**Ảnh có chứa quảng trường

Mô tả được tạo tự độngĐẠI HỌC QUỐC GIA TP HCM**

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

🙞🕮🙝



**ĐỒ ÁN**

**Môn học: KHO DỮ LIỆU VÀ OLAP**

**Đề tài: Phân tích dữ liệu cầu thủ bóng rổ**

**từ năm 1999 đến năm 2020**

Giáo viên hướng dẫn: **Đỗ Thị Minh Phụng**

Nhóm thực hiện: **Nhóm 7**

Thành Viên:

Nguyễn Cao Khoa - 19521694

Hồ Trọng Khang - 19521661

**TP Hồ Chí Minh tháng 3 năm 2022**

**LỜI CÁM ƠN**

Nhóm xin chân thành cảm ơn cô Đỗ Thị Minh Phụng - giảng viên lý thuyết Kho dữ liệu và OLAP đã nhiệt tình giúp đỡ, chỉ bảo và hướng dẫn tận tình trong suốt thời gian thực hiện đồ án môn học. Nhờ đó, chúng em đã có được nhiều kiến ​​thức bổ ích trong việc ứng dụng cũng như kỹ năng làm đồ án.

Bắt đầu từ việc tìm hiểu mục đích của kho dữ liệu, phương pháp xây dựng kho dữ liệu và phân tích dữ liệu trên kho dữ liệu của cầu thủ bóng rổ. Dựa trên những kiến ​​thức được giáo viên cung cấp trên lớp, kết hợp với việc tự nghiên cứu các công cụ và kiến ​​thức mới, nhóm cố gắng thực hiện dự án một cách tốt nhất có thể. Trong một học kỳ thực hiện dự án, nhóm chúng em đã sử dụng kiến ​​thức nền tảng đã tích lũy được, kết hợp học tập và nghiên cứu kiến ​​thức mới. Từ đó, nhóm của chúng em sử dụng đầy đủ các thông tin thu thập được để đưa ra báo cáo dự án tốt nhất có thể. Tuy nhiên, trong quá trình thực hiện, nhóm chúng em không tránh khỏi những thiếu sót. Vì vậy, nhóm rất mong nhận được sự góp ý của các cô để nhóm em hoàn thiện hơn kiến ​​thức và chuẩn bị cho các đề tài khác trong tương lai.

Sau cùng, nhóm xin kính chúc cô thật dồi dào sức khỏe, niềm tin để tiếp tục thực hiện sứ mệnh cao đẹp là truyền đạt kiến thức cho các bạn sinh viên.

Xin chân thành cảm ơn Cô!

**MỤC LỤC**

# **CHƯƠNG I: GIỚI THIỆU TỔNG QUAN VỀ DỮ LIỆU**

## **PHÁT BIỂU VỀ DỮ LIỆU**

* 1. **Lý do chọn đề tài**
* Nói về thể thao thì bóng đá được coi là môn thể thao hấp dẫn nhất hành tinh và được coi là môn thể thao vua nhưng bên cạnh đó bóng rổ cũng không kém phần kịch tính. Hiện nay bóng rổ đã phổ biến rộng rãi khắp nơi trên thế giới. Ngoài được xem là một môn thể thao về mặt giải trí nó còn tạo ra nguồn kinh tế rất lớn vì mỗi đội bóng thi đấu chuyên nghiệp và thi đấu càng tốt sẽ được nguồn lợi càng cao. Nhưng một đội bóng hay thì phải có những cầu thủ giỏi và những đội bóng sẽ cần chuyển nhượng hoặc tuyển chọn những cầu thủ tiềm năng về đội nhà.
* Xác định được vấn đề chính, nhóm nhận thấy việc phân tích các thông số của một cầu thủ là cần thiết. Từ các phân tích này các nhà tuyển chọn sẽ có cái nhìn tổng thể nhất của một tuyển thủ từ đó có thể lên mục tiêu chuyển nhượng tốt nhất và mang lại nhiều lợi ích cho đội bóng.

### **Mô tả về dữ liệu**

* Kho dữ liệu Basketball Player Stats per Season là một kho dữ liệu thu thập các thông tin về các cầu thủ bóng rổ (thông tin cá nhân, thành tích,…) qua từng mùa giải tại các giải đấu trên thế giới (từ năm 1999-2020).
* Thông qua kho dữ liệu người dung có thể biết được các thông tin:
  + Thông tin của 49 giải đấu mùa giải từ 1999-2020.
  + Thông tin của gần 11 ngàn cầu thủ.
  + Thống kê số liệu cầu thủ theo mỗi trận đấu trong mỗi mùa giải.
* Kho dữ liệu gồm 53950 dòng và 34 thuộc tính.
* Link dữ liệu gốc: [Basketball Players Stats per Season - 49 Leagues | Kaggle](https://www.kaggle.com/jacobbaruch/basketball-players-stats-per-season-49-leagues).
  1. **Thuộc tính kho dữ liệu**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu** | **Ý nghĩa thuộc tính** |
| 1 | League | String | Tên của giải bóng rổ  Có 49 giải |
| 2 | Season | String | Năm diễn ra của mùa giải  Từ năm 1999-2020 |
| 3 | Stage | String | Tên thể thức của giải  Có 3 thể thức:   1. Regular\_Season: Mùa giải chính.   (chỉ có tại NBA)   1. Playoffs: Vòng đấu cuối mùa.   (chỉ có tại NBA)   1. International: Vòng đấu quốc tế.   (cho các giải còn lại) |
| 4 | Player | String | Tên của cầu thủ |
| 5 | Team | String | Tên viết tắt đội bóng của cầu thủ |
| 6 | GP | Integer | Game Played: Số trận cầu thủ đã chơi. |
| 7 | MIN | Decimal | Minutes Played: Số phút mà cầu thủ đã chơi. |
| 8 | FGM | Integer | Field Goals Made: Số cú ném ghi điểm của cầu thủ bao gồm cú ném 2 và 3 điểm nhưng không tính ném phạt. |
| 9 | FGA | Integer | Field Goals Attempts: Số cú ném của cầu thủ đã thực hiện bao gồm cú ném 2 và 3 nhưng không tính ném phạt. |
| 10 | 3PM | Integer | Three Points Made: Số cú ném 3 điểm mà cầu thủ ghi được. |
| 11 | 3PA | Integer | Three Points Attempts: Số cú ném 3 điểm mà cầu thủ thực hiện. |
| 12 | FTM | Integer | Free Throws Made: Số cú ném phạt mà cầu thủ ghi được. |
| 13 | FTA | Integer | Free Throws Attempts: Số cú ném phạt mà cầu thủ thực hiện. |
| 14 | TOV | Integer | TurnOvers: Số lần mất bóng trước khi thực hiện cú ném của cầu thủ. |
| 15 | PF | Integer | Personal Fouls: Số lỗi cá nhân của cầu thủ. |
| 16 | ORB | Integer | Offensive Rebounds: Số lần cầu thủ bật nhảy để bắt lại bóng sau khi cá nhân hoặc đồng đội ném rổ hụt khi tấn công. |
| 17 | DRB | Integer | Defensive Rebounds: Số lần cầu thủ bật nhảy bắt lại bóng để phản công khi đội đang phòng thủ. |
| 18 | REB | Integer | Rebounds: Tổng số lần cầu thủ bật nhảy bắt lại bóng bao gồm cả tấn công và phòng thủ. |
| 19 | AST | Integer | Assists: Số lần hỗ trợ hay còn gọi là kiến tạo cho đồng đội ghi điểm. |
| 20 | STL | Integer | Steals: Số lần cướp bóng từ đối thủ. |
| 21 | BLK | Integer | Blocks: Số lần chắn bóng trên không. |
| 22 | PTS | Integer | Points: Tổng số điểm cầu thủ ghi được. |
| 23 | Birth\_year | Integer | Năm sinh của cầu thủ |
| 24 | Birth\_month | String | Tháng sinh của cầu thủ |
| 25 | Birth\_date | String | Ngày tháng năm sinh của cầu thủ |
| 26 | Height | String | Chiều cao của cầu thủ tính theo Feet |
| 27 | Height\_cm | Integer | Chiều cao của cầu thủ tính theo Cm |
| 28 | Weight | Integer | Cân nặng của cầu thủ tính theo Pound |
| 29 | Weight\_kg | Integer | Cân nặng của cầu thủ tính theo Kg |
| 30 | Nationality | String | Quốc tịch của cầu thủ |
| 31 | High\_school | String | Tên trường trung học của cầu thủ |
| 32 | Draft\_round | Integer | Vòng mà cầu thủ được chọn vào đội đầu tiên của giải chuyên nghiệp |
| 33 | Draft\_pick | Integer | Thứ tự mà cầu thủ được chọn vào đội đầu tiên của giải chuyên nghiệp |
| 34 | Draft\_team | String | Tên đội đã chọn cầu thủ trong kì tuyển chọn |

Ảnh có chứa bàn

Mô tả được tạo tự động

*Hình 1.1 Kho dữ liệu khi chưa xử lý*

* 1. **Kho dữ liệu đã được xử lý**
* Sau khi lọc dữ liệu ta được 53950 dòng và 30 thuộc tính.

**Ảnh có chứa bàn

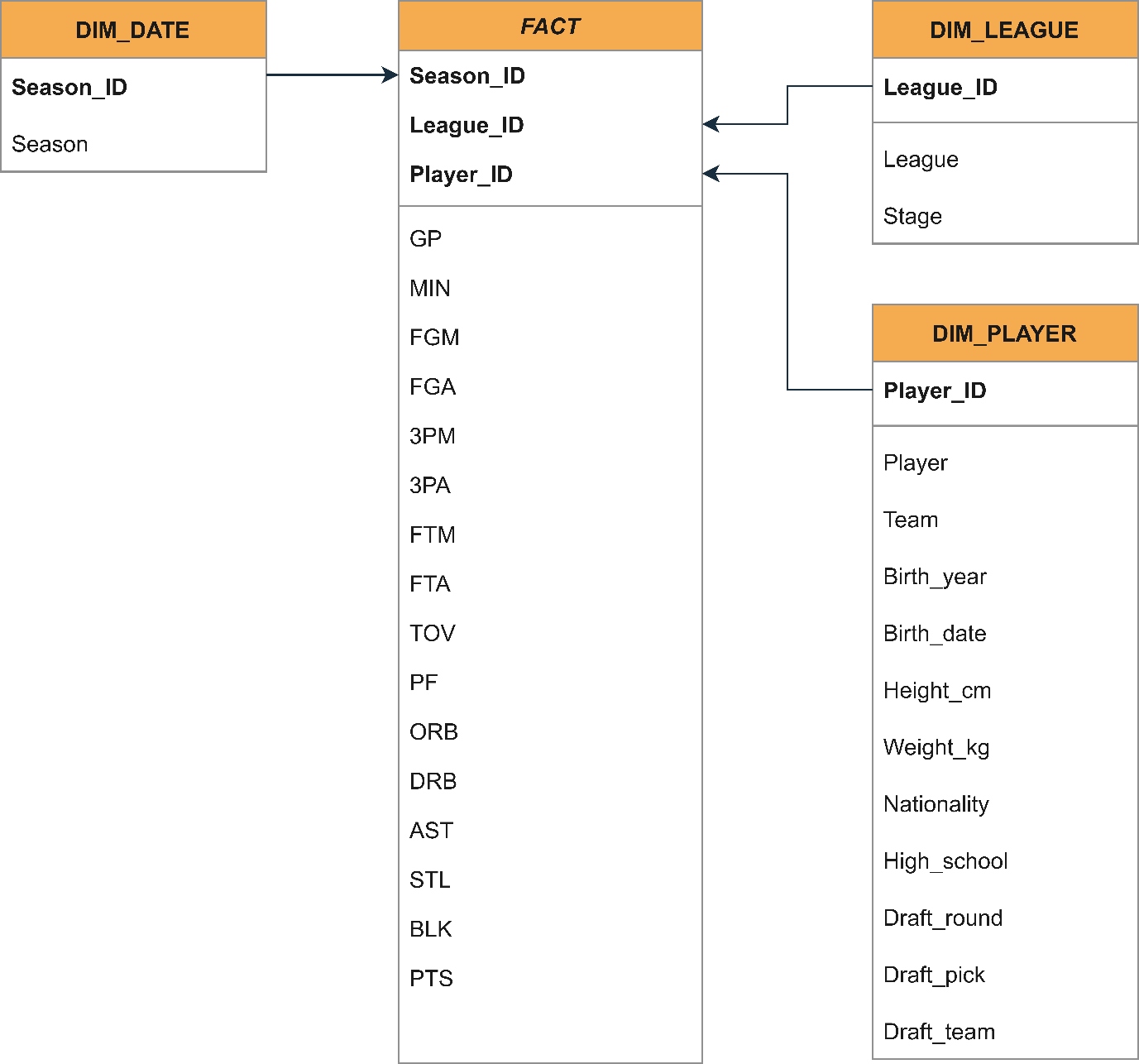
Mô tả được tạo tự động**

*Hình 1.2 Kho dữ liệu đã được lọc*

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu** | **Ý nghĩa thuộc tính** |
| 1 | League | String | Tên của giải bóng rổ  Có 49 giải |
| 2 | Season | String | Năm diễn ra của mùa giải  Từ năm 1999-2020 |
| 3 | Stage | String | Tên thể thức của giải  Có 3 thể thức:   1. Regular\_Season: Mùa giải chính.   (chỉ có tại NBA)   1. Playoffs: Vòng đấu cuối mùa.   (chỉ có tại NBA)   1. International: Vòng đấu quốc tế.   (cho các giải còn lại) |
| 4 | Player | String | Tên của cầu thủ |
| 5 | Team | String | Tên viết tắt đội bóng của cầu thủ |
| 6 | GP | Integer | Game Played: Số trận cầu thủ đã chơi. |
| 7 | MIN | Decimal | Minutes Played: Số phút mà cầu thủ đã chơi. |
| 8 | FGM | Integer | Field Goals Made: Số cú ném ghi điểm của cầu thủ bao gồm cú ném 2 và 3 điểm nhưng không tính ném phạt. |
| 9 | FGA | Integer | Field Goals Attempts: Số cú ném của cầu thủ đã thực hiện bao gồm cú ném 2 và 3 nhưng không tính ném phạt. |
| 10 | 3PM | Integer | Three Points Made: Số cú ném 3 điểm mà cầu thủ ghi được. |
| 11 | 3PA | Integer | Three Points Attempts: Số cú ném 3 điểm mà cầu thủ thực hiện. |
| 12 | FTM | Integer | Free Throws Made: Số cú ném phạt mà cầu thủ ghi được. |
| 13 | FTA | Integer | Free Throws Attempts: Số cú ném phạt mà cầu thủ thực hiện. |
| 14 | TOV | Integer | TurnOvers: Số lần mất bóng trước khi thực hiện cú ném của cầu thủ. |
| 15 | PF | Integer | Personal Fouls: Số lỗi cá nhân của cầu thủ. |
| 16 | ORB | Integer | Offensive Rebounds: Số lần cầu thủ bật nhảy để bắt lại bóng sau khi cá nhân hoặc đồng đội ném rổ hụt khi tấn công. |
| 17 | DRB | Integer | Defensive Rebounds: Số lần cầu thủ bật nhảy bắt lại bóng để phản công khi đội đang phòng thủ. |
| 18 | AST | Integer | Assists: Số lần hỗ trợ hay còn gọi là kiến tạo cho đồng đội ghi điểm. |
| 19 | STL | Integer | Steals: Số lần cướp bóng từ đối thủ. |
| 20 | BLK | Integer | Blocks: Số lần chắn bóng trên không. |
| 21 | PTS | Integer | Points: Tổng số điểm cầu thủ ghi được. |
| 22 | Birth\_year | Integer | Năm sinh của cầu thủ. |
| 23 | Birth\_date | String | Ngày tháng năm sinh của cầu thủ. |
| 24 | Height\_cm | Integer | Chiều cao của cầu thủ tính theo Cm. |
| 25 | Weight\_kg | Integer | Cân nặng của cầu thủ tính theo Kg. |
| 26 | Nationality | String | Quốc tịch của cầu thủ. |
| 27 | High\_school | String | Tên trường trung học của cầu thủ. |
| 28 | Draft\_round | Integer | Vòng mà cầu thủ được chọn vào đội đầu tiên của giải chuyên nghiệp. |
| 29 | Draft\_pick | Integer | Thứ tự mà cầu thủ được chọn vào đội đầu tiên của giải chuyên nghiệp. |
| 30 | Draft\_team | String | Tên đội đã chọn cầu thủ trong kì tuyển chọn. |

## **XÂY DỰNG KHO DỮ LIỆU**

* 1. **Lược đồ kho dữ liệu (Star Schema)**



* 1. **Chi tiết các bảng**
* **Bảng FACT**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** |
| 1 | Season\_ID | Integer | Khóa chính Khóa ngoại |
| 2 | League\_ID | Integer | Khóa chính Khóa ngoại |
| 3 | Player\_ID | Integer | Khóa chính Khóa ngoại |
| 4 | GP | Integer |  |
| 5 | MIN | Decimal |  |
| 6 | FGM | Integer |  |
| 7 | FGA | Integer |  |
| 8 | 3PM | Integer |  |
| 9 | 3PA | Integer |  |
| 10 | FTM | Integer |  |
| 11 | FTA | Integer |  |
| 12 | TOV | Integer |  |
| 13 | PF | Integer |  |
| 14 | ORB | Integer |  |
| 15 | DRB | Integer |  |
| 16 | REB | Integer |  |
| 17 | AST | Integer |  |
| 18 | STL | Integer |  |
| 19 | BLK | Integer |  |
| 20 | PTS | Integer |  |

* **Bảng DIM\_DATE**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** |
| 1 | Season\_ID | Integer | Khóa chính |
| 2 | Season | String |  |

* **Bảng DIM\_LEAGUE**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** |
| 1 | League\_ID | Integer | Khóa chính |
| 2 | Stage | String |  |

* **Bảng DIM\_PLAYER**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** |
| 1 | Player\_ID | Integer | Khóa chính |
| 2 | Player | String |  |
| 3 | Team | String |  |
| 4 | Birth\_year | Integer |  |
| 5 | Birth\_date | String |  |
| 6 | Height\_cm | Integer |  |
| 7 | Weight\_kg | Integer |  |
| 8 | Nationality | String |  |
| 9 | High\_school | String |  |
| 10 | Draft\_round | Integer |  |
| 11 | Draft\_pick | Integer |  |

* 1. **Nội dung 15 câu hỏi truy vấn**

1. Cho biết số lượng cầu thủ không học trường trung học.
2. Cho biết số lượng cầu thủ không phải quốc tịch Mỹ chơi tại giải NBA.
3. Mùa giải 19-20, giải đấu nào có số lượng cầu thủ cao trên 1m9 nhiều nhất.
4. Cho biết thông tin cầu thủ có nhiều kiến tạo nhất qua từng năm tại giải NBL (Úc).
5. Cho biết thông tin cầu thủ tại NBA được draft tại round 1 và pick 1 và tổng số điểm ghi được qua các năm.
6. Cho biết top 10 cầu thủ có tỉ lệ ném 3 điểm thành công cao nhất.
7. Cho biết thông tin các cầu thủ có chỉ số BMI là thừa cân nhưng có số cú chặn bóng trên không > 100.
8. Cho biết 3 đội bóng nào có số lần được vào vòng Playoffs ít nhất.
9. Top 5 cầu thủ trẻ (<= 25 tuổi) có tổng số phút ra sân nhiều nhất.
10. Cho biết cầu thủ có số lần mất bóng nhiều nhất và đồng thời cũng lấy lại bóng từ đối thủ nhiều nhất.
11. Cho biết số lượng cầu thủ có tỉ lệ ném 2 và 3 điểm cao hơn tỉ lệ ném phạt.
12. Mùa giải nào có số lần cầu thủ phạm lỗi thấp nhất.
13. Cho biết cầu thủ nào tham gia nhiều giải đấu khác nhau.
14. Cho biết số lượng cầu thủ có 2 quốc tịch.
15. Cho biết số năm nhiều nhất mà một cầu thủ chơi cho một đội.

# CHƯƠNG 2: XÂY DỰNG KHO DỮ LIỆU(SSIS)

## Tổng quan mô hình SSIS

**Bước 1- Clean DataGoc and load Dataset:** Thực hiện xóa toàn bộ dữ liệu trong Data gốc, tiến hành đưa toàn bộ dữ liệu thô vào bảng tương ứng trong DataGoc, thực hiện tiền xử lý dữ liệu và đổ vào các bảng dữ liệu tương ứng.

**Bước 2- ETL Data to Dimension Table:** Thực hiện đổ dữ liệu vào các bảng Dimension trong Data warehouse.

**Bước 3- Fact Table:** Thực hiện đổ dữ liệu vào bảng fact trong Data warehouse.

**Bước 4- Set foreign keys:** Cài đặt các khóa ngoại tương ứng theo thiết kế lược đồ.

## Tạo project SSIS và thiết lập kết nối

### Tạo project SSIS

**Bước 1:** Mở Visual Studio 2019 -> Create a new project

Ảnh có chứa văn bản

Mô tả được tạo tự động

**Bước 2:** Tìm Integration Services Project -> Click chọn và nhấn Next

Ảnh có chứa văn bản

Mô tả được tạo tự động

**Bước 3: Đặt tên cho Project và nhấn Create**

**Graphical user interface, application

Description automatically generated**

A screenshot of a computer

Description automatically generated with medium confidence

**Hình 2.1.1 Màn hình giao diện của SSIS Project**

Giao diện chính của SSIS project, các chức năng chính nằm ở cột bên trái SSIS ToolBox

### Tạo cơ sở dữ liệu vào thiết lập kết nối

#### Tạo cơ sở dữ liệu

Bước 1: Vào MS SQL Server Management ->Tạo New Query

Graphical user interface, text, application, Word

Description automatically generated

Bước 2: Viết câu lệnh tạo Database



Kết quả sau khi chạy 2 câu lệnh trên

Graphical user interface, text, application

Description automatically generated

DataGoc là database chứa các table sau

* Original\_Data: Table chứa dữ liệu gốc, dữ liệu chưa qua bước tiền xử lý
* Clean\_Data: Table chứa dữ liệu đã trải qua bước tiền xử lý (lọc Null, đổi tên column, thêm column ,..)
* Null\_Data: Table chứa những dòng dữ liệu chứa giá trị Null được lọc từ bảng Original\_Data

Basketball\_DW là database đóng vai trò là một kho dữ liệu, chứa các table theo đúng thiết kế star schema ở mục Xây dựng lược đồ kho dữ liệu

* Dim\_Date: Bảng chiều chứa dữ liệu về thông tin năm của giải đấu
* Dim\_League: Bảng chiều chứa dữ liệu về thông tin của giải đấu
* Dim\_Player: Bảng chiều chứa dữ liệu về thông tin của cầu thủ
* Fact: Bảng sự kiên chứa thông tin và các thuộc tính để tính toán

#### Thiết lập kết nối

#### Thiết lập kết nối đến database DataGoc

Bước 1: Mở Project SSIS trong Visual Studio -> Chuột phải Connection Manager -> New Connection Manager

Graphical user interface, text

Description automatically generated

Bước 2: Chọn OLEDB -> Nhấn Add

Graphical user interface, text, application, email

Description automatically generated

Bước 3: Chọn New -> Vào MS SQL Server Management để lấy được Server Name -> Điền Server Name -> Tại bước select database name, chọn DataGoc -> Chọn Test Connection -> Chọn OK nếu test connection thành công

Graphical user interface, application, email

Description automatically generated

Kết quả thu được

Graphical user interface, text

Description automatically generated

#### Thiết lập kết nối đến database Basketbal\_DW

Bước 1: Thực hiện tương tự 2.2.2.1

Bước 2: Thực hiện tương tự 2.2.2.1

Bước 3: : Chọn New -> Vào MS SQL Server Management để lấy được Server Name -> Điền Server Name -> Tại bước select database name, chọn Basketball\_DW -> Chọn Test Connection -> Chọn OK nếu test connection thành công

Graphical user interface, text, application, email

Description automatically generated

Kết quả thu được

Graphical user interface, text

Description automatically generated

## Làm sạch và nạp dữ liệu và DataGoc

**Mục tiêu: Xóa tất cả dữ liệu đã được Insert vào 3 bảng (Original\_Data, Clean\_Data, Null\_Data) trong DataGoc, tiến hành nạp dữ liệu vào bảng Original\_Data, làm sạch dữ liệu và nạp dữ liệu vào 2 bảng (Clean\_Data, Null\_Data) tương ứng**

**Graphical user interface, text, application, chat or text message

Description automatically generated**

**Tổng quan Sequence Container Clean DataGoc and Load Dataset**

Bước 1: Kéo thả để tạo mới Sequence Container từ thanh công cụ SSIS toolbox -> Đổi tên thành Clean DataGoc and Load Dataset.

Text

Description automatically generated

Bước 2:  Kéo thả công cụ Execute SQL Task vào Container -> Đổi tên thành Make Connection

Graphical user interface, text, application, chat or text message

Description automatically generated

Bước 3: Chuột phải vào Load Dataset -> Edit -> Định nghĩa Connection là tới DataGoc-> Chỉnh sửa SQLStatement -> OK

Graphical user interface, application

Description automatically generated

*Chỉnh sửa connection và add SQLStatement*

Ảnh có chứa văn bản

Mô tả được tạo tự động

*Câu lệnh SQL để add vào SQLStatement, thực hiện việc clean các table trong DataLake*

**Bước 4:** Kéo thả công cụ Data Flow Task vào Container -> Đổi tên thành Load Data

Graphical user interface, text, application, chat or text message

Description automatically generated

**Bước 5:** Click chuột trái vào Load Data để chỉnh sửa Data Flow Task -> Kéo thả công cụ Flat File Source để định nghĩa dữ liệu nguồn

A screenshot of a computer

Description automatically generated

**Bước 6:** Click chuột phải và chọn Edit -> Chọn New -> Browse đến đường dẫn chứa file dữ liệu với định dạng csv

Graphical user interface, application

Description automatically generated

**Bước 7:** Chọn Columns -> Kiểm tra xem tình trạng dữ liệu đã đúng theo mong muốn hay chưa -> Chọn tiếp Advanced -> Chọn Suggest Types để Visual Studio tự động format lại kiểu dữ liệu của các cột -> OK

Graphical user interface, application

Description automatically generated

**Bước 8:** Click chọn Retain Null để giữ lại các dòng dữ liệu có giá trị Null cho các bước xử lý sau -> OK

Ảnh có chứa văn bản

Mô tả được tạo tự động

**Bước 9:** Kéo thả công cụ OLE DB Destination để đưa dữ liệu vào bảng Original\_Data -> Đổi tên thành Original\_Data Table

A screenshot of a computer

Description automatically generated

**Bước 10:** Click chuột phải vào Original\_Data Table -> Chọn Edit -> Chọn connection đến Data Lake -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Chỉnh sửa data type của các cột thuộc tính theo đúng báo cáo ngay trên đoạn mã đã được tự động sinh ra -> Click chọn OK

Graphical user interface, application

Description automatically generated

**Bước 11:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK

Graphical user interface, table

Description automatically generated

**Bước 12:** Kéo thả công cụ Data Flow Task vào Container -> Đổi tên thành Clean Data

Graphical user interface, text, application, chat or text message

Description automatically generated

**Bước 13:** Click chuột trái vào Preprocess Data để chỉnh sửa Data Flow Task -> Kéo thả công cụ OLE DB Source để định nghĩa dữ liệu nguồn

Ảnh có chứa văn bản

Mô tả được tạo tự động

**Bước 14:** Click chuột phải và chọn Edit -> Chọn connection đến Data Lake -> Chọn table Original\_Data

Graphical user interface, text, application, email

Description automatically generated

**Bước 15:** Click chọn Columns để lấy ra những cột cần thiết -> OK

Graphical user interface, table

Description automatically generated

**Bước 16:** Kéo thả công cụ Conditional Split để lọc dữ liệu

Graphical user interface

Description automatically generated

**Bước 17:** Click chuột phải vào Conditional Split -> Chọn Edit -> Chỉnh sửa Output Name thành Null\_Output -> Thêm điều kiện lọc Null trên toàn bộ cột -> Chỉnh sửa Default Output Name thành NotNull\_Output -> OK

Graphical user interface

Description automatically generated

**Bước 18:** Kéo thả công cụ OLE DB Destination để định nghĩa bảng sẽ đưa dữ liệu Null vào -> Đổi tên thành Null\_Data Table -> Nối Data Flow -> Chọn Null\_Output -> OK

Ảnh có chứa văn bản, ảnh chụp màn hình, màn hình

Mô tả được tạo tự động

**Bước 19:** Click chuột phải vào Null\_Data Table -> Chọn Edit -> Chọn connection đến Data Lake -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Click chọn OK

Graphical user interface, application

Description automatically generated

**Bước 20:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK

Table

Description automatically generated

**Bước 21:** Kéo thả công cụ Sort để sắp xếp lại các dòng dữ liệu

Ảnh có chứa văn bản, ảnh chụp màn hình, màn hình

Mô tả được tạo tự động

**Bước 22:** Click chuột phải vào Sort -> Chọn Edit -> Click chọn sort theo tất cả các cột và phương thức là Ascending -> OK

Table

Description automatically generated

**Bước 23:** Kéo thả công cụ OLE DB Destination để định nghĩa bảng sẽ đưa dữ liệu đã Clean vào -> Đổi tên thành Clean\_Data Table

Ảnh có chứa văn bản, ảnh chụp màn hình, màn hình

Mô tả được tạo tự động

**Bước 24:** Click chuột phải vào Clean\_Data Table -> Chọn Edit -> Chọn connection đến Data Lake -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Click chọn OK

Graphical user interface, application

Description automatically generated

**Bước 25:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK

Graphical user interface, table

Description automatically generated

Kết quả thu được sau khi hoàn thành định nghĩa các Data Flow Task cho Sequence Container Clean Data Lake and Load Dataset

Graphical user interface, text, application, chat or text message

Description automatically generated