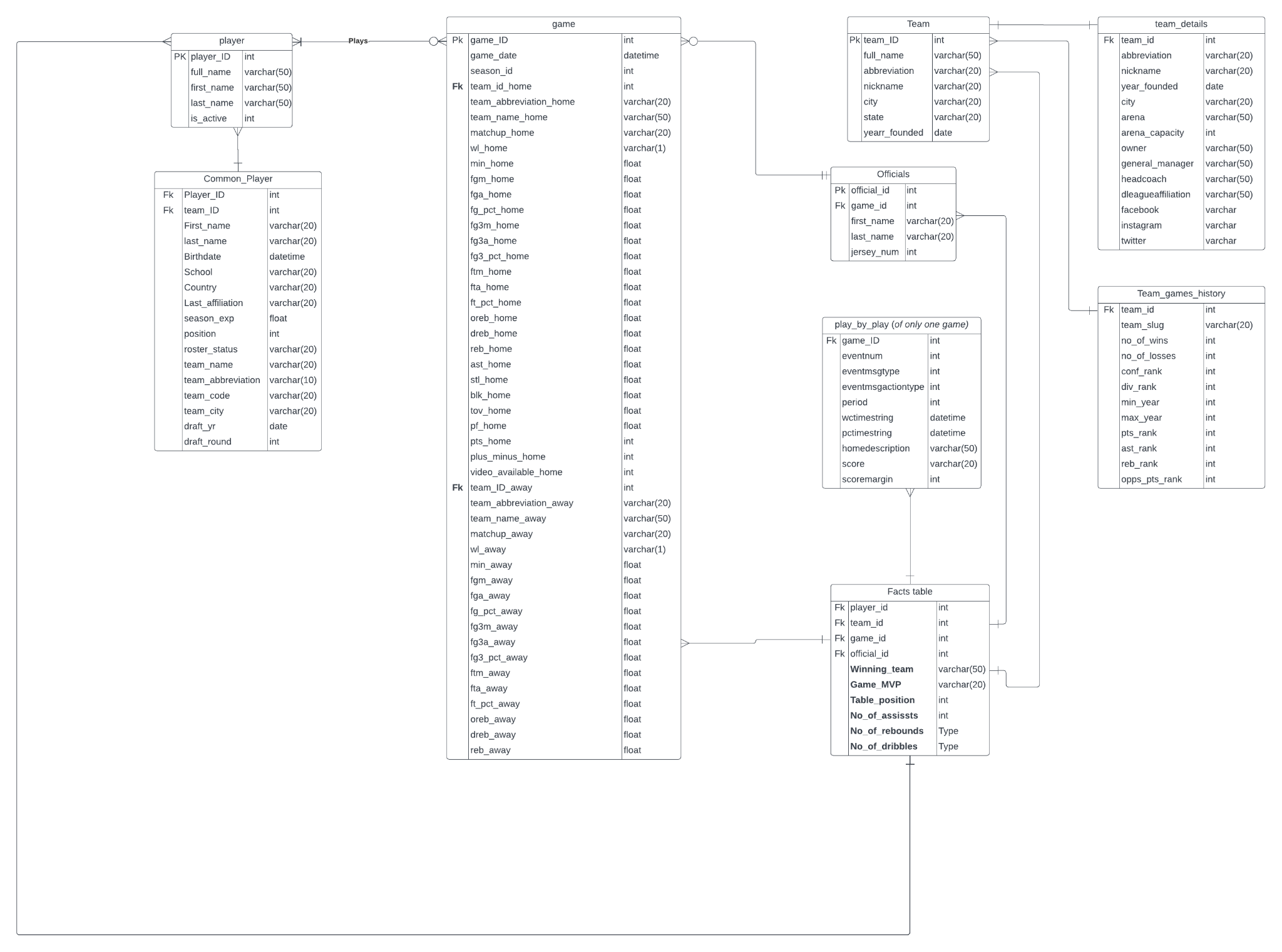
## **Measures:**

1. Winning Team
2. Inactive Player Statistics
3. Officials information based on games
4. No. of teams based per sub-division

**Hierarchies:**

1. Player > Common Player
2. Team > team history

**Multidimensional Conceptual Model Representation:  
  
**

**Multidimensional Logical Model:  
**

**Schema Implementation:**

CREATE TABLE "Facts table" (

"player\_id" int,

"team\_id" int,

"game\_id" int,

"official\_id" int,

"Winning\_team" varchar(50),

"Game\_MVP" varchar(20),

"Table\_position" int,

"No\_of\_assissts" int,

"No\_of\_rebounds" Type,

"No\_of\_dribbles" Type

);

CREATE INDEX "Fk" ON "Facts table" ("player\_id", "team\_id", "game\_id", "official\_id");

CREATE TABLE "Common\_Player" (

"Player\_ID" int,

"team\_ID" int,

"First\_name" varchar(20),

"last\_name" varchar(20),

"Birthdate" datetime,

"School" varchar(20),

"Country" varchar(20),

"Last\_affiliation" varchar(20),

"height" varchar(10),

"season\_exp" float,

"season\_exp" int,

"position" int,

"roster\_status" varchar(20),

"games\_played\_current\_season" varchar(1),

"team\_name" varchar(20),

"team\_abbreviation" varchar(10),

"team\_code" varchar(20),

"team\_city" varchar(20),

"player\_code" varchar(50),

"from\_yr" date,

"to\_yr" date,

"draft\_yr" date,

"draft\_round" int,

"draft\_number" int

);

CREATE INDEX "Fk" ON "Common\_Player" ("Player\_ID", "team\_ID");

CREATE TABLE "player" (

"player\_ID" int,

"full\_name" varchar(50),

"first\_name" varchar(50),

"last\_name" varchar(50),

"is\_active" int

);

CREATE INDEX "Pk" ON "player" ("player\_ID");

CREATE TABLE "game" (

"game\_ID" int,

"game\_date" datetime,

"season\_id" int,

"team\_id\_home" int,

"team\_abbreviation\_home" varchar(20),

"team\_name\_home" varchar(50),

"matchup\_home" varchar(20),

"wl\_home" varchar(1),

"min\_home" float,

"fgm\_home" float,

"fga\_home" float,

"fg\_pct\_home" float,

"fg3m\_home" float,

"fg3a\_home" float,

"fg3\_pct\_home" float,

"ftm\_home" float,

"fta\_home" float,

"ft\_pct\_home" float,

"oreb\_home" float,

"dreb\_home" float,

"reb\_home" float,

"ast\_home" float,

"stl\_home" float,

"blk\_home" float,

"tov\_home" float,

"pf\_home" float,

"pts\_home" int,

"plus\_minus\_home" int,

"video\_available\_home" int,

"team\_ID\_away" int,

"team\_abbreviation\_away" varchar(20),

"team\_name\_away" varchar(50),

"matchup\_away" varchar(20),

"wl\_away" varchar(1),

"min\_away" float,

"fgm\_away" float,

"fga\_away" float,

"fg\_pct\_away" float,

"fg3m\_away" float,

"fg3a\_away" float,

"fg3\_pct\_away" float,

"ftm\_away" float,

"fta\_away" float,

"ft\_pct\_away" float,

"oreb\_away" float,

"dreb\_away" float,

"reb\_away" float,

"ast\_away" float,

"stl\_away" float,

"blk\_away" float,

"tov\_away" float,

"pf\_away" float,

"pts\_away" float,

"plus\_minus\_away" float,

"video\_avaialable\_away" float

);

CREATE INDEX "Pk" ON "game" ("game\_ID");

CREATE INDEX "Fk" ON "game" ("team\_id\_home", "team\_ID\_away");

CREATE TABLE "Team" (

"team\_ID" int,

"full\_name" varchar(50),

"abbreviation" varchar(20),

"nickname" varchar(20),

"city" varchar(20),

"state" varchar(20),

"yearr\_founded" date

);

CREATE INDEX "Pk" ON "Team" ("team\_ID");

CREATE TABLE "Officials" (

"official\_id" int,

"game\_id" int,

"first\_name" varchar(20),

"last\_name" varchar(20),

"jersey\_num" int

);

CREATE INDEX "Pk" ON "Officials" ("official\_id");

CREATE INDEX "Fk" ON "Officials" ("game\_id");

CREATE TABLE "play\_by\_play (of only one game)" (

"game\_ID" int,

"eventnum" int,

"eventmsgtype" int,

"eventmsgactiontype" int,

"period" int,

"wctimestring" datetime,

"pctimestring" datetime,

"homedescription" varchar(50),

"score" varchar(20),

"scoremargin" int,

"person1type" int,

"player1\_ID" int,

"player1\_name" varchar(20),

"player1\_team\_ID" int,

"player1\_team\_city" varchar(20),

"player1\_team\_nickname" varchar(20),

"player1\_team\_abbreviation" varchar(20),

"player2type" int,

"player2\_ID" int,

"player2\_name" varchar(20),

"player2\_team\_ID" int,

"player2\_team\_city" varchar(20),

"player2\_team\_nickname" varchar(20),

"player2\_team\_abbreviation" varchar(20),

"player3type" int,

"player3\_ID" int,

"player3\_name" varchar(20),

"player3\_team\_ID" int,

"player3\_team\_city" varchar(20),

"player3\_team\_nickname" varchar(20),

"player3\_team\_abbreviation" varchar(20)

);

CREATE INDEX "Fk" ON "play\_by\_play (of only one game)" ("game\_ID", "player1\_ID", "player2\_ID", "player3\_ID");

CREATE TABLE "team\_details" (

"team\_id" int,

"abbreviation" varchar(20),

"nickname" varchar(20),

"year\_founded" date,

"city" varchar(20),

"arena" varchar(50),

"arena\_capacity" int,

"owner" varchar(50),

"general\_manager" varchar(50),

"headcoach" varchar(50),

"dleagueaffiliation" varchar(50),

"facebook" varchar,

"instagram" varchar,

"twitter" varchar

);

CREATE INDEX "Fk" ON "team\_details" ("team\_id");

CREATE TABLE "team\_games\_history" (

"team\_id" int,

"season" varchar(20),

"team\_city" varchar(20),

"team\_name" varchar(50),

"team\_abbreviation" varchar(10),

"team\_conference" varchar(10),

"team\_division" varchar(20),

"team\_code" varchar(20),

"team\_slug" varchar(20),

"no\_of\_win" int,

"no\_of\_loss" int,

"pct" float,

"conf\_rank" int,

"div\_rank" int,

"min\_year" date,

"max\_year" date,

"league\_ID" int,

"season\_ID" int,

"pts\_rank" int,

"pts\_pg" float,

"reb\_rank" int,

"reb\_pg" float,

"ast\_rank" int,

"ast\_pg" float,

"opps\_pts\_rank" int,

"opps\_pts\_pg" float

);

CREATE INDEX "Fk" ON "team\_info" ("team\_id");

**OLAP Queries:**

1. Total no of games played by each team in the current season:

SELECT Team.team\_name, Team.game\_history\_no\_of\_wins + Team.game\_history\_no\_of\_loss AS Total\_Games\_Played

FROM Team

WHERE Team.game\_history\_max\_year = 2023

1. Total number of wins & losses for each team in current season:

SELECT Team.team\_name, Team.game\_history\_no\_of\_wins, Team.game\_history\_no\_of\_loss

FROM Team

WHERE Team.game\_history\_max\_year = 2023

1. Total no. of games played by each player in current season:

SELECT Common\_player.first\_name, Common\_player.last\_name, COUNT(\*) AS Total\_Games\_Played

FROM Fact\_table

JOIN Common\_player ON Fact\_table.player\_id = Common\_player.player\_id

JOIN Game ON Fact\_table.game\_id = Game.game\_id

WHERE YEAR(Game.game\_date) = 2023

GROUP BY Common\_player.first\_name, Common\_player.last\_name

1. Total points scored by each team in current season:

SELECT Team.team\_name, SUM(Game.fga\_home \* Game.pct\_home + Game.fga\_away \* Game.pct\_away) AS Total\_Points\_Scored

FROM Team

JOIN Game ON Team.team\_id = Game.home\_team\_id OR Team.team\_id = Game.away\_team\_id

WHERE Game.game\_date BETWEEN '2023-01-01' AND '2023-12-31'

GROUP BY Team.team\_name

1. Total assists made by each player in current season:

SELECT Common\_player.first\_name, Common\_player.last\_name, COUNT(\*) AS Total\_Assists

FROM Fact\_table

JOIN Common\_player ON Fact\_table.player\_id = Common\_player.player\_id

JOIN Play\_by\_play ON Fact\_table.game\_id = Play\_by\_play.game\_id AND Fact\_table.table\_position = Play\_by\_play.home\_description

WHERE YEAR(Game.game\_date) = 2023 AND Play\_by\_play.event\_msg\_type = 'assist'

GROUP BY Common\_player.first\_name, Common\_player.last\_name

1. Total no. of fouls committed by each official in current season:

SELECT Officials.first\_name, Officials.last\_name, COUNT(\*) AS Total\_Fouls\_Committed

FROM Fact\_table

JOIN Officials ON Fact\_table.official\_id = Officials.official\_id

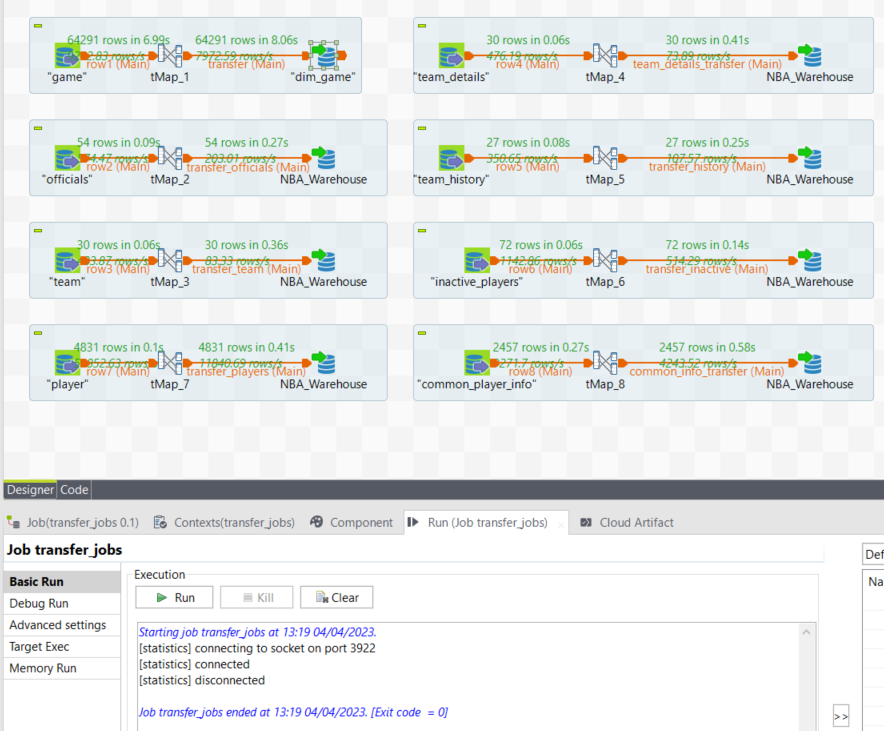
JOIN Play\_by\_play ON Fact\_table.game\_id = Play\_by\_play.game\_id AND Fact\_table.table\_position = Play\_by\_play.home\_description

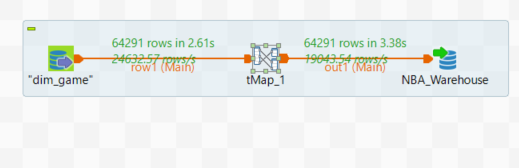
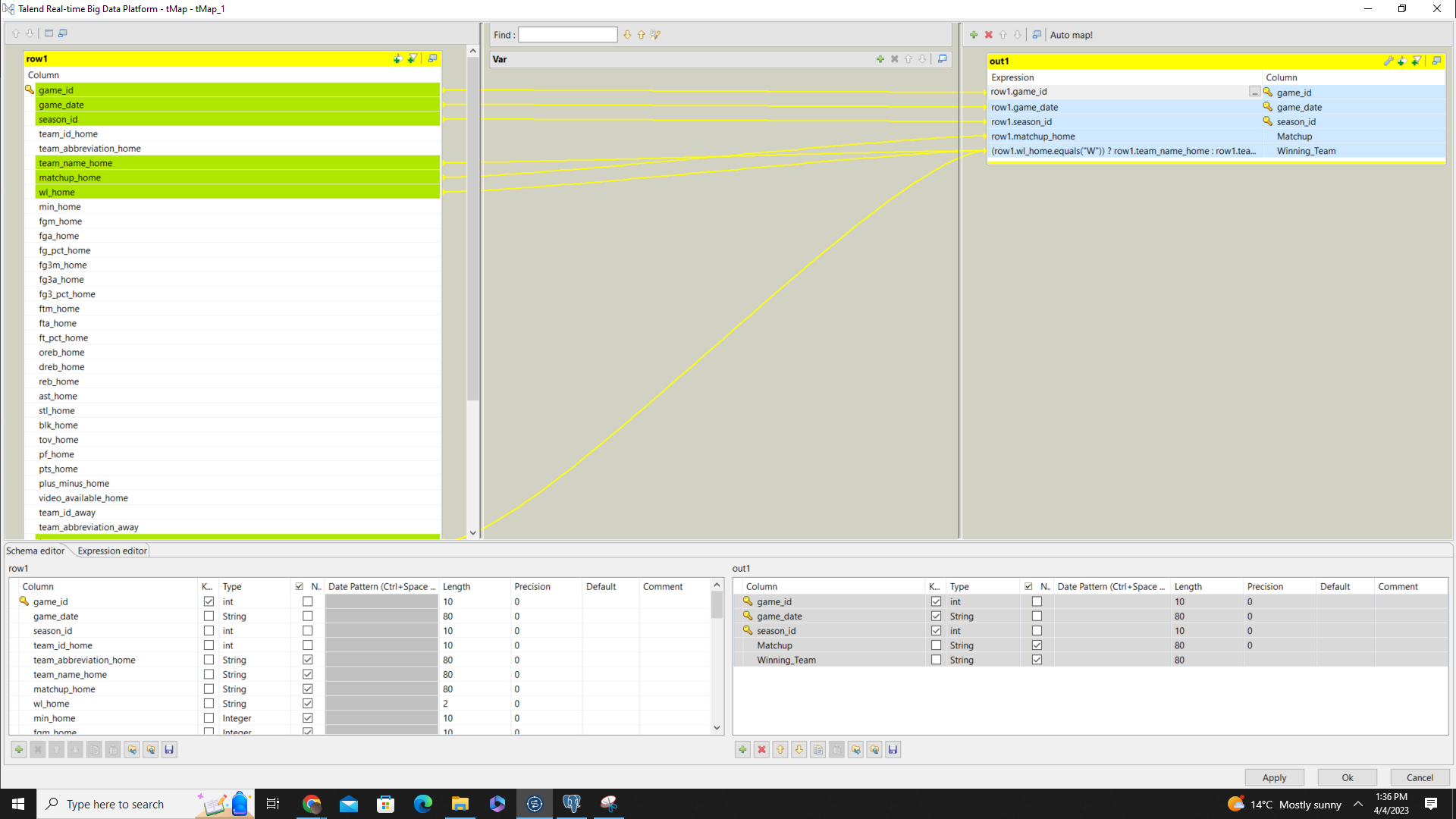
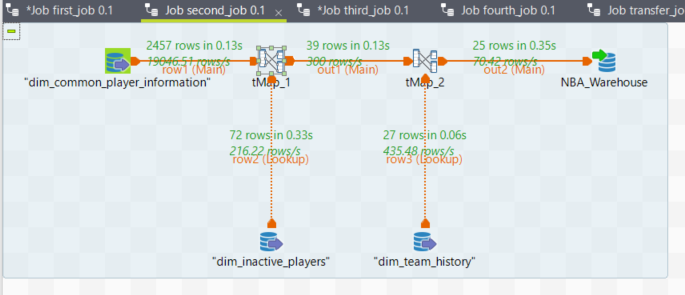
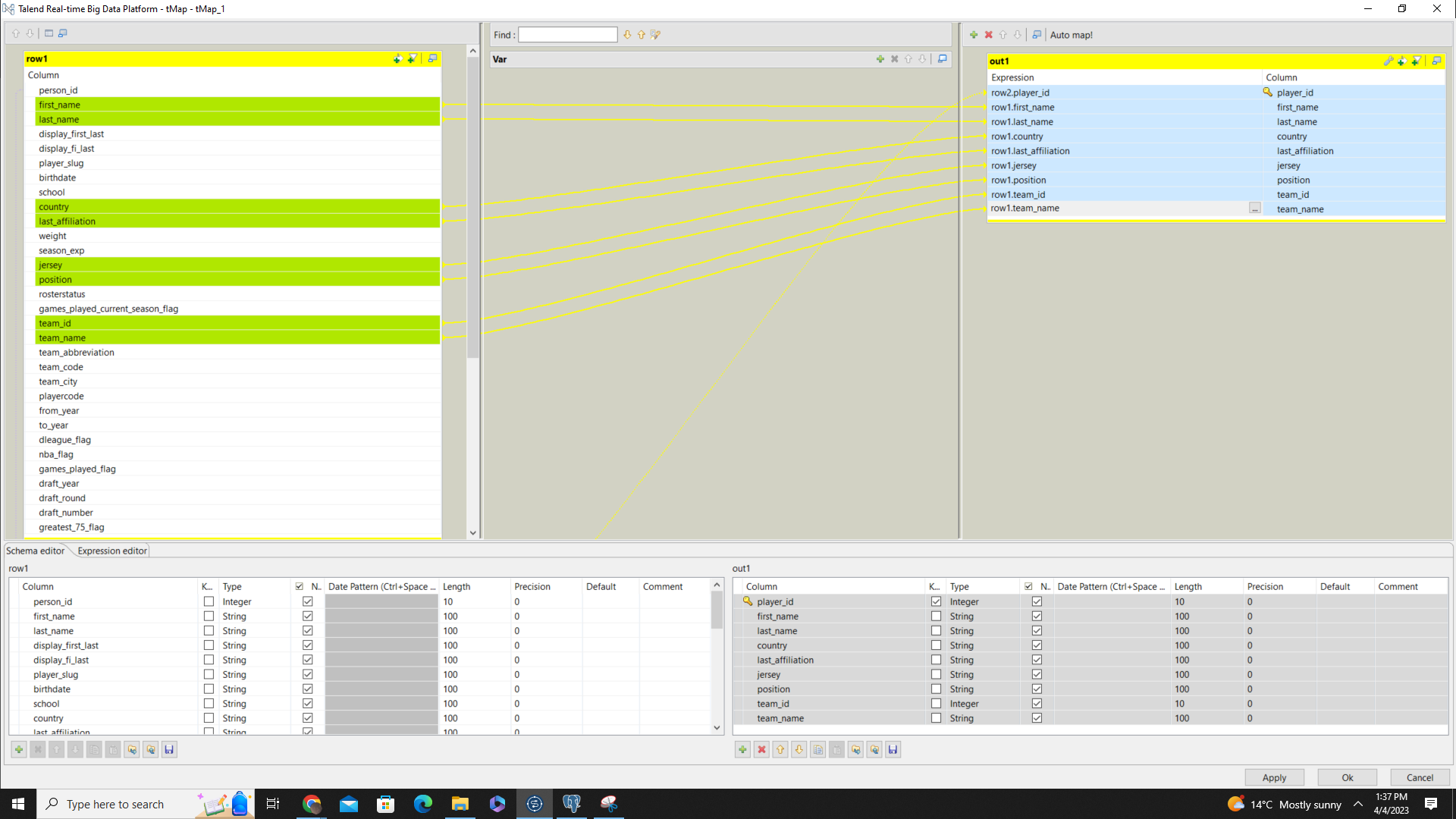
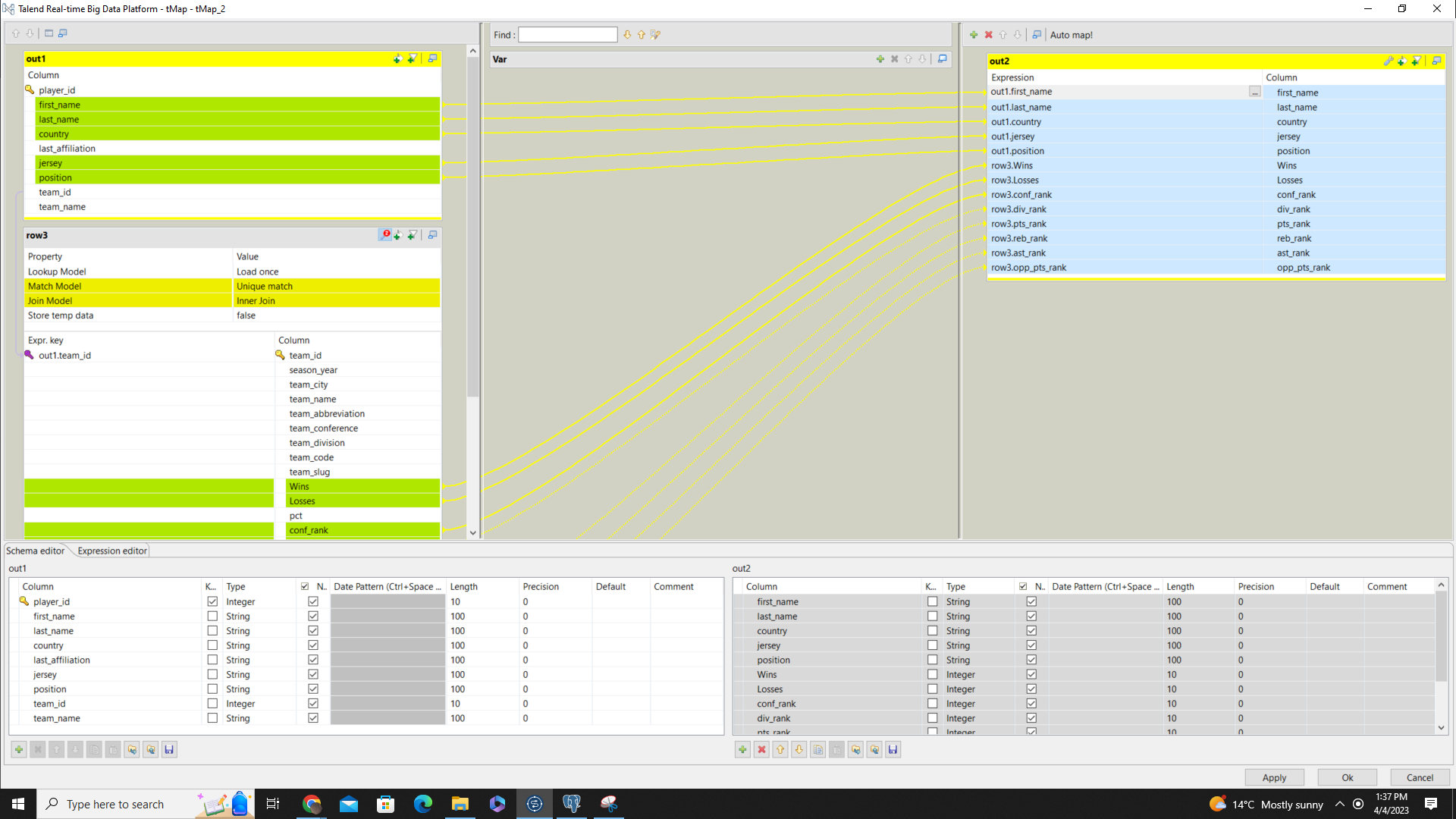
WHERE YEAR(Game.game\_date) = 2023 AND Play\_by\_play.event\_msg\_type = 'foul'

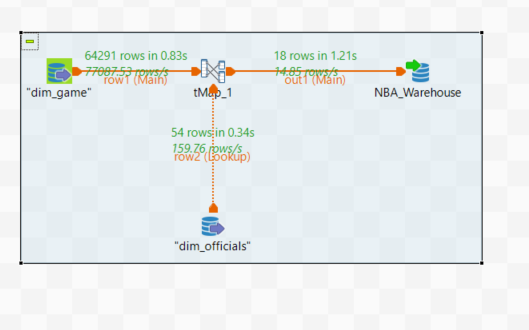
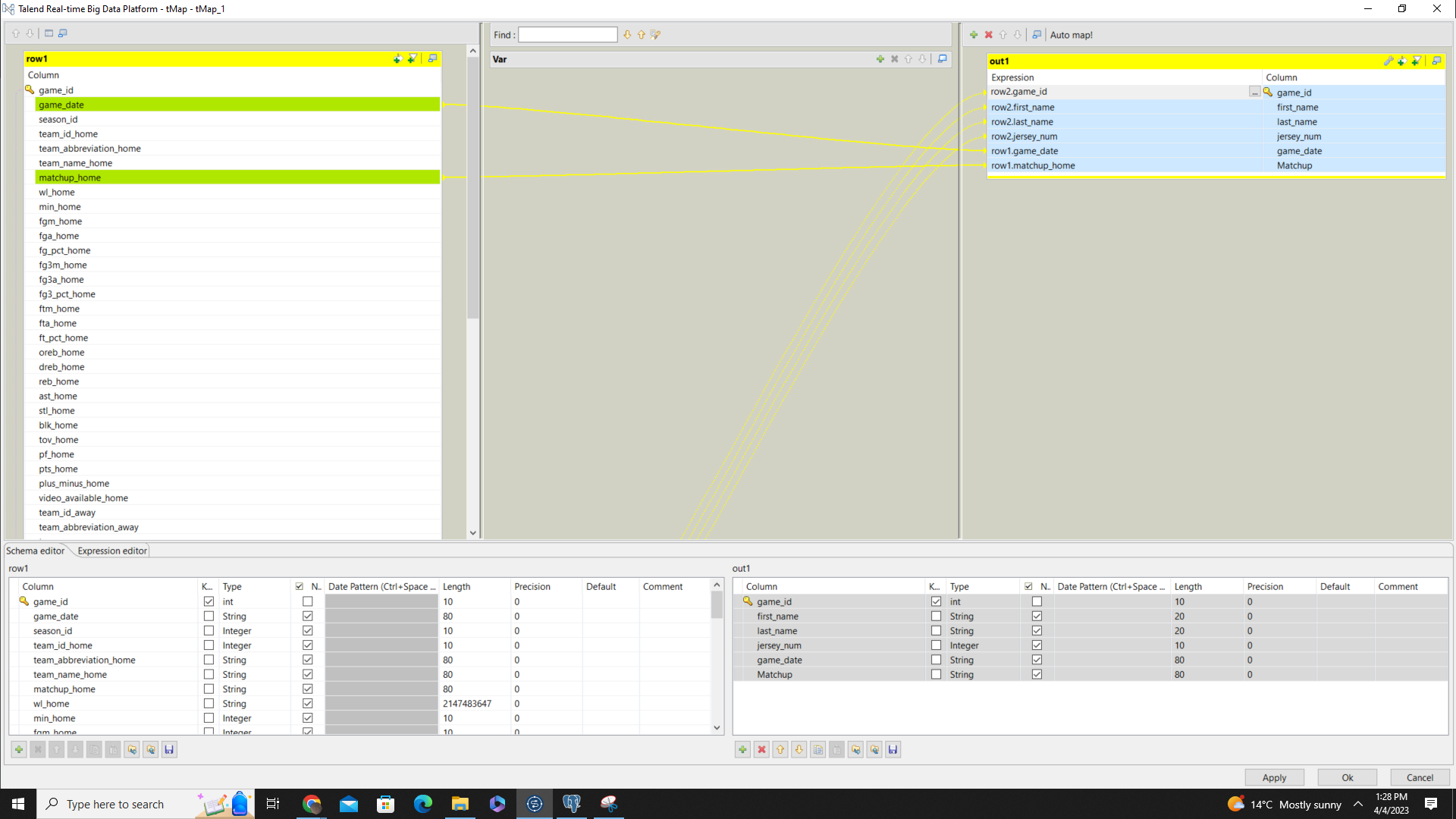
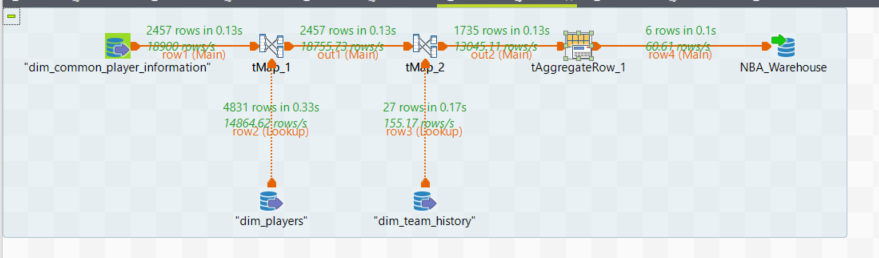
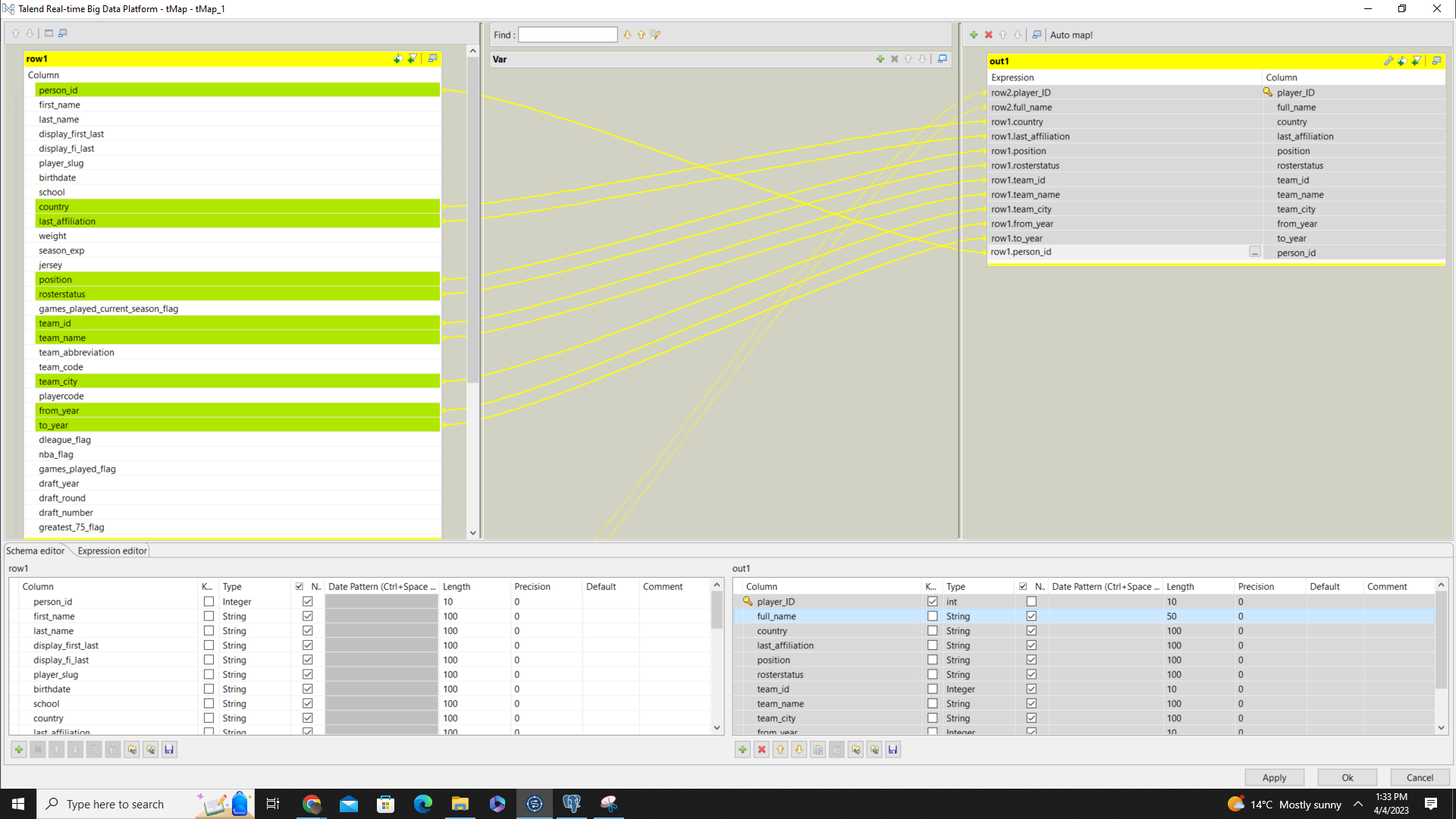
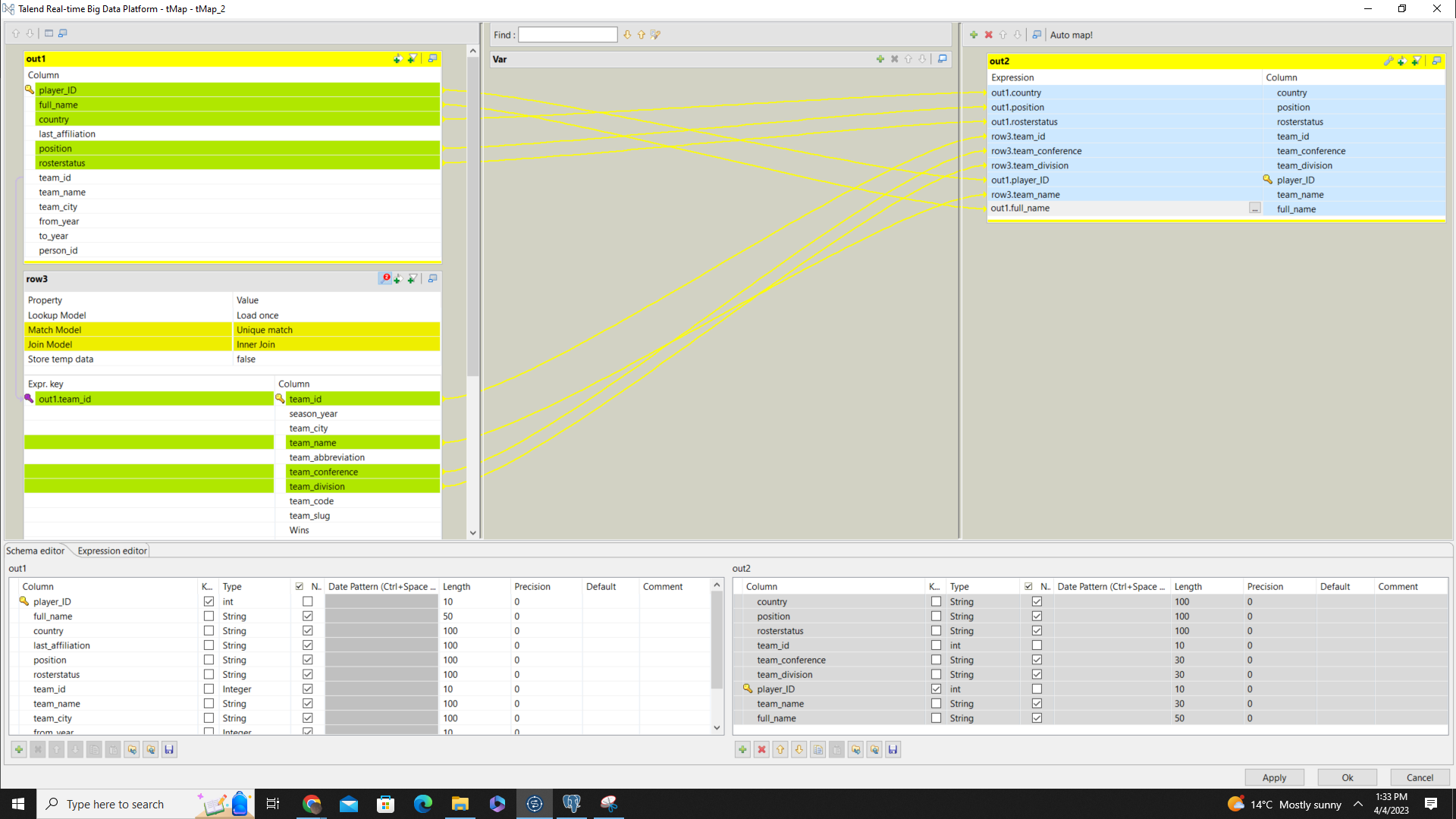
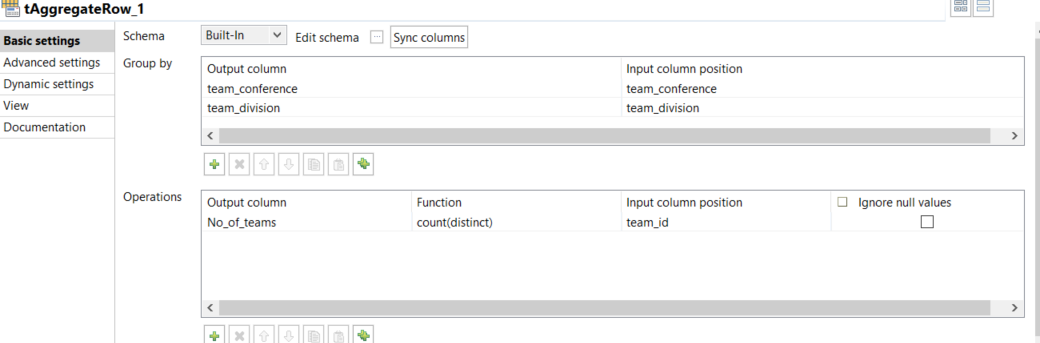
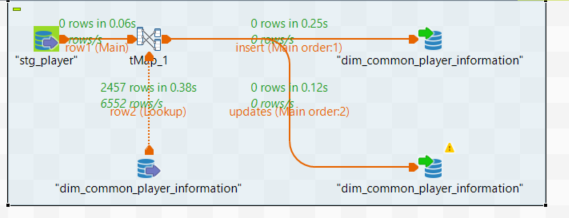
GROUP BY Officials.first\_name, Officials.last\_name

Milestone - 4

* Extracting the data from the csv files to create dimensions for our proposed data warehouse:



* Identifying and retrieving the winning team from the game dimension and storing it in the facts table..  
    
  
* Retrieval of statistics related to the players present in the current roster but are inactive due to injuries or other reasons.  
    
    
  

* Identification and retrieval of officials information and the games officiated by them.  
    
  
* Calculation of the number of teams that are present in each subdivision and grouped by the conference they belong to.  
    
    
    
  
* **Updating Records (Slowly changing Dimension - Type 2)**:  
  

We have taken into account the possibility that the player information changes over time and created a job to accommodate those changes. We have added sequence queries in postgres for stg\_player, incremented them and changed it to pk so that the new table can hold two records of the same player id. If the lookup is successful, it will duplicate the existing record and create a new record with new flags or else it will insert the new data in sequential manner