

ĐẠI HỌC BÁCH KHOA HÀ NỘI
KHOA TOÁN - TIN

ĐỒ ÁN II

Xây dựng kho dữ liệu
và hệ thống báo cáo thông minh

ĐỒNG VĂN SỸ HOÀNG

Email: hoang.dvs227231@sis.hust.edu.vn

Mã sinh viên: 20227231

Chuyên ngành Hệ thống Thông tin Quản lý

Giảng viên hướng dẫn: TS. Lê Hải Hà

Chữ ký GVHD

HÀ NỘI, 1/2026

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. Mục tiêu và nội dung của đồ án

- a. Mục tiêu: Xây dựng hệ thống kho dữ liệu và hệ thống báo cáo thông minh
- b. Nội dung:
 - i. Tìm hiểu về kho dữ liệu và các báo cáo thông minh
 - ii. Xây dựng được hệ thống tự động lấy dữ liệu, cào dữ liệu và lập được các báo cáo tự động

2. Kết quả đạt được

- a. Tìm hiểu được các hệ thống kho dữ liệu, các hệ thống báo cáo tự động
- b. Xây dựng được hệ thống và áp dụng vào bài toán thực tế

3. Ý thức làm việc của sinh viên:

- a. Sinh viên có tinh thần cầu tiến, có ý thức làm việc tốt, tham gia đầy đủ các buổi họp và đánh giá đồ án

Hà Nội, ngày 8 tháng 01 năm 2026

Giảng viên hướng dẫn

TS. Lê Hải Hà

PHIẾU BÁO CÁO TIẾN ĐỘ ĐỒ ÁN

Danh sách đánh giá đồ án

Ngày đánh giá	Lần	Nội dung kế hoạch	Nội dung đã thực hiện	Điểm tích cực	Điểm nội dung	Ghi chú
26/11/2025	1	<ul style="list-style-type: none">Nghiên cứu lý thuyết kho dữ liệu, kinh doanh thông minhNghiên cứu PowerBIPhân tích, thiết kế ứng dụngViết khung báo cáo	<ul style="list-style-type: none">Hoàn thành các nội dung theo kế hoạchĐã viết báo cáo sơ bộ	10	10	
22/12/2025	2	<ul style="list-style-type: none">Hoàn thiện chương trìnhHoàn thiện sơ bộ báo cáo	<ul style="list-style-type: none">Đã hoàn thành các nội dung kế hoạch	10	10	

Hình 1: Phiếu báo cáo tiến độ đồ án

Tóm tắt nội dung Đồ án

Đồ án tập trung về việc xây dựng một hệ thống kho dữ liệu và thiết kế các hệ thống báo cáo thông minh, hướng tới việc xây dựng một giải pháp dữ liệu trọn vẹn.

Nội dung của đồ án được trình bày qua 4 chương chính:

Chương 1: Cơ sở lý thuyết Trình bày các khái niệm nền tảng về kho dữ liệu (Data Warehouse), quy trình ETL và hệ thống báo cáo thông minh (BI)

Chương 2: Khảo sát đề tài Phân tích thực trạng dữ liệu bóng đá giải Ngoại hạng Anh (EPL), xác định nhu cầu khai thác thông tin và các bài toán thực tế cần giải quyết.

Chương 3: Ứng dụng DW & BI trong phân tích cầu thủ giải đấu EPL Tập trung vào việc thiết kế kiến trúc kho dữ liệu, xây dựng các bảng Fact và Dimension để tối ưu hóa việc truy vấn thông tin cầu thủ.

Chương 4: Xây dựng dashboard Thiết kế và triển khai các báo cáo trực quan hóa dữ liệu (Dashboard) nhằm trình diễn các chỉ số hiệu suất (KPIs) của cầu thủ một cách sinh động. Hệ thống giúp người dùng dễ dàng so sánh, đánh giá phong độ và đưa ra các nhận định chiến thuật dựa trên dữ liệu thực tế.

Cuối cùng, đồ án tổng kết các kết quả đạt được trong việc xây dựng hệ thống DW và BI hoàn chỉnh cho giải đấu EPL, đồng thời đánh giá mức độ đáp ứng mục tiêu đề ra. Đồ án cũng chỉ ra những hạn chế còn tồn tại và đề xuất hướng phát triển mở rộng hệ thống trong tương lai.

Hà Nội, ngày 8 tháng 1 năm 2026

Tác giả đồ án

Đồng Văn Sỹ Hoàng

Mục lục

Bảng thuật ngữ tên viết tắt	2
Danh sách hình ảnh	5
Danh sách bảng	6
Chương 1 Cơ sở lý thuyết	9
1.1 Khái niệm kho dữ liệu	9
1.1.1 Định nghĩa kho dữ liệu	9
1.1.2 So sánh kho dữ liệu và cơ sở dữ liệu	10
1.2 Mô hình thiết kế kho dữ liệu	10
1.2.1 Các bước thiết kế kho dữ liệu:	11
1.2.2 Các lược đồ phổ biến trong thiết kế kho dữ liệu	12
1.3 Quy trình ETL trong hệ thống Data Warehouse	14
1.4 Tổng quan về Web Scraping	15
1.4.1 Web Scraping là gì	15
1.4.2 Cách thức hoạt động	15
1.4.3 Ứng dụng của Web Scraping	15
1.4.4 Vấn đề pháp lý và đạo đức	16
1.5 Kinh doanh thông minh	16
1.5.1 Kinh doanh thông minh là gì?	16
1.5.2 Các thành phần của kinh doanh thông minh	16
1.5.3 Vai trò của kinh doanh thông minh trong doanh nghiệp	17
1.6 Các bước trong kinh doanh thông minh	17
1.7 Lợi ích từ các ứng dụng BI	18
1.8 Công cụ trực quan hóa POWER BI	18

Chương 2 Khảo sát đề tài	20
2.1 Khảo sát nhu cầu	20
2.1.1 Giới thiệu nhu cầu	20
2.1.2 Nhu cầu và đặc điểm người dùng	21
2.1.3 Cảnh tranh và cơ hội	21
2.2 Khảo sát nguồn dữ liệu	22
2.3 Khảo sát các cách thức thu thập dữ liệu	23
2.4 Khảo sát các cách lưu trữ dữ liệu	25
Chương 3 Ứng dụng DW & BI trong phân tích cầu thủ giải đấu bóng đá Ngoại hạng Anh	27
3.1 Mô tả bài toán	27
3.1.1 Giới thiệu về bài toán	27
3.1.2 Giới thiệu về công ty	28
3.2 Mô tả hệ thống	29
3.2.1 Quy trình nghiệp vụ	29
3.2.2 Luồng dữ liệu	30
3.2.3 Hệ thống hiện tại	31
3.2.4 Yêu cầu hệ thống mới	32
3.3 Quy trình thu thập và lưu trữ dữ liệu thô	32
3.3.1 Quy trình cào dữ liệu	32
3.3.2 Về thông tin của các câu lạc bộ	34
3.3.3 Về thông tin cầu thủ	35
3.3.4 Về hiệu suất các cầu thủ trong trận đấu	36
3.4 Khai phá dữ liệu	38
3.4.1 Tổng quan về các bảng	38
3.4.2 Data Taxonomy	40
3.4.3 Tần suất cập nhật dữ liệu	40
3.4.4 Khám phá dữ liệu	42
3.5 Kiến trúc kho dữ liệu	52

3.6 Quy trình ETL	54
3.6.1 Data Pipeline	54
3.6.2 Quy trình ETL	55
3.6.3 Tạo cơ sở dữ liệu và lưu trữ dữ liệu	59
3.6.4 ETL dữ liệu qua Data Warehouse	62
3.7 Lập lịch quy trình tự động	66
3.7.1 Quy trình ETL	66
3.7.2 Thực hiện đầy thêm dữ liệu mới vào cơ sở dữ liệu	70
3.7.3 Lên lịch tự động với Window Task Scheduler	70
Chương 4 Xây dựng dashboard	73
4.1 Yêu cầu phân tích	73
4.2 Hệ thống Dimension	73
4.3 Data Model Logic	75
4.4 Dashboard	76
4.4.1 Dashboard tổng quan về giải đấu	76
4.4.2 Dashboard tổng quan cho câu lạc bộ	78
Kết luận	80

Bảng thuật ngữ tên viết tắt

DW	Data Warehouse
BI	Business Intelligence
SQL	Structured Query Language
ETL	Extract, Transform, Load
DB	Database
OLTP	Online Transaction Processing
OLAP	Online Analytical Processing
3NF	Third Normal Form
CEO	Chief executive officer
KPIs	Key Performance Indicators
ERP	Enterprise Resource Planning
CRM	Customer Relationship Management
HTML	Hypertext Markup Language
EPL	English Primier League
AI	Artificial Intelligence
NoSQL	Not only SQL
JSON	JavaScript Object Notation
ID	Identification
CSV	Comma-separated values
EDA	Exploratory data analysis

Danh sách hình ảnh

1	Phiếu báo cáo tiến độ đồ án	
1.1	Lược đồ hình sao	12
1.2	Lược đồ hình bông tuyết	13
1.3	Quy trình ETL	14
1.4	Các thành phần POWER BI	19
2.1	CIES Football Observatory	22
2.2	Trang web understat	23
2.3	Trang chủ trò chơi fantasy	23
2.4	Lưu trữ dữ liệu bằng excel	25
2.5	Lưu trữ dữ liệu bằng cơ sở dữ liệu	25
2.6	Lưu trữ dữ liệu bằng lưu trữ đám mây	26
3.1	Công ty Opta Sports	28
3.2	Quy trình nghiệp vụ	29
3.3	Luồng dữ liệu	30
3.4	Quy trình cào dữ liệu	33
3.5	Quy trình cào dữ liệu về đội bóng	34
3.6	Kết quả cào được về club info	35
3.7	Quy trình cào dữ liệu về cầu thủ	36
3.8	Kết quả cào được về player info	36
3.9	Quy trình cào dữ liệu hiệu suất cầu thủ	37
3.10	Quy trình cào dữ liệu hiệu suất cầu thủ	37
3.11	Data taxonamy	40
3.12	Tốc độ tăng trưởng dữ liệu bảng Player_Standard	41

3.13 Phân bố cầu thủ theo quốc gia	42
3.14 Phân bố cầu thủ theo đội bóng	43
3.15 Phân bố cầu thủ theo vị trí thi đấu	43
3.16 Phân bố cầu thủ theo thời gian thi đấu	44
3.17 Tương quan giữa bàn thắng kỳ vọng và bàn thắng thực tế	47
3.18 Tương quan giữa số lần cứu thua và số trận thắng	48
3.19 Khuynh hướng thi đấu theo địa điểm của các đội bóng	49
3.20 Khuynh hướng thi đấu theo địa điểm của các cầu thủ	50
3.21 Tỷ lệ thắng của các đội bóng theo tháng	51
3.22 Tỷ lệ thắng của đội Manchester City	51
3.23 Kiến trúc kho dữ liệu	53
3.24 Data Pipeline trong quy trình ETL	54
3.25 Quy trình ETL	55
3.26 Tiền xử lý bảng Club và bảng Season	55
3.27 Tiền xử lý bảng Club_info	56
3.28 Tiền xử lý bảng club_perform	56
3.29 Tiền xử lý bảng Player	57
3.30 Tiền xử lý bảng Club_player	57
3.31 Tiền xử lý bảng Matches	58
3.32 Tiền xử lý bảng Matches_Club_player	59
3.33 Tiền xử lý bảng Player_performance	59
3.34 Tạo cơ sở dữ liệu	60
3.35 Kết nối với cơ sở dữ liệu	61
3.36 Mô hình dữ liệu OLTP	61
3.37 Quá trình mapping các cột để tạo bảng dimension và bảng fact . . .	62
3.38 Bảng dim_date	63
3.39 Ví dụ các bảng Dimesion	64
3.40 Ví dụ các bảng Dimension	64
3.41 Ví dụ các bảng Dimension	65

3.42	Bảng fact Player_Performance	65
3.43	Lập lịch quy trình tự động	66
3.44	Xử lý thông tin cầu thủ mới	67
3.45	Xử lý cào link matchlog mới	67
3.46	Xử lý cào dữ liệu hiệu suất mới	68
3.47	Xử lý dữ liệu các bảng Matches và Club_Player	68
3.48	Xử lý dữ liệu bảng Club_Player_Matchlogs	69
3.49	Xử lý dữ liệu các bảng hiệu suất	70
3.50	Import dữ liệu vào cơ sở dữ liệu	70
3.51	Điều phối thứ tự xử lý	71
3.52	Window Task Scheduler	72
4.1	Ví dụ các bảng Dimesion	74
4.2	Ví dụ các bảng Dimension	74
4.3	Ví dụ các bảng Dimension	74
4.4	Data Model Logic	75
4.5	Data Physic Model	76
4.6	Dashboard tổng quan về giải đấu	76
4.7	Dashboard tổng quan cho câu lạc bộ	78

Danh sách bảng

1.1 So sánh Cơ sở dữ liệu (DB) và Kho dữ liệu (DW)	11
2.1 So sánh các phương pháp thu thập dữ liệu hiệu suất cầu thủ EPL .	24
3.1 Tổng quan về các bảng dữ liệu thô	38
3.2 Bảng thống kê về chiều cao và cân nặng	45
3.3 Bảng thống kê các hiệu số	46

Mở đầu

Trong kỷ nguyên “Big Data” của thể thao hiện đại, dữ liệu không chỉ là những con số thống kê khô khan mà là tài sản vô giá giúp các câu lạc bộ tuyển trạch, huấn luyện viên xây dựng chiến thuật và người hâm mộ đánh giá cầu thủ.

Tuy nhiên, hiện nay đang tồn tại một số bất cập:

- **Dữ liệu phân tán:** Thông tin về chỉ số cầu thủ (bàn thắng, kiến tạo, xG, xA, số lần tắc bóng...) nằm rải rác trên nhiều website khác nhau (như trang chủ Premier League, FBref, Understat) và thường ở dạng phi cấu trúc hoặc bán cấu trúc.
- **Khó khăn trong theo dõi lịch sử:** Các website thường chỉ hiển thị dữ liệu hiện tại (snapshot). Việc lưu trữ lịch sử phong độ theo từng tuần (gameweek) để phân tích xu hướng (trend) dài hạn là rất khó khăn nếu không có hệ thống lưu trữ riêng.
- **Thiếu tính tự động hóa:** Việc thu thập dữ liệu thủ công hàng tuần để làm báo cáo tồn rất nhiều thời gian, dễ sai sót và không đảm bảo tính cập nhật (real-time/near real-time).

Xuất phát từ nhu cầu đó, việc xây dựng một hệ thống **tự động hóa quy trình thu thập dữ liệu (ETL Pipeline)** và **xây dựng Kho dữ liệu (Data Warehouse)** tập trung là vô cùng cần thiết để phục vụ cho việc phân tích chuyên sâu và trực quan hóa thông qua Power BI.

Lời cảm ơn

Báo cáo này được thực hiện và hoàn thành tại Đại học Bách Khoa Hà Nội, nằm trong nội dung học phần *Đồ án II* của kì học 2025-1.

Em xin được dành lời cảm ơn chân thành đến **TS. Lê Hải Hà** là giảng viên đã và đang trực tiếp hướng dẫn cho em trong *Đồ án* này. Đồng thời thầy cũng đã giúp đỡ tận tình và có những đóng góp bổ ích để em có thể hoàn thành báo cáo này một cách tốt nhất. Em xin chúc thầy thật nhiều sức khỏe, vẫn nhiệt huyết với giáo dục. Hy vọng có thể được tiếp tục làm việc với thầy trong những *Đồ án* tiếp theo.

Hà Nội, ngày 8 tháng 1 năm 2026
Tác giả đồ án

Đồng Văn Sỹ Hoàng

Chương 1

Cơ sở lý thuyết

1.1 Khái niệm kho dữ liệu

1.1.1 Định nghĩa kho dữ liệu

Data Warehouse (Kho dữ liệu) là một hệ thống lưu trữ dữ liệu tập trung, được thiết kế để thu thập, lưu trữ và quản lý dữ liệu từ nhiều nguồn khác nhau. Kho dữ liệu thường được sử dụng để kết nối và phân tích dữ liệu kinh doanh từ các nguồn không đồng nhất. Chúng lưu trữ dữ liệu lịch sử và hiện tại ở một nơi duy nhất được sử dụng để tạo báo cáo phân tích cho người lao động trong toàn doanh nghiệp. Dữ liệu được xử lý, chuyển đổi và nhập để người dùng có thể truy cập dữ liệu đã xử lý trong Kho dữ liệu thông qua các công cụ Business Intelligence, ứng dụng khách SQL và bảng tính.

Kho dữ liệu có một số đặc điểm chính sau:

- **Chủ đề:** Dữ liệu trong kho dữ liệu được tổ chức theo các chủ đề hoặc lĩnh vực kinh doanh như bán hàng, tài chính, hoặc khách hàng, thay vì tổ chức theo quá trình giao dịch. Tập trung vào việc mô hình và phân tích dữ liệu cho việc ra quyết định. Cung cấp góc nhìn đơn giản và xúc tích quanh một chủ đề.
- **Tích hợp:** Dữ liệu được thu thập từ nhiều nguồn khác nhau và được tích hợp lại, loại bỏ các khác biệt về định dạng, đơn vị đo lường và chuẩn hóa.
- **Biến đổi theo thời gian:** Kho dữ liệu lưu trữ thông tin theo chiều

thời gian, cho phép theo dõi và phân tích sự thay đổi của dữ liệu theo thời gian.

- **Tính bền vững:** Sau khi dữ liệu đã được nạp vào kho dữ liệu, nó không bị thay đổi hoặc xóa đi mà chỉ có thể được cập nhật bằng cách thêm dữ liệu mới. Chỉ yêu cầu hai thao tác là nạp dữ liệu và truy cập dữ liệu.

1.1.2 So sánh kho dữ liệu và cơ sở dữ liệu

- **Data Warehouse (Kho dữ liệu)** là nơi lưu trữ dữ liệu của một tổ chức, doanh nghiệp. Kho dữ liệu chỉ có nhân viên trong tổ chức, doanh nghiệp được phép sử dụng để phục vụ cho việc phân tích dữ liệu và báo cáo. Dữ liệu đầu vào của kho dữ liệu có thể là một hoặc nhiều cơ sở dữ liệu quan hệ, các file excel, CSV, text, ... lưu trữ thông tin dữ liệu cần phân tích. Tất cả được tổng hợp lại làm đầu vào cho kho dữ liệu tại tầng dưới cùng và sau một quá trình ETL phức tạp, các dữ liệu đó sẽ được chuyển đổi, làm sạch để có thể dễ dàng trích xuất lên báo cáo phân tích.
- **Data Base (Cơ sở dữ liệu)** là nơi tập hợp các dữ liệu có cấu trúc, có tổ chức và mối liên quan với nhau, được dùng để ghi và truy vấn dữ liệu. Cơ sở dữ liệu thường được lưu trữ và truy cập từ một hệ thống máy tính, được nhiều người sử dụng, tương tác với cơ sở dữ liệu và tổ chức theo mô hình. Cơ sở dữ liệu được sử dụng bằng hình thức xử lý trực tuyến, có sự chuẩn hóa (đối với mô hình cơ sở dữ liệu quan hệ) để giảm thiểu dữ liệu dư thừa, tối ưu hóa dung lượng lưu trữ.

1.2 Mô hình thiết kế kho dữ liệu

Mô hình thiết kế kho dữ liệu là quá trình xác định cách dữ liệu sẽ được tổ chức, lưu trữ và quản lý trong kho dữ liệu. Mục tiêu của thiết kế kho dữ liệu là đảm bảo rằng dữ liệu có thể được truy cập một cách hiệu quả để hỗ trợ các hoạt động phân tích và ra quyết định. Thiết kế kho dữ liệu liên

Bảng 1.1: So sánh Cơ sở dữ liệu (DB) và Kho dữ liệu (DW)

Tiêu chí	Cơ sở dữ liệu (OLTP)	Kho dữ liệu (OLAP)
Mục đích chính	Xử lý giao dịch hàng ngày (Ghi đơn hàng, thanh toán...).	Phân tích, báo cáo và hỗ trợ ra quyết định (BI).
Đặc điểm dữ liệu	Dữ liệu chi tiết, hiện tại (Current data).	Dữ liệu tổng hợp, lịch sử (Historical data).
Cấu trúc	Chuẩn hóa cao (3NF) để tránh dư thừa dữ liệu.	Phi chuẩn hóa (Star/Snowflake Schema) để tối ưu truy vấn.
Tính biến động	Thay đổi liên tục (Thêm, Sửa, Xóa).	Ôn định, chủ yếu là Đọc (Read-only), cập nhật định kỳ.
Kích thước	Thường nhỏ đến trung bình (GB).	Rất lớn (TB đến PB).
Người dùng	Nhân viên vận hành, khách hàng, hệ thống.	Nhà quản lý, chuyên viên phân tích dữ liệu, CEO.
Hiệu năng	Tối ưu cho các giao dịch ngắn, nhanh.	Tối ưu cho các truy vấn phức tạp, khối lượng lớn.

quan đến việc chọn cấu trúc dữ liệu phù hợp và các kỹ thuật tổ chức dữ liệu để tối ưu hóa hiệu suất và tính dễ sử dụng.

1.2.1 Các bước thiết kế kho dữ liệu:

Quá trình thiết kế kho dữ liệu thường bao gồm các bước sau:

- Xác định yêu cầu nghiệp vụ:** Bắt đầu bằng việc hiểu rõ các yêu cầu phân tích và báo cáo của doanh nghiệp. Điều này bao gồm việc xác định các chỉ số kinh doanh quan trọng (Key Performance Indicators - KPIs), các câu hỏi phân tích cần trả lời, và các nguồn dữ liệu cần thiết.
- Phân tích nguồn dữ liệu:** Tiến hành phân tích các hệ thống nguồn (cơ sở dữ liệu giao dịch, hệ thống ERP, CRM, v.v.) để xác định các dữ liệu cần thiết và cách dữ liệu đó được tổ chức.
- Thiết kế mô hình dữ liệu:** Dựa trên các yêu cầu nghiệp vụ, tiến

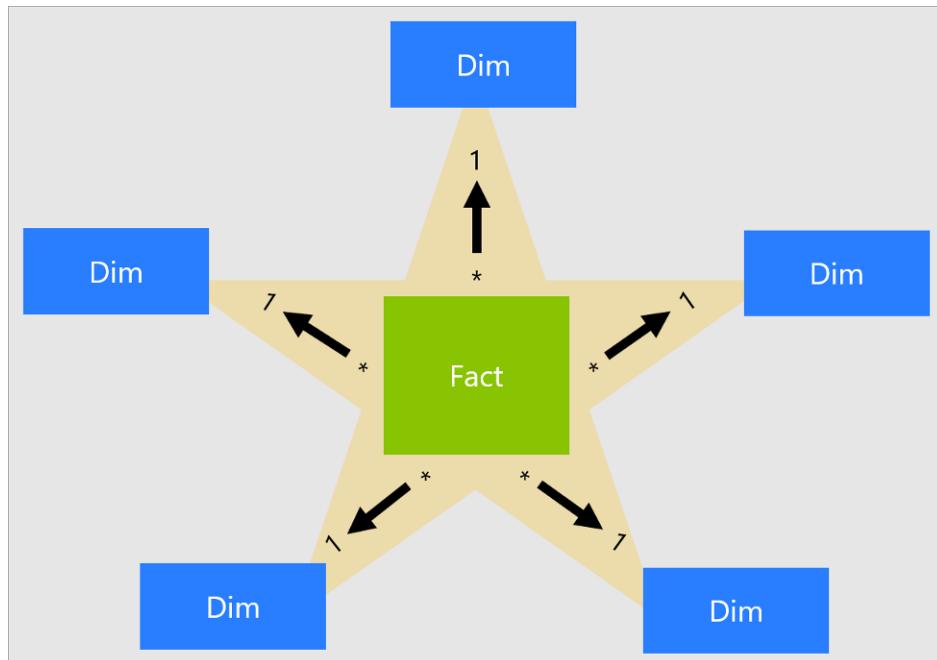
hành thiết kế mô hình dữ liệu để xác định các bảng sự kiện (fact tables), bảng chiều (dimension tables), và mối quan hệ giữa chúng.

- **Lập kế hoạch quy trình ETL:** Thiết lập quy trình ETL để trích xuất dữ liệu từ các nguồn, biến đổi dữ liệu thành định dạng phù hợp, và nạp dữ liệu vào kho dữ liệu.
- **Tối ưu hóa và triển khai:** Thực hiện các kỹ thuật tối ưu hóa để đảm bảo hiệu suất của kho dữ liệu, bao gồm việc lập chỉ mục, phân vùng dữ liệu và sử dụng các công nghệ lưu trữ phù hợp. Sau đó, triển khai kho dữ liệu để hỗ trợ người dùng cuối.

1.2.2 Các lược đồ phổ biến trong thiết kế kho dữ liệu

Thiết kế kho dữ liệu thường sử dụng ba mô hình lược đồ chính để tổ chức dữ liệu:

- **Lược đồ hình sao - Star Schema**



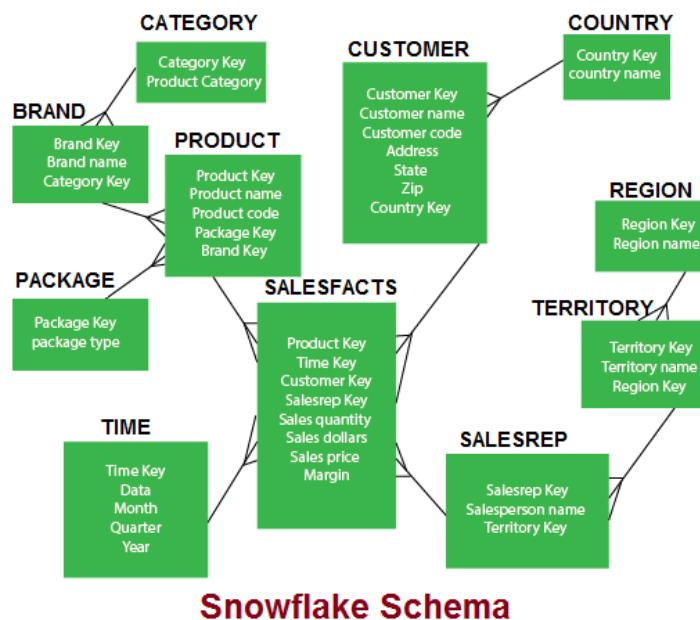
Hình 1.1: Lược đồ hình sao

- Đây là mô hình đơn giản và phổ biến nhất, trong đó bảng sự kiện nằm ở trung tâm và kết nối trực tiếp với các bảng chiều
- Bảng sự kiện chứa các dữ liệu đo lường, chẳng hạn như doanh thu

hoặc số lượng bán hàng, và các khóa ngoại để liên kết với các bảng chiều.

- Các bảng chiều cung cấp thông tin bổ sung, như thông tin về sản phẩm, khách hàng hoặc thời gian, giúp người dùng có thể phân tích dữ liệu từ nhiều góc độ khác nhau.
- Ưu điểm của lược đồ hình sao là dễ hiểu và truy xuất dữ liệu nhanh chóng, do số lượng bảng tham gia trong truy vấn là tối thiểu.

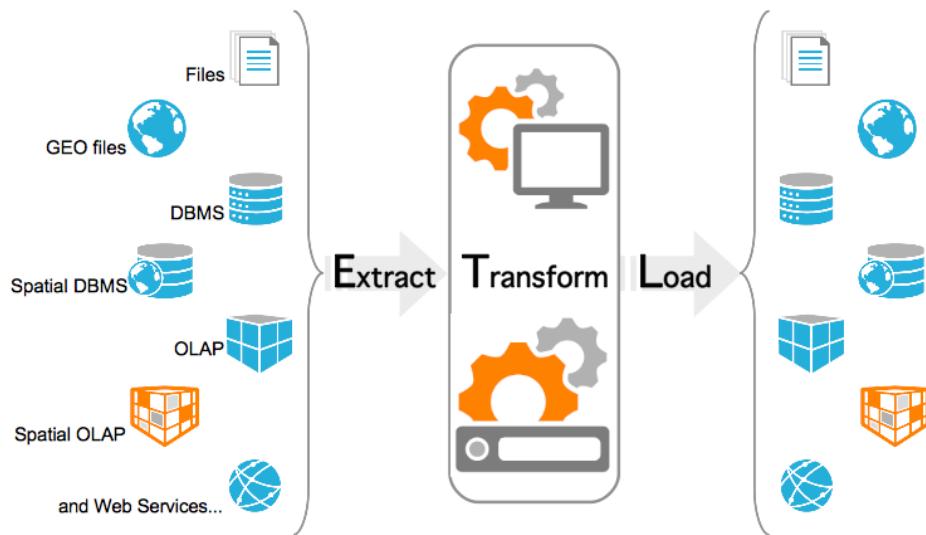
• Lược đồ bông tuyết - Snowflake Schema



Hình 1.2: Lược đồ bông tuyết

- Là một biến thể của lược đồ hình sao, trong đó các bảng chiều được bình thường hóa để loại bỏ sự trùng lặp dữ liệu. Các bảng chiều có thể được chia nhỏ thành các bảng con để lưu trữ các thuộc tính phân cấp chi tiết hơn.
- Ví dụ, thay vì có một bảng chiều "địa lý" chứa cả quốc gia, vùng và thành phố, các thuộc tính này có thể được tách ra thành các bảng riêng biệt và liên kết với nhau.
- Lược đồ bông tuyết giúp tiết kiệm không gian lưu trữ và có thể cải thiện hiệu suất truy vấn trong một số trường hợp, nhưng phức tạp hơn và khó quản lý hơn so với lược đồ hình sao.

1.3 Quy trình ETL trong hệ thống Data Warehouse



Hình 1.3: Quy trình ETL

Quy trình ETL (Extract, Transform, Load)

- **Extract:** Đây là bước khởi đầu của quy trình, nơi dữ liệu thô được thu thập từ nhiều nguồn không đồng nhất như hệ thống ERP, CRM, website hay các file excel. Mục tiêu cốt lõi của giai đoạn này là lấy dữ liệu nhanh chóng và chính xác đưa vào vùng đệm (Staging Area), đồng thời đảm bảo không làm gián đoạn hay ảnh hưởng đến hiệu suất vận hành của các hệ thống nguồn.
- **Trans:** Transform là giai đoạn quan trọng nhất, đóng vai trò như "bộ lọc" tinh chỉnh chất lượng dữ liệu. Tại đây, dữ liệu thô sẽ trải qua quá trình làm sạch (loại bỏ lỗi, dữ liệu rác), chuẩn hóa định dạng và áp dụng các quy tắc nghiệp vụ phức tạp để tính toán các chỉ số. Kết quả là dữ liệu trở nên nhất quán, tin cậy và mang ý nghĩa kinh doanh thực sự.
- **Load:** Load là công đoạn cuối cùng, đưa dữ liệu đã được làm sạch và tối ưu hóa vào kho dữ liệu đích. Dữ liệu được sắp xếp khoa học vào các bảng sự kiện và danh mục, sẵn sàng phục vụ cho các công cụ báo cáo truy xuất với tốc độ cao, giúp doanh nghiệp có cái nhìn toàn cảnh và ra quyết định chính xác.

1.4 Tổng quan về Web Scraping

1.4.1 Web Scraping là gì

Web Scraping là quá trình sử dụng các công cụ phần mềm hoặc mã lập trình để tự động trích xuất dữ liệu từ các trang web. Thông thường, khi truy cập web, con người đọc nội dung bằng mắt. Nhưng với Web Scraping, một con bot sẽ thay thế con người để truy cập vào hàng loạt trang web, đọc mã nguồn (HTML) và bóc tách những thông tin cụ thể rồi lưu trữ chúng dưới dạng cấu trúc (như Excel, CSV, hoặc Cơ sở dữ liệu).

1.4.2 Cách thức hoạt động

Quy trình hoạt động của Web Scraping thường diễn ra qua 4 bước kỹ thuật chính:

- Gửi yêu cầu : Trình thu thập dữ liệu gửi yêu cầu truy cập HTTP đến trang web mục tiêu, tương tự như trình duyệt khi bạn truy cập trang web đó
- Nhận phản hồi: Máy chủ phản hồi bằng cách gửi lại nội dung trang web, thường là dưới dạng mã HTML. Đối với một số trang web mục tiêu hiện đại hơn thì có thể bao gồm cả dữ liệu JSON
- Phân tích cú pháp: Công cụ Scraping sẽ đọc mã HTML vừa nhận được, sử dụng các bộ chọn như CSS Selectors hoặc XPath để tìm đúng dữ liệu như được yêu cầu trong công cụ
- Trích xuất và lưu trữ: Sau khi tìm được đến dữ liệu mong muốn, công cụ sẽ làm sạch nó, và lưu nó về định dạng đã cài đặt từ trước

1.4.3 Ứng dụng của Web Scraping

Web Scraping được ứng dụng rất nhiều trong việc thu thập dữ liệu, phục vụ cho các đề tài nghiên cứu:

- Thu thập dữ liệu từ nhiều trang web khác nhau nhằm tham khảo và giám sát các chỉ số cần thiết cho đề tài từ nhiều nguồn khác nhau

- Việc sử dụng công cụ Web Scraping có thể giúp tự động thu thập dữ liệu một cách liên tục mà không cần phải làm thủ công

1.4.4 Vấn đề pháp lý và đạo đức

Đây là khía cạnh quan trọng nhất mà người làm Scraping cần lưu ý để tránh rắc rối pháp lý, không phải trang web nào cũng cho phép các bot truy cập để thu thập dữ liệu. Và việc thu thập dữ liệu các nhân mà không có sự đồng ý có thể sẽ vi phạm các luật bảo mật dữ liệu và luật an ninh mạng. Hãy luôn thực hiện cào dữ liệu một cách an toàn, chỉ lấy dữ liệu công khai và kiểm tra tính pháp lý trước khi triển khai các đề tài lớn.

1.5 Kinh doanh thông minh

1.5.1 Kinh doanh thông minh là gì?

Kinh doanh thông minh là một tập hợp các phương pháp, quy trình và công cụ giúp doanh nghiệp thu thập, lưu trữ, phân tích và hiển thị dữ liệu để hỗ trợ quá trình ra quyết định. BI tập trung vào việc chuyển đổi dữ liệu thô từ các nguồn khác nhau thành thông tin có ý nghĩa, nhằm cung cấp cái nhìn sâu sắc về hoạt động kinh doanh và giúp doanh nghiệp cải thiện hiệu quả hoạt động, tăng cường khả năng cạnh tranh.

1.5.2 Các thành phần của kinh doanh thông minh

Hệ thống BI được cấu thành từ các thành phần chính sau:

- **Kho dữ liệu:** Hệ thống lưu trữ tập trung, tích hợp dữ liệu từ nhiều nguồn để phục vụ phân tích.
- **Khai phá dữ liệu:** Sử dụng thuật toán để tìm kiếm các mẫu và xu hướng tiềm ẩn trong dữ liệu.
- **Phân tích trực tuyến (OLAP):** Công cụ phân tích đa chiều, giúp xem xét dữ liệu từ nhiều góc độ khác nhau.
- **Trực quan hóa dữ liệu:** Chuyển đổi dữ liệu thành biểu đồ, đồ thị giúp dễ dàng nhận diện thông tin.

- **Báo cáo và Dashboard:** Cung cấp cái nhìn tổng quan về hiệu suất, hỗ trợ ra quyết định kịp thời.

1.5.3 Vai trò của kinh doanh thông minh trong doanh nghiệp

BI giúp doanh nghiệp chuyển đổi dữ liệu thành tri thức để tối ưu hóa hoạt động thông qua các vai trò chính:

- **Hỗ trợ ra quyết định:** Cung cấp phân tích chính xác dựa trên dữ liệu thực tế, giúp giảm thiểu rủi ro và quản lý nguồn lực hiệu quả.
- **Nâng cao hiệu suất:** Theo dõi các chỉ số KPI để kịp thời nhận diện vấn đề và tối ưu hóa các mục tiêu kinh doanh.
- **Dự báo xu hướng:** Phân tích hành vi khách hàng và thị trường, giúp doanh nghiệp chủ động xây dựng chiến lược thích ứng.
- **Tối ưu quy trình:** Phát hiện các điểm nghẽn nội bộ để cải thiện hiệu quả vận hành và tiết kiệm chi phí.

1.6 Các bước trong kinh doanh thông minh

Quy trình BI gồm 5 bước cơ bản nhằm chuyển đổi dữ liệu thô thành giá trị thực tiễn:

- **Bước 1: Xác định yêu cầu:** Xác định mục tiêu và các chỉ số cần phân tích.
- **Bước 2: Thu thập và tích hợp:** Tập hợp dữ liệu từ nhiều nguồn khác nhau.
- **Bước 3: Làm sạch và biến đổi:** Chuẩn hóa dữ liệu để đảm bảo tính chính xác và đồng nhất.
- **Bước 4: Phân tích và trực quan hóa:** Khai phá dữ liệu và trình bày qua các biểu đồ, báo cáo (Dashboard).
- **Bước 5: Ra quyết định:** Sử dụng kết quả phân tích để đưa ra các hành động chiến lược.

1.7 Lợi ích từ các ứng dụng BI

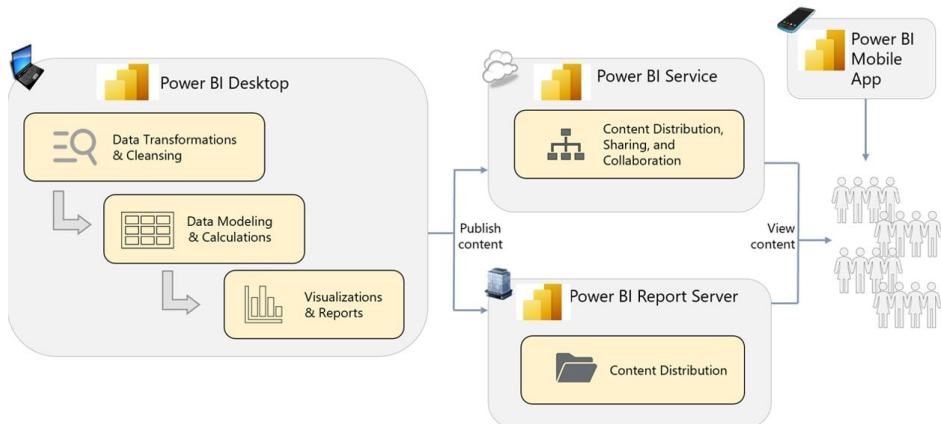
Việc triển khai BI mang lại những lợi ích chiến lược sau cho doanh nghiệp:

- **Tối ưu hóa ra quyết định:** Cung cấp thông tin kịp thời, giúp quản lý đưa ra quyết định dựa trên dữ liệu thực tế thay vì cảm tính.
- **Nâng cao hiệu suất vận hành:** Nhận diện các điểm yếu trong quy trình thông qua chỉ số KPI để tối ưu hóa nguồn lực.
- **Khám phá cơ hội mới:** Phân tích thị trường và hành vi khách hàng để xây dựng chiến lược kinh doanh đột phá.
- **Thấu hiểu khách hàng:** Cung cấp cái nhìn sâu sắc về nhu cầu khách hàng, giúp cá nhân hóa sản phẩm và dịch vụ.
- **Quản trị rủi ro:** Nhận diện sớm các mối đe dọa tiềm ẩn để đưa ra biện pháp phòng ngừa và giảm thiểu tổn thất.

1.8 Công cụ trực quan hóa POWER BI

Power BI là công cụ phân tích và trực quan hóa dữ liệu dành cho lĩnh vực Business Intelligence (BI) của Microsoft. Đây là công cụ thông minh hỗ trợ doanh nghiệp trong việc tạo ra báo cáo quản trị đầy đủ và trực quan, giúp các nhà lãnh đạo có thể đưa ra quyết định chính xác dựa vào kết quả phân tích tình hình kinh doanh.

Power BI có thể kết nối với nhiều dịch vụ phần mềm, hoạt động song song, từ đó có thể kết nối nhiều nguồn dữ liệu với nhau và tạo ra mô hình dữ liệu bao gồm các biểu đồ, con số và các thông tin được tự động tính toán chi tiết, liền mạch và được thể hiện một cách trực quan hóa. Mô hình này có thể được chia sẻ với bất cứ ai trong tổ chức, hoặc những người có tài khoản Power BI. Power BI cũng tối ưu hóa mức tiêu thụ dữ liệu Azure, giúp tiết kiệm đáng kể cho các công ty dựa vào Power BI cho nhu cầu báo cáo hàng ngày.



Hình 1.4: Các thành phần POWER BI

Power BI bao gồm bốn thành phần chính:

- **Power BI Desktop:** Ứng dụng cài đặt trên máy tính cho phép người dùng thiết kế báo cáo và bảng điều khiển.
- **Power BI Service (Online):** Nền tảng trực tuyến để chia sẻ và cộng tác trên các báo cáo và bảng điều khiển.
- **Power BI Report Server:** Cho phép người dùng có thể xuất báo cáo sau khi hoàn thành thao tác trên hệ thống.
- **Power BI App:** Ứng dụng trên thiết bị di động để truy cập và theo dõi báo cáo từ bất cứ đâu.

Chương 2

Khảo sát đề tài

2.1 Khảo sát nhu cầu

2.1.1 Giới thiệu nhu cầu

Thị trường phân tích dữ liệu hiệu suất cầu thủ trong bóng đá, đặc biệt tại **Ngoại hạng Anh (Premier League - EPL)**, đang phát triển mạnh mẽ nhờ sự bùng nổ của công nghệ AI, machine learning và dữ liệu theo dõi thời gian thực. Theo các báo cáo năm 2025, thị trường phân tích thể thao toàn cầu đạt khoảng **5-6 tỷ USD**, với tốc độ tăng trưởng CAGR từ 20-27% đến năm 2030-2033, trong đó bóng đá chiếm tỷ lệ lớn nhất nhờ các giải đấu hàng đầu như EPL.

Riêng phân khúc phần mềm phân tích bóng đá ước tính khoảng **0.44 tỷ USD** năm 2025, dự kiến tăng lên 0.87 tỷ USD vào 2033. Tại EPL, **75% câu lạc bộ** có đội ngũ phân tích chuyên biệt, với quy mô trung bình 6 người (tăng từ 4 người năm 2020), và các ông lớn như Manchester City, Arsenal, Liverpool đều tư 1-5 triệu GBP/năm. EPL dẫn đầu châu Âu về áp dụng analytics, với dữ liệu từ hàng nghìn trận đấu được thu thập bởi các nhà cung cấp như Opta (Stats Perform) và Oracle, hỗ trợ từ tuyển mộ đến quản lý chấn thương.

Quy mô này phản ánh giá trị kinh tế khổng lồ của EPL (doanh thu hàng tỷ USD/năm), nơi phân tích dữ liệu giúp tối ưu hóa hiệu suất và giảm rủi ro tài chính trong thị trường chuyển nhượng cạnh tranh. Quy mô này phản ánh giá trị kinh tế khổng lồ của EPL (doanh thu hàng tỷ USD/năm), nơi phân tích dữ liệu giúp tối ưu hóa hiệu suất và giảm rủi ro tài chính trong

thị trường chuyển nhượng cạnh tranh.

2.1.2 Nhu cầu và đặc điểm người dùng

Đối tượng sử dụng kết quả phân tích hiệu suất cầu thủ EPL rất đa dạng, mỗi nhóm có những đặc điểm và kỳ vọng riêng biệt:

- Nhóm chuyên môn (HLV, chuyên gia phân tích,...): Họ cần các chỉ số nâng cao như xG (Bàn thắng kỳ vọng), xA (Kiến tạo kỳ vọng), tỷ lệ thoát pressing, hay khả năng tinh tiến bóng. Họ cần dữ liệu để so sánh cầu thủ mục tiêu với đội hình hiện tại hoặc để bắt bài lối chơi của đối thủ.
- Nhóm truyền thông và sáng tạo nội dung: Họ cần các biểu đồ trực quan hóa như Heatmap, Radar chart để minh họa cho các bài viết phân tích, giúp khán giả dễ dàng hình dung sự xuất sắc hoặc sa sút của một ngôi sao.
- Nhóm người hâm mộ và người chơi Fantasy Premier League: Họ cần các dự báo về phong độ trong ngắn hạn để đưa ra quyết định chuyển nhượng trong trò chơi FPL hoặc đơn giản là để có cơ sở dữ liệu khách quan cho các cuộc tranh luận bóng đá hàng ngày.

2.1.3 Cạnh tranh và cơ hội

Thị trường phân tích dữ liệu bóng đá EPL cạnh tranh gay gắt với các nhà cung cấp lớn thống trị:

- **Stats Perform (Opta)** → Nhà cung cấp chính thức dữ liệu EPL, mạnh về dữ liệu chi tiết, AI, được hầu hết clubs và truyền thông sử dụng.
- **Wyscout (Hudl)** → Tập trung video phân tích, phủ sóng rộng ở nhiều giải đấu, phổ biến cho tuyển mộ.
- **Các đối thủ khác:** Catapult (wearable tracking), Genius Sports (betting và official data), Sportradar, Oracle (cloud analytics cho EPL), StatsBomb (advanced models).

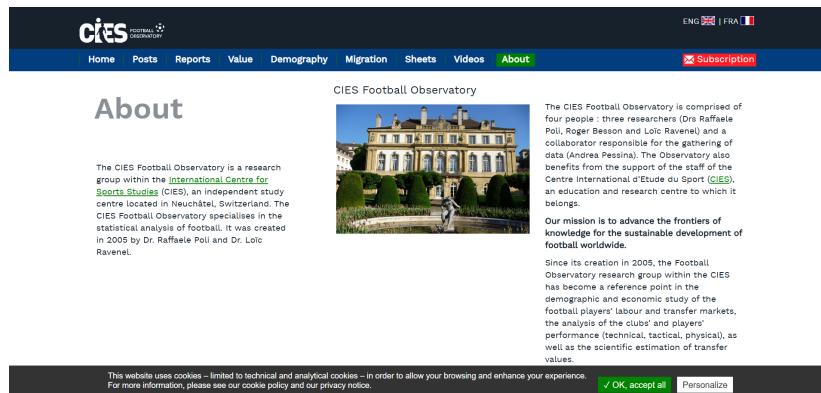
Cạnh tranh tập trung vào độ chính xác, tốc độ real-time và tích hợp AI/video. Các ông lớn như Opta và Wyscout chiếm thị phần lớn, nhưng chi phí cao tạo rào cản cho clubs nhỏ.

Cơ hội: Thị trường tăng trưởng nhanh (CAGR >20%), mở rộng cho startup Việt Nam hoặc khu vực với giải pháp chi phí thấp, tích hợp AI predictive hoặc tùy chỉnh cho giải đấu châu Á. Với EPL dẫn đầu xu hướng, các sản phẩm mới tập trung real-time và AI có tiềm năng thâm nhập, đặc biệt khi nhu cầu tiếp cận cho clubs nhỏ đang tăng.

2.2 Khảo sát nguồn dữ liệu

Các nguồn cung cấp dữ liệu về hiệu suất cầu thủ tại giải đấu Ngoại hạng Anh bao gồm:

- Báo cáo của các tổ chức nghiên cứu bóng đá uy tín như CIES Football Observatory, Football Benchmark và các báo cáo kỹ thuật của giải đấu. Các báo cáo này được phát hành hàng tuần về hiệu suất cầu thủ, giá trị chuyển nhượng, xu hướng chiến thuật....



Hình 2.1: CIES Football Observatory

- Dữ liệu từ các trang web công khai và miễn phí như FBref, WhoScore, Understat.. Đây là những trang web cung cấp các chỉ số đã qua xử lý, rất phù hợp để thu thập dữ liệu thủ công hoặc sử dụng công cụ cào dữ liệu.
- Nguồn dữ liệu thông qua các trò chơi như Fantasy Primier League. Có thể dùng API của trò chơi để lấy các chỉ số về phong độ, điểm số, và

Nº	Team	W	D	L	G	GA	PTS	xG	xGA	xPTS
1	Arsenal	17	12	3	2	31	10	39	32.53 +1.03	12.74 -0.76
2	Manchester City	17	12	1	4	41	16	37	35.68 +0.07	18.09 -0.01
3	Aston Villa	17	11	3	3	27	18	36	21.27 +0.73	24.72 -0.72
4	Chelsea	17	8	5	4	29	17	29	31.04 -0.16	23.40 +0.40
5	Liverpool	17	8	2	6	28	25	28	30.12 +2.03	24.62 +0.38
6	Sunderland	17	7	5	6	31	28	27	17.00 +0.01	23.79 -0.79
7	Manchester United	17	7	5	6	31	28	26	24.80 +0.00	24.97 +0.03
8	Crystal Palace	17	7	5	5	21	19	28	30.34 +0.16	23.86 +0.06
9	Brighton	17	6	6	6	25	23	24	28.42 +1.62	26.06 -0.06
10	Everton	17	7	3	7	18	20	24	20.89 +2.88	28.13 -0.13
11	Newcastle United	17	6	5	6	23	22	23	26.16 +1.18	18.27 +0.03
12	Brentford	17	7	2	8	24	25	23	28.82 +0.02	23.27 +0.73
13	Fulham	17	7	2	8	24	28	23	20.54 +0.48	24.82 -0.38
14	Tottenham	17	6	4	7	26	23	22	18.03 +0.07	23.55 -0.56
15	Bournemouth	17	5	7	5	28	29	22	28.89 +0.00	22.81 +0.09
16	Leeds	17	5	4	8	24	31	10	25.50 +1.00	23.76 +0.24
17	Nottingham Forest	17	5	5	9	17	26	18	19.94 +0.48	27.21 +0.71
18	West Ham	17	3	4	10	19	35	13	17.94 +0.08	34.90 +0.10

Hình 2.2: Trang web understat

biến động giá của các cầu thủ có dữ liệu ở trong trò chơi

Goals >	Assists >	Total Passes >	Clean Sheets >
1 Erling Haaland	19	1 Bruno Fernandes	1 David Raya
2 Igor Thago	11	2 Rayan Cherki	2 Robert Sánchez
3 Hugo Ekitike	8	3 Mohammed Kudus	3 Gianluigi Donnarumma
4 Antoine Semenyo	8	4 Granit Xhaka	4 Dean Henderson

Hình 2.3: Trang chủ trò chơi fantasy

2.3 Khảo sát các cách thức thu thập dữ liệu

Các cách thu thập dữ liệu về hiệu suất cầu thủ:

- Thu thập dữ liệu thứ cấp: Sử dụng các loại báo cáo của các tổ chức nghiên cứu bóng đá uy tín như CIES Football Observatory, Football Benchmark và các báo cáo kỹ thuật của giải đấu. Các báo cáo này được phát hành hàng tuần về hiệu suất cầu thủ, giá trị chuyển nhượng, xu hướng chiến thuật... Phương pháp này tiết kiệm thời gian nhưng có thể dữ liệu không phù hợp với yêu cầu phân tích.
- Web Scraping: Tự động thu thập dữ liệu thông qua các trang web miễn phí. Truy cập các trang web và cào xuồng dữ liệu về thông tin cầu thủ,

hiệu suất, giá cả, đây là một cách thức hiệu quả để khai thác dữ liệu lớn và được cập nhật liên tục theo thời gian

- Thông qua các công ty chuyên thu thập dữ liệu: Liên kết với các công ty chuyên về việc thu thập dữ liệu liên quan đến đề tài. Cách làm này có thể sẽ có dữ liệu phù hợp để phân tích nhưng đi kèm với đó là chi phí lớn để có thể thực hiện

Bảng 2.1: So sánh các phương pháp thu thập dữ liệu hiệu suất cầu thủ EPL

Tiêu chí	Thu thập dữ liệu thứ cấp (Báo cáo)	Web Scraping (Cào dữ liệu)	Công ty dữ liệu chuyên nghiệp
Chi phí	Thường miễn phí hoặc phí thấp cho bản tóm tắt.	Rẻ (chỉ tốn chi phí vận hành, máy chủ).	Rất cao (vài nghìn đến chục nghìn USD).
Độ chi tiết	Thấp - Trung bình (Dữ liệu đã được tổng hợp).	Cao (Lấy được từng thông số nhỏ của từng trận).	Rất cao (Dữ liệu thô, tọa độ di chuyển, nhịp tim...).
Tính cập nhật	Chậm (Theo tuần, tháng hoặc sau mùa giải).	Rất nhanh (Cập nhật ngay sau khi trận đấu kết thúc).	Thời gian thực (Real-time).
Độ tin cậy	Rất cao (Đã qua kiểm chứng bởi chuyên gia).	Trung bình (Phụ thuộc vào nguồn web và thuật toán cao).	Tuyệt đối (Là nguồn chuẩn cho các CLB và nhà cái).
Kỹ năng yêu cầu	Đọc hiểu, tổng hợp và phân tích báo cáo.	Kỹ năng lập trình (Python, R, Selenium...).	Kỹ năng xử lý dữ liệu lớn (Big Data), SQL.
Khả năng tùy biến	Thấp (Phụ thuộc vào nội dung báo cáo có sẵn).	Rất cao (Muốn lấy chỉ số nào, giai đoạn nào cũng được).	Cao (Cung cấp theo gói yêu cầu).

2.4 Khảo sát các cách lưu trữ dữ liệu

- File Excel và CSV Đây là công cụ phổ biến để lưu trữ và quản lý dữ liệu dạng bảng như danh mục cầu thủ, câu lạc bộ, giải đấu, trận đấu và hiệu suất cầu thủ. Ngoài ra Excel còn hỗ trợ các tính năng tính toán, lọc, tạo báo cáo và biểu đồ giúp phân tích dữ liệu hiệu quả



Hình 2.4: Lưu trữ dữ liệu bằng excel

- Cơ sở dữ liệu Dữ liệu được lưu trữ có cấu trúc trong các hệ quản trị cơ sở dữ liệu như SQL server, MySQL, Oracle, PostgreSQL hoặc NoSQL như MongoDB để quản lý lượng lớn dữ liệu, truy vấn nhanh và bảo mật cao hơn



Hình 2.5: Lưu trữ dữ liệu bằng cơ sở dữ liệu

- File word và PDF Dùng để lưu trữ các báo cáo phân tích, các báo cáo chuyên môn của giải đấu, dễ dàng chia sẻ và đảm bảo tính toàn vẹn
- Dịch vụ lưu trữ đám mây OneDrive, Google Drive, Amazon S3 giúp lưu trữ linh hoạt, dễ dàng truy cập và chia sẻ dữ liệu đến nhiều thiết bị ở những vị trí khác nhau.



Hình 2.6: Lưu trữ dữ liệu bằng lưu trữ đám mây

Chương 3

Ứng dụng DW & BI trong phân tích cầu thủ giải đấu bóng đá Ngoại hạng Anh

3.1 Mô tả bài toán

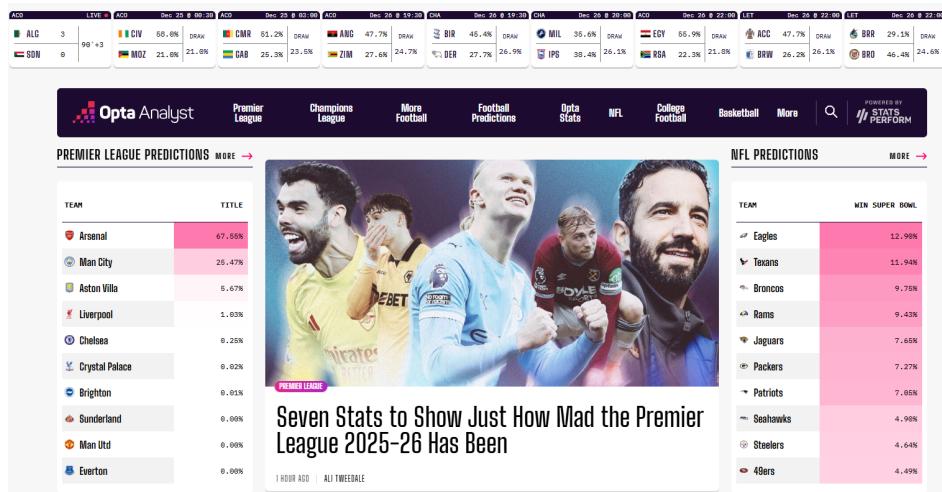
3.1.1 Giới thiệu về bài toán

Bài toán Trong bối cảnh bóng đá hiện đại, đặc biệt là giải Ngoại hạng Anh (EPL), việc đánh giá cầu thủ đang chuyển dịch mạnh mẽ từ cảm tính cá nhân sang phân tích dữ liệu chuyên sâu. Thay vì chỉ dựa vào các chỉ số cơ bản như bàn thắng hay kiến tạo, sự bùng nổ của "big data" đòi hỏi những phương pháp tiếp cận khoa học để khai thác các giá trị ẩn như đóng góp phòng ngự, khả năng pressing hay các chỉ số kỳ vọng (xG, xA). Điều này giúp loại bỏ tính chủ quan, mang lại cái nhìn khách quan và chính xác hơn về năng lực thực sự cũng như tiềm năng của mỗi cầu thủ.

Việc ứng dụng các công cụ hiện đại như xử lý dữ liệu với Pandas, trực quan hóa hay Machine Learning không chỉ giúp phát hiện những tài năng bị định giá thấp mà còn hỗ trợ tối ưu hóa chiến thuật và dự báo rủi ro chấn thương. Đề tài nghiên cứu phân tích dữ liệu cầu thủ EPL vì vậy mang lại giá trị thực tiễn cao, cung cấp những insight quan trọng cho các câu lạc bộ và người hâm mộ. Trong một giải đấu khắc nghiệt nhất hành tinh, các giải pháp dựa trên dữ liệu chính là chìa khóa để nâng tầm chuyên môn và làm phong phú thêm trải nghiệm bóng đá cho cộng đồng.

3.1.2 Giới thiệu về công ty

Opta Sports là một doanh nghiệp hàng đầu thế giới chuyên cung cấp các giải pháp dữ liệu và trí tuệ nhân tạo trong lĩnh vực thể thao, với mục tiêu biến những diễn biến trên sân cỏ thành những thông tin chi tiết và giá trị. Với sự kết hợp giữa công nghệ thu thập dữ liệu thời gian thực tiên tiến và đội ngũ chuyên gia phân tích giàu kinh nghiệm, Opta Sports đã khẳng định vị thế là "tiêu chuẩn vàng" trong ngành thống kê thể thao toàn cầu. Doanh nghiệp này không chỉ cung cấp các con số thô, mà còn chú trọng vào việc xây dựng các chỉ số nâng cao và mô hình dự đoán, giúp các đối tác tối ưu hóa hiệu suất và nâng cao trải nghiệm cho người hâm mộ. Với mạng lưới phủ sóng rộng khắp và dịch vụ chuyên nghiệp, Opta Sports luôn nỗ lực không ngừng để định hình lại cách thế giới theo dõi, phân tích và tận hưởng thể thao.



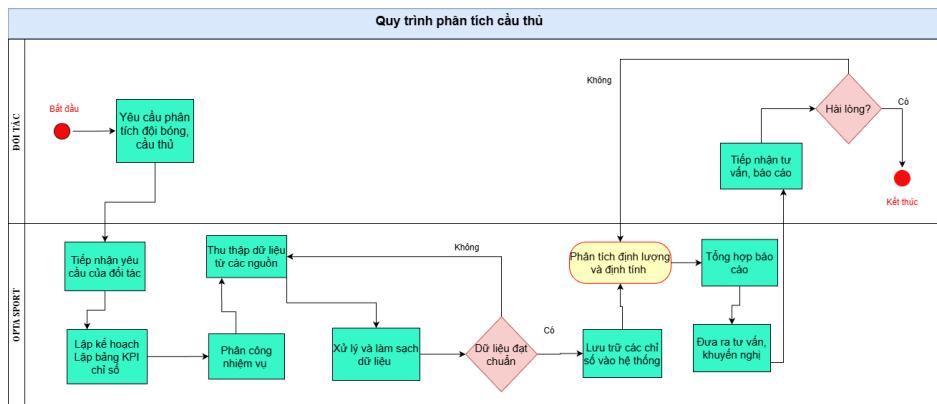
Hình 3.1: Công ty Opta Sports

Các dịch vụ chính của Opta Sports

- Thu thập và phân phối dữ liệu: Ghi lại và cung cấp dữ liệu chi tiết của hàng ngàn trận đấu theo thời gian thực với độ chính xác tuyệt đối.
- Phân tích và Dự đoán (AI): Sử dụng trí tuệ nhân tạo để tạo ra các chỉ số chuyên sâu như Bàn thắng kỳ vọng (xG), Kiến tạo kỳ vọng (xA) và các mô hình dự đoán kết quả trận đấu.

3.2 Mô tả hệ thống

3.2.1 Quy trình nghiệp vụ



Hình 3.2: Quy trình nghiệp vụ

1) Giai đoạn Tiếp nhận và Lập kế hoạch

- Đối tác gửi yêu cầu phân tích cụ thể về đội bóng hoặc cầu thủ mục tiêu.
- Opta Sport tiếp nhận yêu cầu và xác định các mục tiêu phân tích trọng tâm.
- Thiết lập bảng chỉ số KPI (Key Performance Indicators) phù hợp với yêu cầu.
- Phân công nhiệm vụ cho đội ngũ chuyên viên phân tích và kiểm định.

2) Giai đoạn Thu thập và Quản lý Dữ liệu

- Thu thập dữ liệu thô từ các nguồn (video trực tiếp, cảm biến, vệ tinh).
- Thực hiện xử lý, chuẩn hóa và làm sạch dữ liệu để loại bỏ sai sót.
- Kiểm soát chất lượng:** Dánh giá dữ liệu dựa trên các tiêu chuẩn nghiêm ngặt.
 - Nếu không đạt: Quay lại bước thu thập và làm sạch.
 - Nếu đạt: Chuyển sang lưu trữ vào hệ thống cơ sở dữ liệu tập trung.

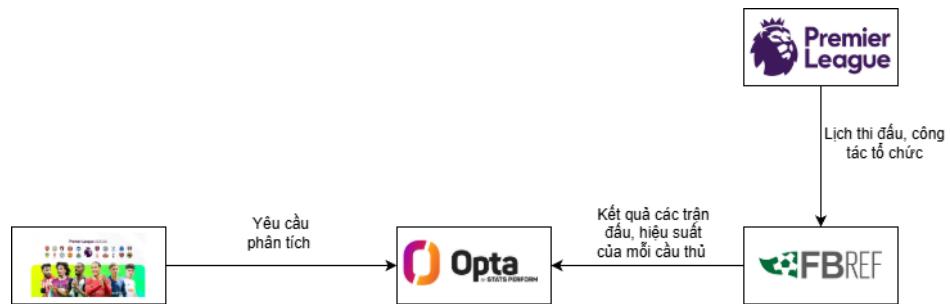
3) Giai đoạn Phân tích chuyên sâu

- Thực hiện phân tích hiệu suất bằng các mô hình toán học và trí tuệ nhân tạo (AI).
- Kết hợp phân tích định tính từ các chuyên gia bóng đá để đánh giá tư duy chiến thuật.
- Tổng hợp các kết quả phân tích thành báo cáo trực quan (Dashboard, Heatmap).

4) Giai đoạn Chuyển giao và Phản hồi

- Đưa ra các tư vấn, khuyến nghị chiến thuật hoặc chuyển nhượng cho đối tác.
- Đối tác tiếp nhận báo cáo và đánh giá mức độ đáp ứng của thông tin.
- **Vòng lặp tối ưu:** Nếu đối tác chưa hài lòng hoặc cần làm rõ, hệ thống sẽ quay lại bước phân tích hiệu suất để đào sâu dữ liệu.
- Kết thúc quy trình khi đạt được sự đồng thuận và hài lòng từ đối tác.

3.2.2 Luồng dữ liệu



Hình 3.3: Luồng dữ liệu

Qua sơ đồ luồng dữ liệu, ta có thể thấy Opta Sports đóng vai trò là nơi cuối cùng lưu trữ dữ liệu, tổng hợp dữ liệu từ nhiều nguồn khác nhau, nhằm tạo ra một cơ sở dữ liệu lưu trữ đầy đủ và toàn diện về các cầu thủ và câu lạc bộ đang thi đấu tại Ngoại Hạng Anh

- Từ Priemer League đến FBRef: Ban tổ chức giải đấu sẽ cung cấp cho FBRef các dữ liệu ban đầu như cặp thi đấu, lịch thi đấu, sân vận động tổ chức.
- Từ FBRef đến Opta Sports: Sau khi nhận được thông tin từ ban tổ chức, FBRef sẽ chuẩn bị công cụ, máy móc và con người để tiến hành thu thập các dữ liệu về hiệu suất trong trận đấu. Sau đó sẽ tổng hợp và gửi dữ liệu được yêu cầu về cho Opta Sports.
- Bên cạnh đó, Opta cũng nhận thông tin, yêu cầu phân tích từ các đối tác để nắm được nhu cầu, đề ra các chỉ số cần thu thập dùng cho việc phân tích.

3.2.3 Hệ thống hiện tại

Hệ thống hiện tại của Opta Sports đã đáp ứng được một số chức năng cơ bản như lưu trữ và quản lý dữ liệu, đã đảm bảo về:

- Tính hiệu quả: Hệ thống cho phép lưu trữ dữ liệu một cách hiệu quả, cho phép các bên đối tác sử dụng, tạo một vài báo cáo đơn giản
- Tính ổn định: Hệ thống hoạt động ổn định trong một thời gian dài mà không gặp nhiều sự cố kỹ thuật
- Tính bảo mật: Hệ thống lưu trữ dữ liệu được cung cấp từ FBRef một cách an toàn, không bị các bên khác tấn công và khai thác sử dụng trái phép

Tuy nhiên hệ thống vẫn còn nhiều hạn chế, tồn đọng một vài vấn đề:

- Hệ thống lưu trữ dữ liệu hiện tại chỉ là các file dữ liệu thô, không nằm trong các cơ sở dữ liệu chuẩn chỉnh, không đảm bảo được tính toàn vẹn khi chia sẻ dữ liệu cho các bên liên quan
- Hệ thống hiện tại chưa có thể tạo các báo cáo chuyên sâu do các dữ liệu được lưu trữ chỉ mang tính thống kê, không thể hỗ trợ để được ra các phân tích chính xác về hiệu suất các cầu thủ, từ đó chưa thể đưa ra quyết định quan trọng cho các bên liên quan

3.2.4 Yêu cầu hệ thống mới

Dựa trên những hạn chế của hệ thống cũ, khi cải thiện hệ thống mới thì cần đảm bảo được một số yêu cầu sau đây:

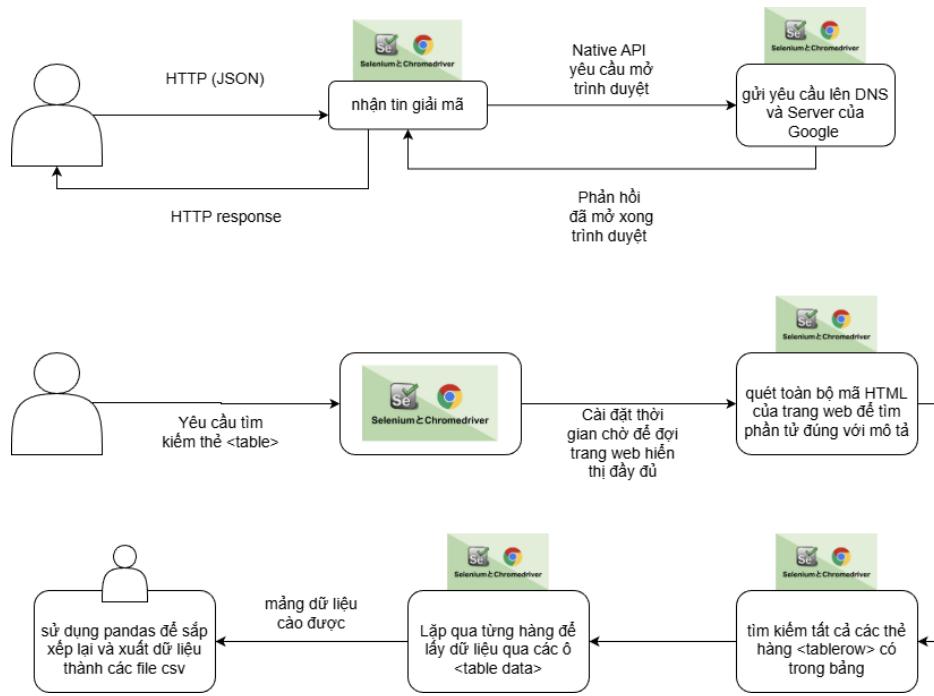
- Cập nhật hệ thống lưu trữ với các cơ sở dữ liệu và kho dữ liệu, đảm bảo tính toàn vẹn khi chia sẻ dữ liệu cho các bên liên quan.
- Đảm bảo được khả năng bảo mật cho các thông tin được các bên cung cấp.
- Tích hợp các công cụ phân tích, đảm bảo việc phân tích các insight quan trọng, cung cấp thông tin kịp thời liên quan đến hiệu suất cầu thủ
- Xây dựng các mô hình dự đoán, dự báo để hỗ trợ ra các quyết định quan trọng, đảm bảo người dùng có được các quyết định đúng đắn nhất
- Đảm bảo được khả năng thu thập dữ liệu liên tục và ổn định, đồng thời cải thiện tốc độ truy vấn dữ liệu, tạo báo cáo thống kê

3.3 Quy trình thu thập và lưu trữ dữ liệu

Dữ liệu của công ty Opta Sports hiện đang được lấy bằng cách cào dữ liệu công khai từ trang FbRef. Bao gồm các thông tin về câu lạc bộ, thông tin cầu thủ và hiệu suất của cầu thủ qua mỗi trận.

3.3.1 Quy trình cào dữ liệu

Ta sử dụng selenium để cào dữ liệu từ trang web FbRef. Việc trang web FBRef thường hay xuất hiện pop-up cookies thì việc sử dụng Selenium là cách hợp lý nhất để cào được dữ liệu từ trang web về.



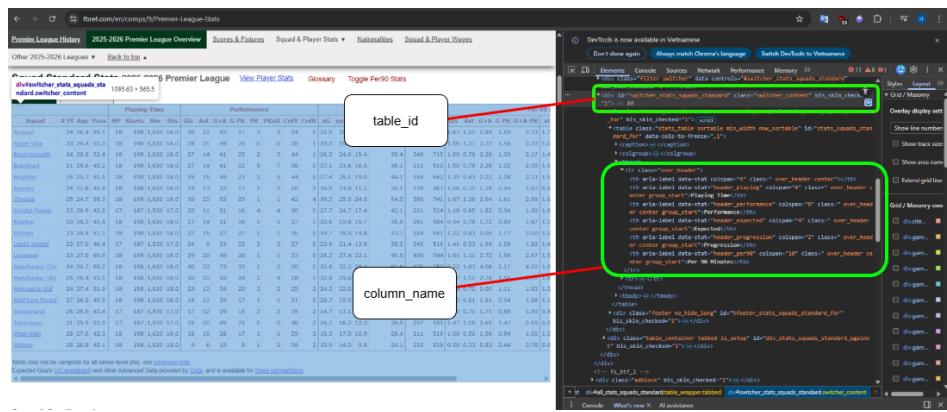
Hình 3.4: Quy trình cào dữ liệu

- Khởi tạo và kết nối: Ban đầu, người dùng thông qua các đoạn code gửi một yêu cầu dưới dạng HTTP (JSON) đến WEBDriver (ở đây là Chrome), hệ thống nhận được yêu cầu, tiến hành giải mã, sau đó thông qua native API để ra lệnh cho ChromeDriver mở trình duyệt. Trình duyệt gửi yêu cầu đến DNS và Server của Google để tải trang web. Sau khi tải được trang web, thông báo phản hồi sẽ được gửi về cho Client để xác nhận trạng thái sẵn sàng
- Điều hướng và quét dữ liệu: Người dùng gửi yêu cầu tìm kiếm thẻ <table> trong đoạn code HTML của trang web. Chương trình được cài đặt sử dụng thời gian chờ để chờ trang web hiển thị lên đầy đủ, sau đó selenium quét toàn bộ mã nguồn để xác định đúng phần tử khớp với mô tả của thẻ <table>, ví dụ như ID, Class, hoặc XPath của bảng.
- Trích xuất và lưu trữ: Sau khi tìm được các bảng, tiếp tục sẽ tìm kiếm các thẻ hàng, truy cập vào các ô dữ liệu để lấy dữ liệu về. Toàn bộ dữ liệu sau khi quét sẽ được gom lại thành một mảng dữ liệu, cuối cùng sử dụng thư viện pandas để tạo thành dataframe, tại đây dữ liệu sẽ được sắp xếp và lưu ra file CSV để sử dụng.

3.3.2 Về thông tin của các câu lạc bộ

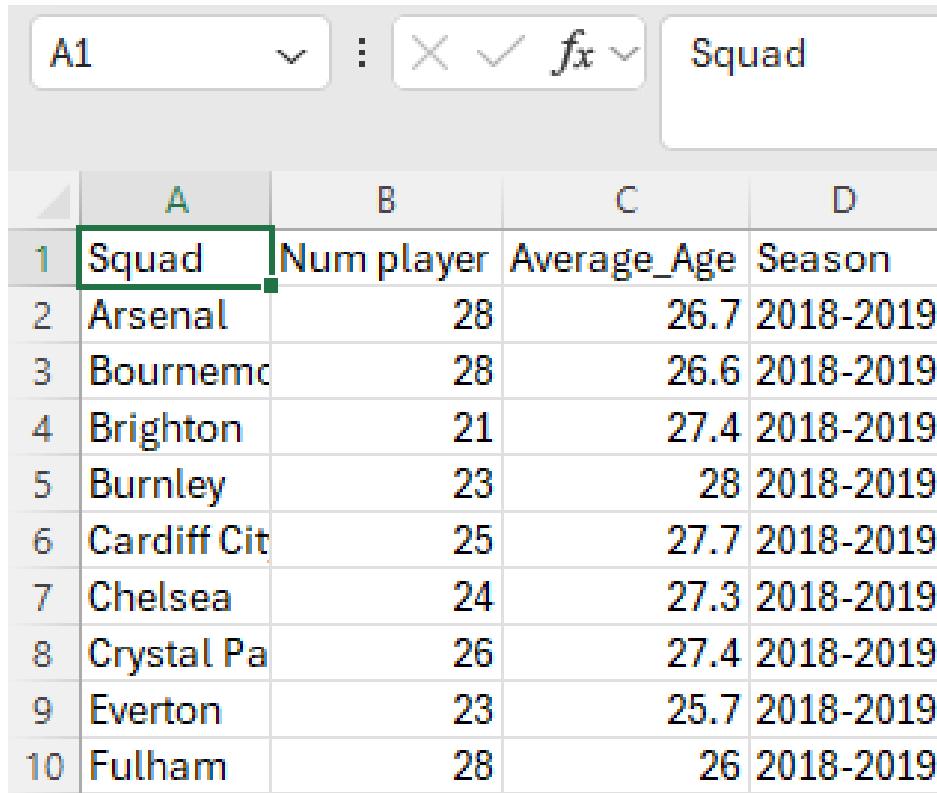
Ta tiến hành cào dữ liệu về thông tin các câu lạc bộ thi đấu tại giải đấu Ngoại hạng Anh từ năm 2018 đến thời điểm hiện tại:

- Sử dụng vòng lặp để cào dữ liệu với năm bắt đầu là 2018 và năm kết thúc là 2025. Khi cào đến năm nào thì trong dataframe sẽ có thêm 1 cột là cột của mùa giải đó.
- Tìm đến thẻ `<table>` có ID là "squad_stat_standard", sau đó tìm các thông tin trong bảng và lấy về.



Hình 3.5: Quy trình cào dữ liệu về đội bóng

- Trích xuất dữ liệu vừa cào được và lưu về 1 file csv ở trong máy tính.



	A	B	C	D
1	Squad	Num player	Average_Age	Season
2	Arsenal	28	26.7	2018-2019
3	Bournemouth	28	26.6	2018-2019
4	Brighton	21	27.4	2018-2019
5	Burnley	23	28	2018-2019
6	Cardiff City	25	27.7	2018-2019
7	Chelsea	24	27.3	2018-2019
8	Crystal Palace	26	27.4	2018-2019
9	Everton	23	25.7	2018-2019
10	Fulham	28	26	2018-2019

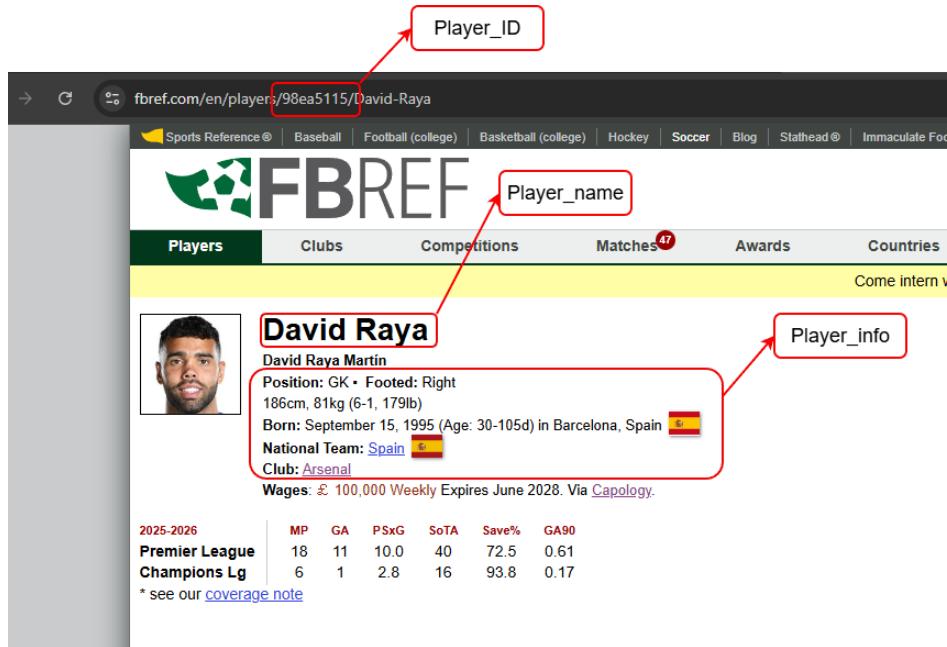
Hình 3.6: Kết quả cào được về club info

- Quá trình cào dữ liệu về thông tin các câu lạc bộ sẽ được lặp lại vào mỗi đầu mùa giải.

3.3.3 Về thông tin cầu thủ

Ta tiến hành cào dữ liệu về thông tin các cầu thủ, ở đây với dữ liệu trang web cung cấp thì sẽ là những cầu thủ ra sân ít nhất 1 phút trong khoảng thời gian các mùa giải từ năm 2018 đến bây giờ:

- Sử dụng vòng lặp để cào dữ liệu với năm bắt đầu là 2018 và năm kết thúc là 2025. Tuy nhiên sẽ có nhiều cầu thủ thi đấu ở giải từ năm này qua năm khác, cho nên với mỗi player được cào sẽ lưu cả ID của mỗi cầu thủ, khi cào mùa giải sau thì sẽ kiểm tra các ID đã có sẵn trước, nếu chưa có thì sẽ cào mới.
- Sau khi mở xong trang web, tiến hành tìm các đoạn mã HTML tương ứng với từng thông tin cầu thủ để tiến hành lấy dữ liệu về



Hình 3.7: Quy trình cào dữ liệu về cầu thủ

- Trích xuất dữ liệu cào được và lưu về 1 file csv ở trong máy tính.

A	B	C	D	E	F	G	
1	Player ID	Name	Born Date	National Team	Footed	Height (cm)	Weight (kg)
2	5f09991f	Patrick van	1990-08-25	Netherlands	Left	175	67
3	4d034881	Sergio Agüero		Argentina	Right	172	68
4	eaeca114	Nathan Ake	1995-02-18	Netherlands	Left	180	74
5	b827d5b3	Marc Albrighton		England	Right	175	67
6	f7d50789	Toby Alderweireld		Belgium	Right	187	79
7	cd1acf9d	Trent Alexander-Arnold	1998-10-07	England	Right	175	68
8	7a2e46a8	Alisson	1992-10-02	Brazil	Right	193	91
9	cea4ee8f	Dele Alli		England	Right	188	76
10	862a1c15	Miguel Almiron	1994-02-10	Paraguay	Left	174	69

Hình 3.8: Kết quả cào được về player info

- Quá trình cào dữ liệu về thông tin các cầu thủ sẽ được lặp lại hàng tuần sau mỗi vòng đấu.

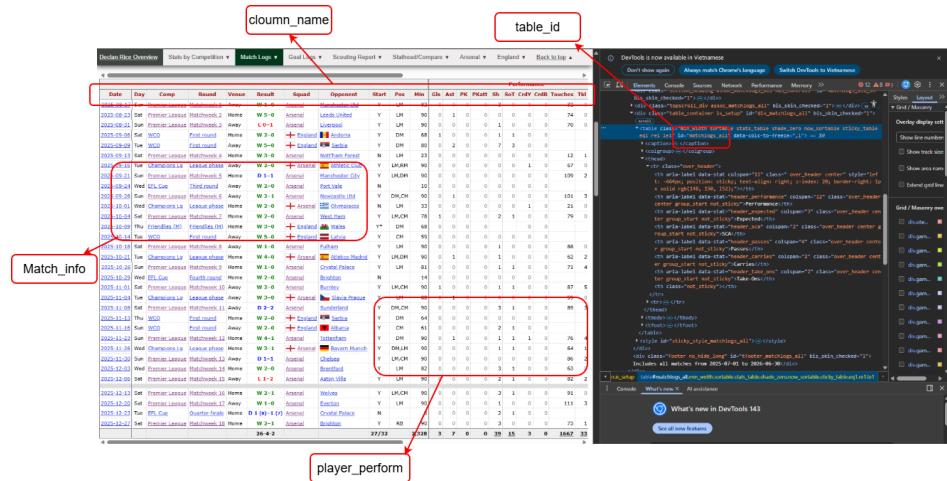
3.3.4 Về hiệu suất các cầu thủ trong trận đấu

Ta tiến hành cào dữ liệu về hiệu suất các cầu thủ trong mỗi trận đấu. Dữ liệu được cung cấp ở đây là hiệu suất của mỗi cầu thủ qua từng trận, chứ không phải mỗi trận có những cầu thủ nào, cho nên sẽ có những thông tin bị trùng lặp là thông số của các trận, sẽ xử lý sau:

- Sử dụng vòng lặp để cào dữ liệu với năm bắt đầu là 2018 và năm kết

thúc là 2025. Với mỗi cầu thủ thì chúng ta file đi qua 6 links tương ứng với 6 chỉ số được thu thập "standard", "passing", "passtype", "goal and shot", "possession", "defend actions" và thêm dữ liệu "goalkeeper" của thủ môn.

- Sau khi mở xong trang web, tiến hành tìm đến các thẻ `<table>` trong mỗi chỉ số để cào được dữ liệu.



Hình 3.9: Quy trình cào dữ liệu hiệu suất cầu thủ

- Trích xuất dữ liệu cào được và lưu vào các file, với mỗi file là 1 loại chỉ số tương ứng.

Name	Date modified	Type	Size
all_gca_premier_league_matchlogs_final.csv	12/19/2025 3:21 AM	Microsoft Excel Comma ...	12,672 KB
ETLipynb	12/24/2025 2:04 AM	Jupyter Source File	56 KB
player_defensive_actions_matchlogs.csv	12/19/2025 3:21 AM	Microsoft Excel Comma ...	13,012 KB
player_goalkeeper_matchlogs.csv	12/19/2025 3:20 AM	Microsoft Excel Comma ...	1,136 KB
player_pass_types_matchlogs.csv	12/19/2025 3:18 AM	Microsoft Excel Comma ...	13,006 KB
player_passing_matchlogs.csv	12/19/2025 3:18 AM	Microsoft Excel Comma ...	15,373 KB
player_possession_matchlogs.csv	12/19/2025 3:17 AM	Microsoft Excel Comma ...	14,832 KB
players_standard_matchlogs.csv	12/19/2025 3:16 AM	Microsoft Excel Comma ...	15,095 KB

Hình 3.10: Quy trình cào dữ liệu hiệu suất cầu thủ

- Quá trình cào dữ liệu hiệu suất của các cầu thủ sẽ được lặp lại hàng tuần sau mỗi vòng đấu.

3.4 Khai phá dữ liệu

3.4.1 Tổng quan về các bảng

Bảng 3.1: Tổng quan về các bảng dữ liệu thô

Tên bảng	Dung lượng	Hàng	Cột	Tần suất cập nhật
Squad_Standard	21KB	141	33	Hàng năm
Squad_Defend	13 KB	141	20	Hàng năm
Squad_GCA	11 KB	141	20	Hàng năm
Squad_goalkeeping	12 KB	141	22	Hàng năm
Squad_pass_type	13 KB	141	19	Hàng năm
Squad_passing	20 KB	141	27	Hàng năm
Squad_possession	20 KB	141	27	Hàng năm
Player_info	84 KB	1588	7	Hàng tuần
Player_Standard	6873 KB	81086	26	Hàng tuần
Player_Defend	4193 KB	80998	17	Hàng tuần
Player_GCA	3269 KB	81070	15	Hàng tuần
Player_goalkeeping	575 KB	5702	26	Hàng tuần
Player_pass_type	4348 KB	80779	16	Hàng tuần
Player_passing	7392 KB	81380	23	Hàng tuần
Player_possession	6921 KB	81091	23	Hàng tuần

Một số trường dữ liệu quan trọng:

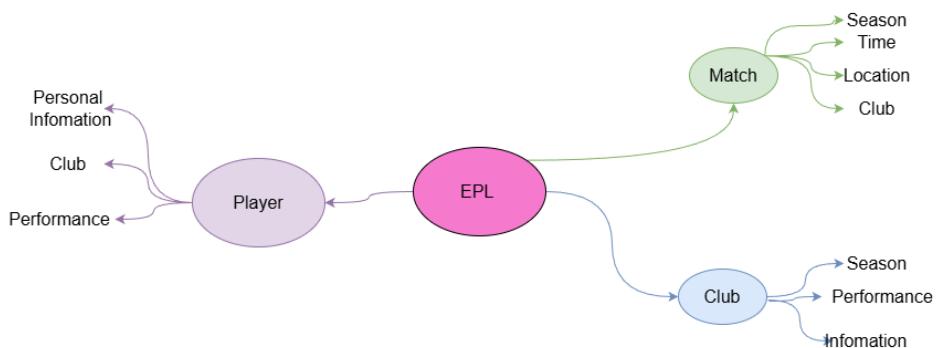
- Squad: thông tin về tên câu lạc bộ
- Season: Thông tin về mùa giải thi đấu
- Num Player: Số lượng cầu thủ của đội bóng
- Average Age: Thông tin về độ tuổi trung bình của câu lạc bộ.
- Player ID: Thông tin về mã cầu thủ.
- Name: Thông tin về tên của cầu thủ.
- Born Date: Thông tin về ngày tháng năm sinh của cầu thủ.
- National Team: Thông tin về quốc tịch của cầu thủ.
- Footed: Thông tin về chân thuận của cầu thủ.

- Height: Thông tin về chiều cao của cầu thủ.
- Weight: Thông tin về cân nặng của cầu thủ.
- Date: Thông tin về ngày diễn ra trận đấu.
- Round: Thông tin về vòng thi đấu của trận đấu đó.
- Result: Thông tin về kết quả của trận thi đấu.
- Opponent: Thông tin về đối thủ của trận thi đấu.
- Start: Thông tin về việc cầu thủ có ra sân ngay từ đầu hay không.
- Pos: Thông tin về vị trí thi đấu của cầu thủ.
- **Phân cấp dữ liệu:** Bộ dữ liệu được phân chia rõ rệt thành hai cấp độ: cấp độ đội bóng (*Squad*) và cấp độ cầu thủ (*Player*). Điều này cho thấy cấu trúc cơ sở dữ liệu được thiết kế để phục vụ cả phân tích tổng quan lẫn chi tiết cá nhân.
- **Quy mô và Dung lượng:**
 - Các bảng nhóm **Squad** có quy mô nhỏ, đồng nhất về số hàng (141 hàng), dung lượng thấp (dưới 25KB), tập trung vào các chỉ số tổng kết.
 - Các bảng nhóm **Player** có quy mô cực lớn, với số lượng hàng lên tới hơn 81.000 và dung lượng đạt mức MB (cao nhất là *Player_passing* với 7392 KB). Điều này phản ánh khối lượng thông tin chi tiết khổng lồ được thu thập cho từng cá nhân qua mỗi trận đấu.
- **Tính biến động và Cập nhật:**
 - Nhóm dữ liệu đội bóng chỉ cập nhật **hàng năm**, phù hợp cho việc lưu trữ lịch sử hoặc tổng kết sau mỗi mùa giải.
 - Nhóm dữ liệu cầu thủ được cập nhật **hàng tuần**, cho thấy đây là dữ liệu động, phục vụ việc theo dõi phong độ tức thời, phân tích chiến thuật và cập nhật trạng thái cầu thủ theo thời gian thực của mùa giải.

- **Độ phức tạp:** Các bảng như *Squad_Standard* (33 cột) và *Player_Standard* (26 cột) chứa nhiều trường thông tin nhất, đóng vai trò là các bảng dữ liệu gốc quan trọng trong hệ thống thống kê bóng đá này.

3.4.2 Data Taxonomy

Data Taxonomy là một hệ thống tổ chức và phân loại các dữ liệu thành các nhóm, lớp hoặc loại dựa trên các đặc điểm và thuộc tính chung của chúng. Mục tiêu của việc xây dựng data taxonomy là để giúp người dùng dễ dàng tìm kiếm, quản lý, và sử dụng dữ liệu một cách có hệ thống và hiệu quả hơn.

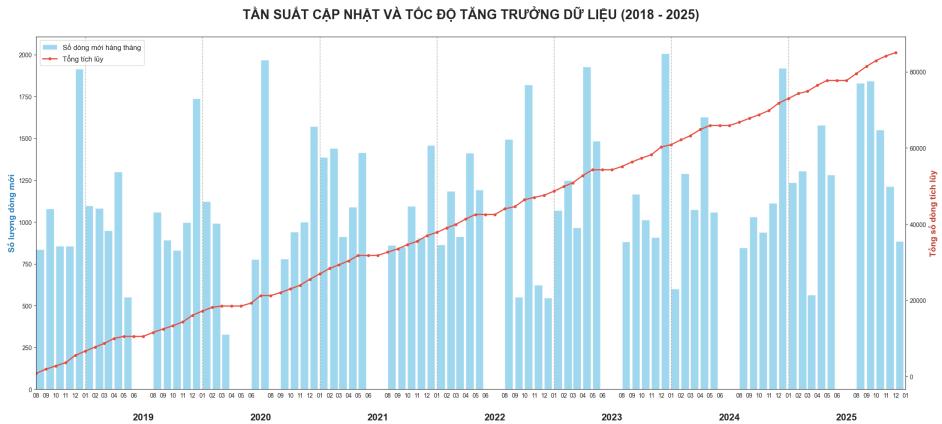


Hình 3.11: Data taxonomy

Dựa vào hình trên, ta có thể thấy bộ dữ liệu thu thập được xoay quanh chủ điểm về giải đấu English Premier League và được chia thành 3 nhóm gồm : Nhóm cầu thủ, nhóm trận đấu và nhóm đội bóng. Trong các nhóm chính đó sẽ tiếp tục được chia thành các nhóm nhỏ hơn, là cơ sở để chúng ta xây dựng hệ thống dimension sau này.

3.4.3 Tần suất cập nhật dữ liệu

Ta tiến hành đo tốc độ tăng trưởng của bộ dữ liệu, bộ dữ liệu chủ yếu được cập nhật hàng tuần với 7 bảng hiệu suất của cầu thủ. Bảng *Player_Standard* có thể xem như là nhật ký của cả bộ dữ liệu, tần suất cập nhật hàng tuần tương đương với các bảng còn lại, vậy ta tiến hành khảo sát bảng này, từ đó có thể đo được tần suất của cả bộ dữ liệu:



Hình 3.12: Tốc độ tăng trưởng dữ liệu bảng Player _ Standard

Dựa trên biểu đồ kết hợp giữa cột (lượng dữ liệu mới) và đường (tổng tích lũy), ta có các nhận xét chi tiết sau:

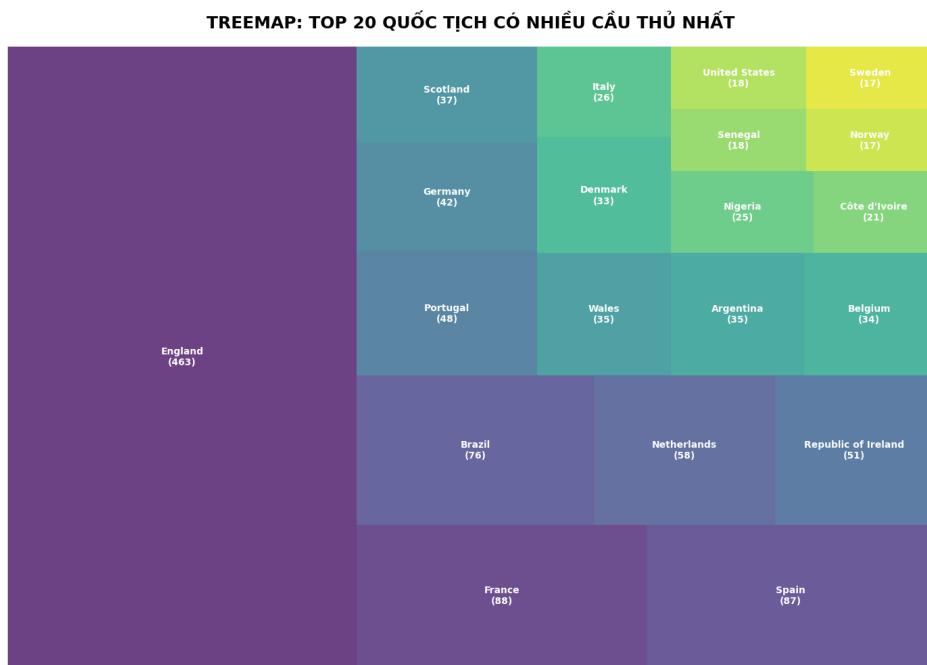
- **Xu hướng tăng trưởng bền vững:** Đường tổng tích lũy (màu đỏ) duy trì độ dốc đi lên ổn định trong suốt giai đoạn 7 năm. Đến cuối năm 2025, hệ thống đã ghi nhận tổng cộng hơn **80.000 dòng dữ liệu**, cho thấy một kho lưu trữ dữ liệu lịch sử có chiều sâu và giá trị phân tích cao.
- **Đặc thù chu kỳ mùa giải:** Biểu đồ phản ánh rõ nét đặc thù của ngành bóng đá thông qua các *khoảng trống dữ liệu* định kỳ vào tháng 6 và tháng 7 hàng năm. Đây là giai đoạn nghỉ giữa hai mùa giải (off-season), khi không có các trận đấu chính thức diễn ra, dẫn đến việc không phát sinh dữ liệu mới.
- **Biến động dữ liệu hàng tháng:**
 - Lượng dòng mới hàng tháng (cột màu xanh) dao động mạnh, thường đạt đỉnh vào các giai đoạn cao điểm của mùa giải (tháng 8-9 và tháng 1-2) với mức từ **1.500 đến gần 2.000 dòng/tháng**.
 - Giai đoạn từ năm 2023 đến 2025 cho thấy mật độ các cột xanh dày hơn và cao hơn so với giai đoạn đầu (2018-2019), minh chứng cho việc mở rộng quy mô thu thập dữ liệu hoặc tăng cường độ chi tiết trong việc ghi chép các chỉ số.
- **Độ tin cậy của hệ thống:** Việc dữ liệu được cập nhật đều đặn ngay khi mùa giải bắt đầu trở lại cho thấy quy trình vận hành và thu thập

dữ liệu có tính kỷ luật cao, đảm bảo tính cập nhật liên tục cho các mô hình dự báo hoặc báo cáo thống kê.

Với tổng toàn bộ dữ liệu có mức độ tăng trưởng gấp 7 lần như thế này thì công cụ sử dụng để tiến hành xây dựng Data Warehouse hay xử lý dữ liệu phù hợp là Python.

3.4.4 Khám phá dữ liệu

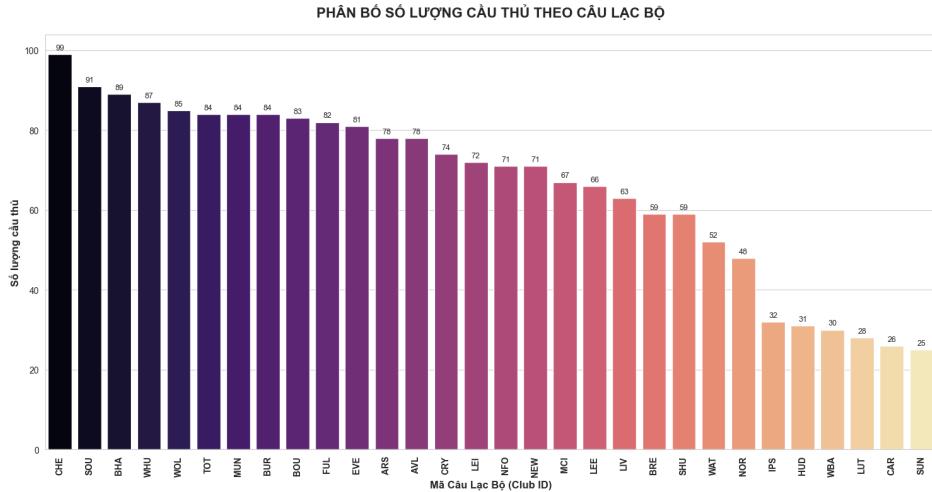
1) Phân bố cầu thủ theo quốc gia



Hình 3.13: Phân bố cầu thủ theo quốc gia

- Số lượng các cầu thủ có quốc tịch Anh là nhiều nhất, bởi vì ở đây chúng ta đang có dữ liệu về giải thi đấu bóng đá ở Anh
- Các quốc gia khác cũng có nhiều cầu thủ thi đấu ở đây, đặc biệt là các quốc gia châu Âu như Pháp, Tây Ban Nha...

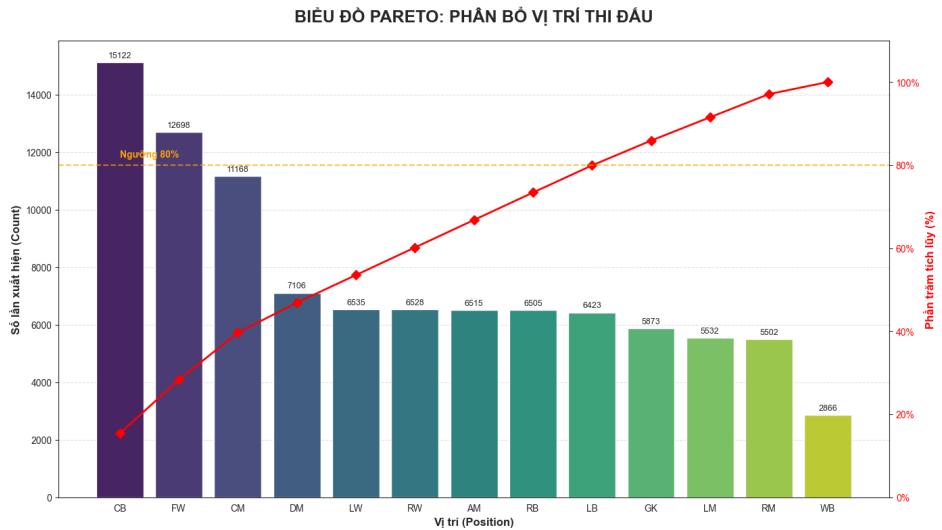
2) Phân bố cầu thủ theo câu lạc bộ



Hình 3.14: Phân bố cầu thủ theo đội bóng

- Các đội bóng có nhiều cầu thủ từng thi đấu qua như Chelsea, Southampton là những đội thường xuyên thay đổi đội hình, mua bán cầu thủ liên tục
- Có những đội bóng mới chỉ xuất hiện 1 mùa nên số lượng cầu thủ được thống kê thi đấu cho đội bóng đó khá ít, ví dụ như Sunderland, Luton Town

3) Phân bố theo vị trí thi đấu

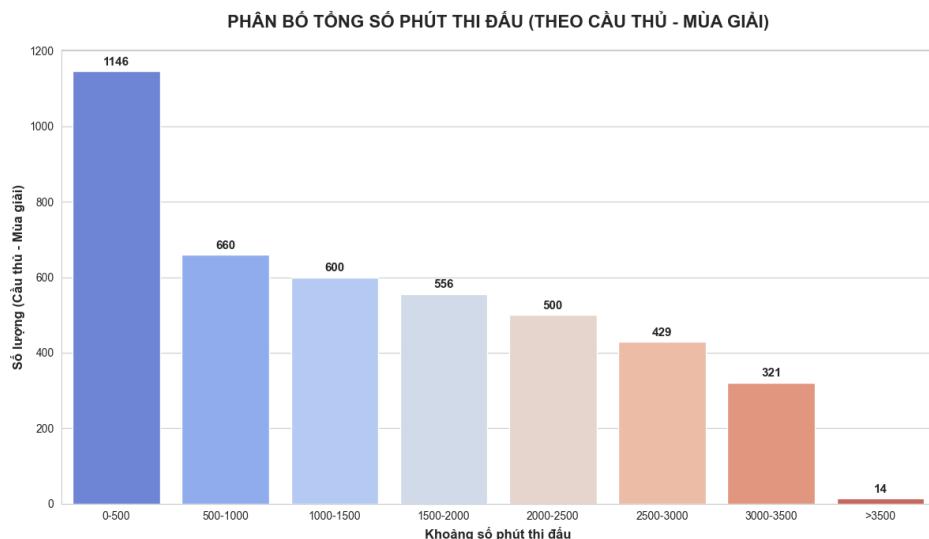


Hình 3.15: Phân bố cầu thủ theo vị trí thi đấu

- Nhóm vị trí chủ đạo: CB, FW và CM là ba vị trí có tần suất xuất hiện cao nhất, đóng vai trò nòng cốt và chiếm tỷ trọng lớn nhất trong bộ dữ liệu.

- Quy luật tích lũy: Ngưỡng 80% dữ liệu tập trung ở 9 nhóm vị trí đầu tiên (từ CB đến LB), cho thấy sự phân bổ tập trung vào các vai trò thi đấu phổ biến.
- Vị trí đặc thù: WB (Hậu vệ biên tấn công) có tần suất thấp nhất, phản ánh đây là vị trí chuyên biệt hoặc ít được sử dụng nhất trong các sơ đồ chiến thuật.

4) Phân bổ theo thời gian thi đấu



Hình 3.16: Phân bố cầu thủ theo thời gian thi đấu

- Tập trung ở nhóm ít phút: Đa số cầu thủ (1146 trường hợp) thi đấu dưới 500 phút, cho thấy một lượng lớn nhân sự trong bộ dữ liệu thuộc nhóm dự bị hoặc có thời gian ra sân hạn chế.
- Xu hướng giảm dần: Số lượng cầu thủ giảm dần tỷ lệ nghịch với số phút thi đấu, minh chứng cho sự cạnh tranh khốc liệt để có suất đá chính thường xuyên.
- Nhóm trụ cột đặc biệt: Chỉ có duy nhất 14 trường hợp thi đấu trên 3500 phút, đại diện cho nhóm cầu thủ nòng cốt, có thể lực và phong độ cực kỳ ổn định xuyên suốt mùa giải.

5) Đặc trưng về thông tin cầu thủ

- Bộ dữ liệu có quy mô mẫu lớn với hơn 1.300 quan sát, mang lại độ tin cậy thống kê cao cho kết quả phân tích. Sai số chuẩn rất

Bảng 3.2: Bảng thống kê về chiều cao và cân nặng

Thông số	Height (cm)	Weight (kg)
Mean	181.983	74.082
Standard Error	0.183	0.192
Median	182.000	73.000
Mode	180.000	69.000
Standard Deviation	6.969	7.141
Sample Variance	48.560	50.993
Kurtosis	-0.554	0.129
Skewness	0.077	0.278
Range	39.000	59.000
Minimum	163.000	52.000
Maximum	202.000	111.000
Sum	263,148.000	102,381.000
Count	1,446.000	1,382.000
Confidence Level(95.0%)	0.359	0.377

thấp và khoảng tin cậy 95% hẹp cho thấy các giá trị trung bình tính toán được có độ chính xác lớn và đại diện tốt cho tổng thể.

- Chiều cao trung bình đạt 181,98 cm, cho thấy đối tượng khảo sát có thể hình cao lớn. Do các chỉ số Mean, Median và Mode xấp xỉ nhau cùng hệ số Skewness gần bằng 0 (0,077), phân phối của chiều cao mang tính đối xứng cao và tiệm cận phân phối chuẩn.
- Cân nặng trung bình là 74,08 kg với xu hướng lệch phải nhẹ ($Skewness = 0,278$), phản ánh sự hiện diện của một số cá thể có trọng lượng vượt trội trong mẫu. Nhìn chung, sự kết hợp giữa chiều cao và cân nặng tạo ra chỉ số BMI lý tưởng (khoảng 22,36), cho thấy thể trạng của nhóm mẫu rất cân đối.

6) Đặc trưng về hiệu suất thi đấu

- Với quy mô mẫu cực lớn (gần 68.000 quan sát), các kết quả thống kê có độ tin cậy rất cao và sai số chuẩn cực thấp. Chỉ số phút thi đấu (Min) có Median và Mode đều đạt 90 cùng hệ số Skewness âm

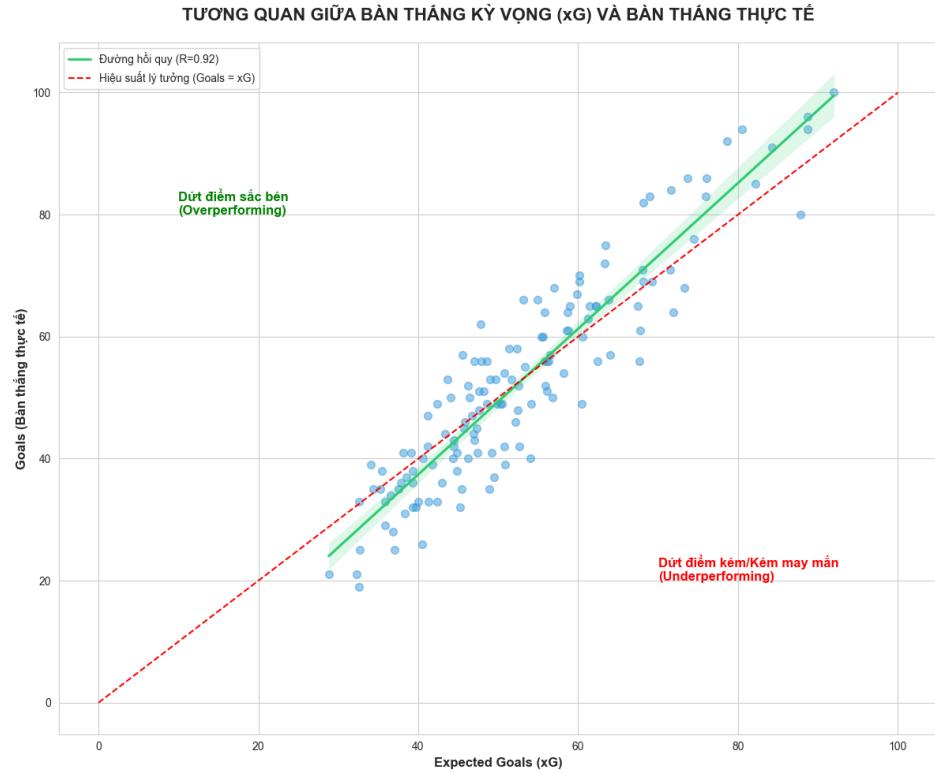
Bảng 3.3: Bảng thống kê các hiệu số

Thông số	Min	Touches	Cmp%
Mean	81.897	50.391	76.593
Standard Error	0.056	0.088	0.049
Median	90.000	47.000	78.400
Mode	90.000	39.000	75.000
Standard Deviation	14.678	22.996	12.742
Sample Variance	215.439	528.821	162.363
Kurtosis	2.564	0.750	1.153
Skewness	-1.880	0.819	-0.891
Range	59.000	191.000	100.000
Minimum	31.000	2.000	0.000
Maximum	90.000	193.000	100.000
Sum	5,573,832.000	3,426,522.000	5,208,003.500
Count	68,059.000	67,999.000	67,996.000
Confidence Level(95.0%)	0.110	0.173	0.096

(-1,88) cho thấy đa số cầu thủ trong mẫu thi đấu trọn vẹn cả trận, chỉ một số ít bị thay ra sớm hoặc vào sân từ ghế dự bị.

- Chỉ số Touches (chạm bóng) có sự biến thiên rất rộng từ 2 đến 193 lần, với phân phối lệch phải rõ rệt (Skewness = 0,819). Điều này phản ánh thực tế trên sân cỏ khi phần lớn cầu thủ có mức chạm bóng trung bình (khoảng 47-50 lần), trong khi một nhóm nhỏ các tiền vệ kiến thiết hoặc hậu vệ triển khai bóng có số lần chạm bóng vượt trội hẳn so với phần còn lại.
- Tỷ lệ chuyền bóng chính xác (Cmp%) đạt mức trung bình khá ổn tương là 76,59%, với Median đạt tới 78,4%. Phân phối của biến này lệch trái (-0,891), cho thấy mặt bằng chung kỹ thuật của các cầu thủ là khá tốt và ổn định, phần lớn đều thực hiện chính xác trên 3/4 số đường chuyền của mình.

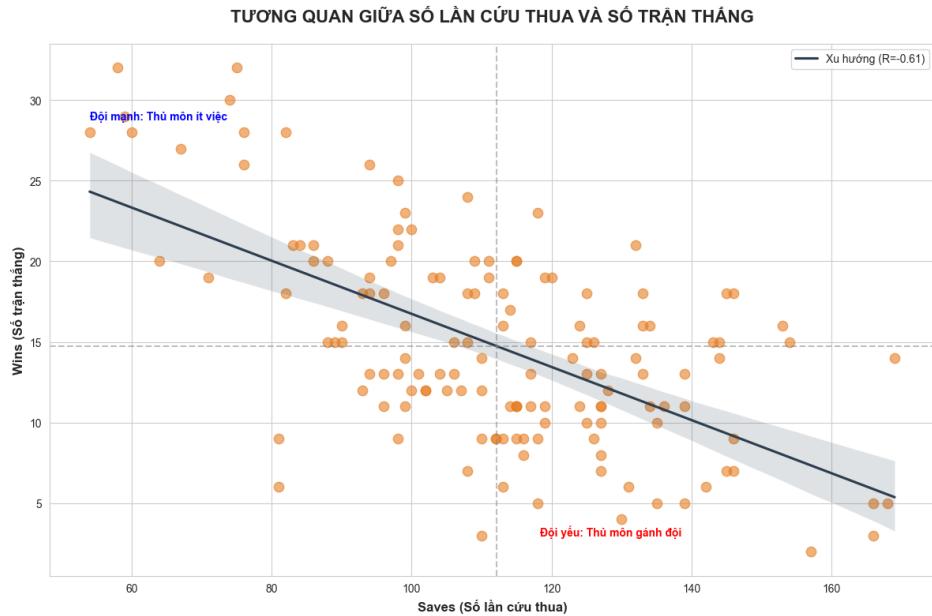
7) Tương quan giữa bàn thắng kỳ vọng và bàn thắng thực tế



Hình 3.17: Tương quan giữa bàn thắng kỳ vọng và bàn thắng thực tế

- Tương quan cực kỳ mạnh mẽ: Hệ số tương quan $R = 0.92$ khẳng định mối quan hệ tuyến tính rất chặt chẽ; chỉ số bàn thắng kỳ vọng (xG) là một thước đo có độ tin cậy cực cao để dự báo kết quả ghi bàn thực tế.
- Sự phân hóa hiệu suất: Các điểm dữ liệu tập trung quanh đường lý tưởng nhưng vẫn có sự phân hóa rõ rệt: nhóm nằm trên đường đỏ thể hiện kỹ năng dứt điểm sắc bén (Overperforming), ngược lại nhóm nằm dưới đang dứt điểm kém hiệu quả hoặc thiếu may mắn (Underperforming).
- Độ tin cậy của mô hình: Đường hồi quy màu xanh tiệm cận sát với đường lý tưởng $Goals = xG$, cho thấy mô hình tính toán xG đang sử dụng rất sát với thực tế khách quan của các trận đấu.

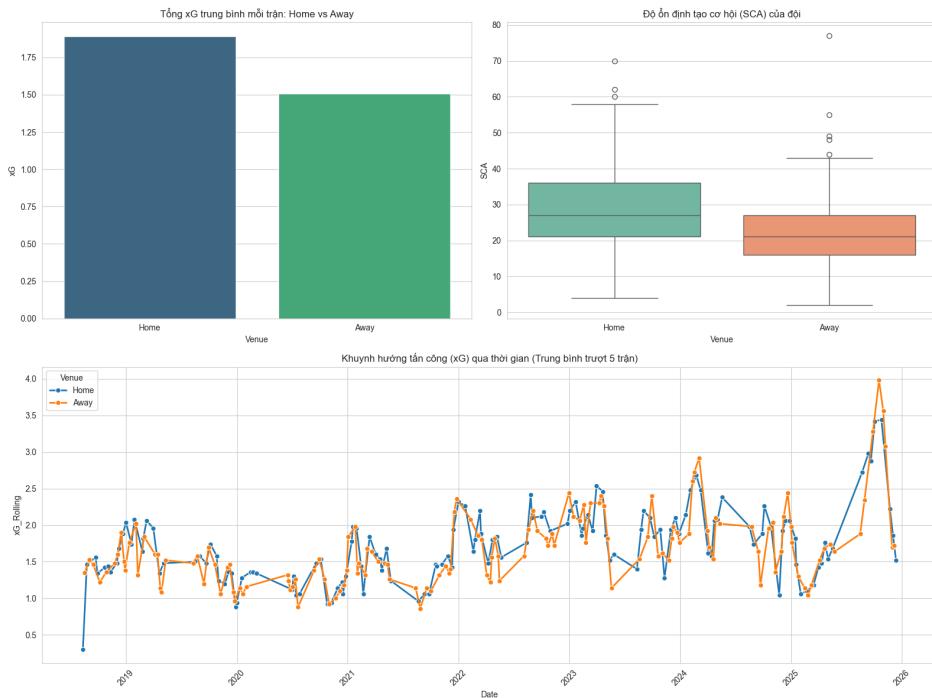
8) Tương quan giữa số lần cứu thua và số trận thắng



Hình 3.18: Tương quan giữa số lần cứu thua và số trận thắng

- Tương quan nghịch biến rõ rệt: Hệ số $R = -0.61$ cho thấy mối quan hệ nghịch biến giữa hai đại lượng. Các đội bóng có số trận thắng cao thường kiểm soát thế trận tốt hơn, dẫn đến việc thủ môn ít phải thực hiện các pha cứu thua.
- Sự phân hóa giữa các nhóm đội:
 - Nhóm đội mạnh: Tập trung ở góc trên bên trái với tỷ lệ thắng cao và số lần cứu thua thấp, minh chứng cho một hàng phòng ngự vững chắc.
 - Nhóm đội yếu: Tập trung ở góc dưới bên phải, nơi thủ môn phải "gánh đội" với số lần cứu thua cực lớn nhưng vẫn không đủ để bù đắp cho kết quả trận đấu.
- Giá trị phản ánh: Số lần cứu thua cao thường là hệ quả của việc hàng phòng ngự để đổi phương dứt điểm quá nhiều, thay vì là một chỉ số tích cực phản ánh khả năng giành chiến thắng của toàn đội.

9) Khuynh hướng thi đấu theo địa điểm của các đội bóng

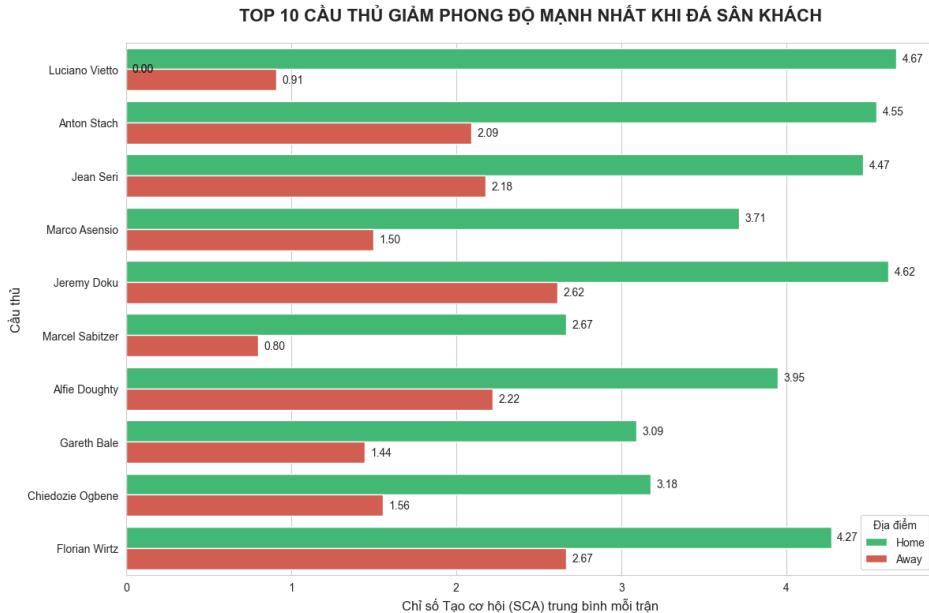


Hình 3.19: Khuynh hướng thi đấu theo địa điểm của các đội bóng

- **Ưu thế sân nhà (Home Advantage):** Lợi thế sân bãi được thể hiện rõ nét qua chỉ số xG trung bình tại sân nhà (1.9) vượt trội so với sân khách (1.5). Điều này cho thấy đội bóng luôn chủ động áp đặt lối chơi và tạo ra nhiều cơ hội nguy hiểm hơn khi được thi đấu tại tổ ấm.
- **Độ ổn định và khả năng bùng nổ:** Biểu đồ Boxplot về SCA chỉ ra rằng tại sân nhà, đội bóng không chỉ có mức hiệu suất trung bình cao hơn mà còn có khả năng tạo ra các trận đấu "hủy diệt" với số lượng cơ hội cực lớn (các điểm ngoại lai đạt ngưỡng 70-80 SCA). Trong khi đó, hiệu suất sân khách thường bị giới hạn trong phạm vi hẹp và thấp hơn.
- **Xu hướng phát triển dài hạn:** Nhìn vào biểu đồ trung bình trượt (Rolling Average), có thể thấy một lộ trình cải thiện chất lượng tấn công xuyên suốt từ năm 2018.
 - Giai đoạn 2019 - 2021: Duy trì sự ổn định ở mức trung bình.
 - Giai đoạn 2022 - 2024: Bắt đầu có những nhịp tăng trưởng rõ rệt.
 - Giai đoạn cuối 2025 - đầu 2026: Chúng kiến sự bùng nổ mạnh

mẽ nhất trong lịch sử dữ liệu khi chỉ số xG đạt đỉnh gần 4.0, minh chứng cho một hàng công đang ở trạng thái cực thịnh.

10) Khuynh hướng thi đấu theo địa điểm của các cầu thủ

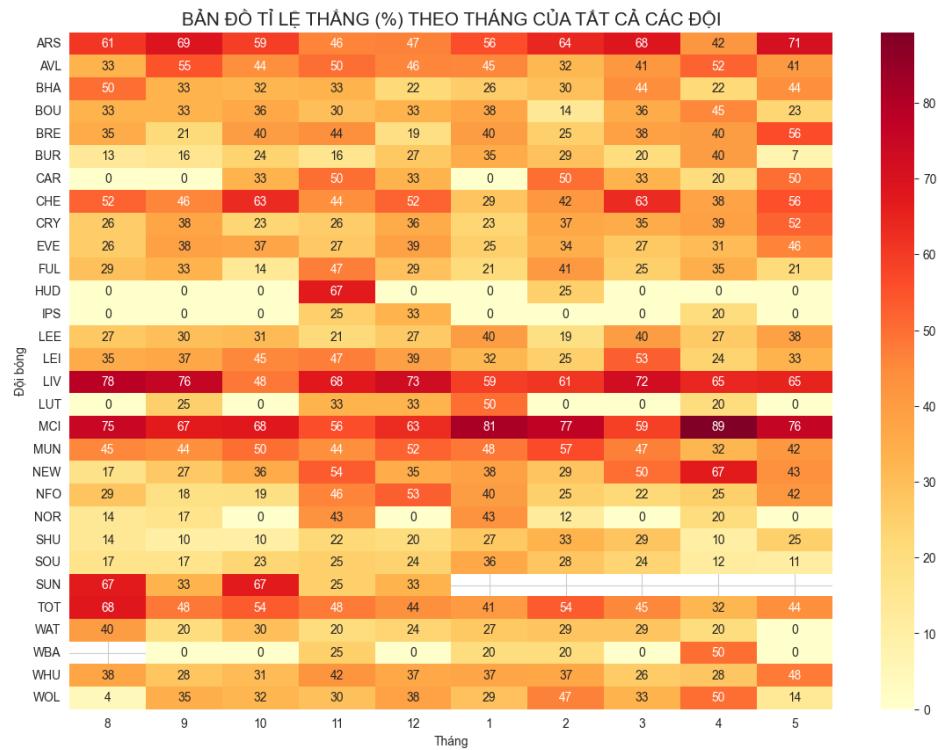


Hình 3.20: Khuynh hướng thi đấu theo địa điểm của các cầu thủ

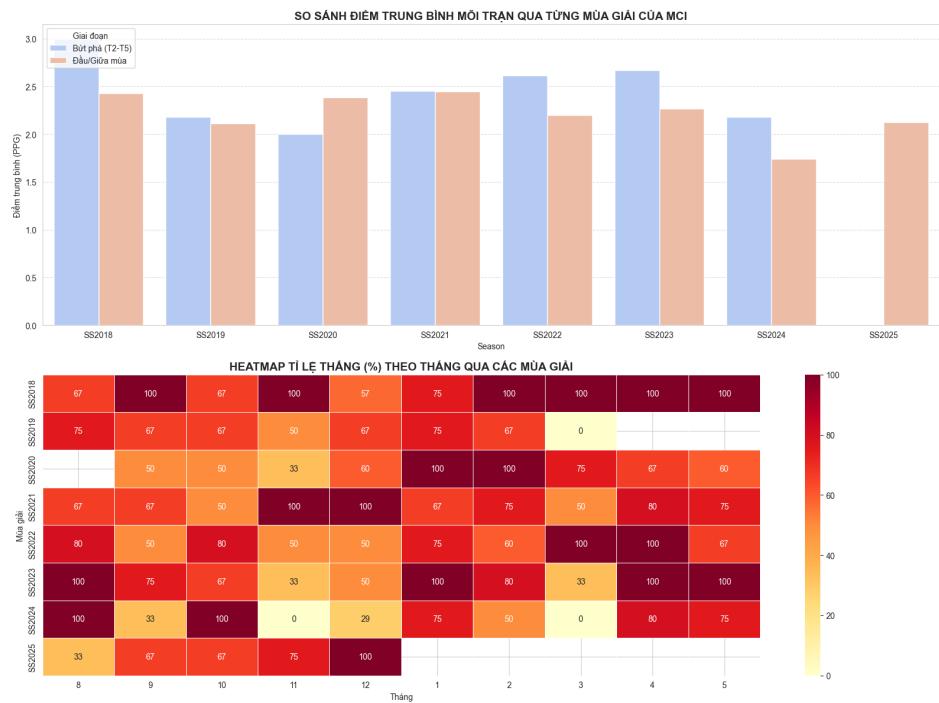
Dựa trên chỉ số Tạo cơ hội (SCA) trung bình mỗi trận, ta có các nhận xét sau về nhóm cầu thủ nhạy cảm với địa điểm thi đấu:

- Sự sụt giảm hiệu suất nghiêm trọng: Biểu đồ cho thấy một khoảng cách lớn giữa phong độ sân nhà (cột xanh) và sân khách (cột đỏ). Đặc biệt, **Luciano Vietto** là trường hợp tiêu biểu nhất khi chỉ số SCA giảm gần 5 lần khi phải thi đấu xa nhà (từ 4.67 xuống 0.91).
- Nhóm cầu thủ sáng tạo bị hạn chế: Các cầu thủ có thiên hướng kỹ thuật và bùng nổ như **Florian Wirtz, Jeremy Doku** và **Jean Seri** đều mất đi khoảng 40-50% khả năng gây đột biến khi đá sân khách. Điều này cho thấy lối chơi của họ bị ảnh hưởng mạnh bởi không gian thi đấu và áp lực từ khán giả đối phương.
- Đặc điểm "Home Specialist": Danh sách này chỉ ra những cầu thủ đóng vai trò "ngôi sao sân nhà". Việc hiểu rõ sự biến động này giúp ban huấn luyện có những điều chỉnh chiến thuật hợp lý, hoặc cân nhắc xoay tua đội hình trong các chuyến làm khách để duy trì tính đột biến cho hàng công.

11) Tỷ lệ thắng của các đội bóng theo tháng



Hình 3.21: Tỷ lệ thắng của các đội bóng theo tháng



Hình 3.22: Tỷ lệ thắng của đội Manchester City

- Sự thống trị tuyệt đối: Trên bản đồ nhiệt tỉ lệ thắng của tất cả các

đội, MCI là đội bóng có sắc đỏ đậm nhất và ổn định nhất. Tỉ lệ thắng của họ hiếm khi rơi xuống dưới 60%, vượt trội hoàn toàn so với các đối thủ lớn như CHE, MUN hay thậm chí là ARS.

- **Điểm rơi phong độ:** Trong khi các đội bóng khác thường hụt hơi vào giai đoạn khốc liệt đầu năm mới (tháng 1, tháng 2), MCI lại bắt đầu chu kỳ tăng tốc mạnh mẽ với các chỉ số thường xuyên đạt mức trên 80%.
- **Chỉ số PPG vượt trội:** Biểu đồ so sánh điểm trung bình mỗi trận (PPG) cho thấy một quy luật lặp đi lặp lại: Giai đoạn bứt phá (tháng 2 - tháng 5) luôn có hiệu suất cao hơn giai đoạn đầu mùa. Điều này khẳng định khả năng tính toán điểm rơi phong độ cực tốt của ban huấn luyện.
- **Sức mạnh hủy diệt ở giai đoạn "Run-in":**
 - Heatmap theo từng mùa giải cho thấy vào các tháng quyết định (tháng 4 và tháng 5), MCI thường xuyên đạt tỉ lệ thắng tuyệt đối **100%**.
 - Diễn hình như mùa giải SS2018 và SS2023, đội bóng đã quét sạch mọi đối thủ trong 2 tháng cuối để lên ngôi vô địch.
- **Kết luận:** MCI không chỉ là một đội bóng mạnh, mà còn là một "cỗ máy về đích". Tính mùa vụ trong lối chơi giúp họ luôn làm chủ cuộc đua đường dài, biến áp lực ở những vòng đấu cuối thành lợi thế nhờ sự ổn định và bản lĩnh kinh ngạc.

3.5 Kiến trúc kho dữ liệu

Sau khi khai phá dữ liệu, ta lựa chọn được kiến trúc phù hợp với kho dữ liệu sẽ được thiết kế như sau:

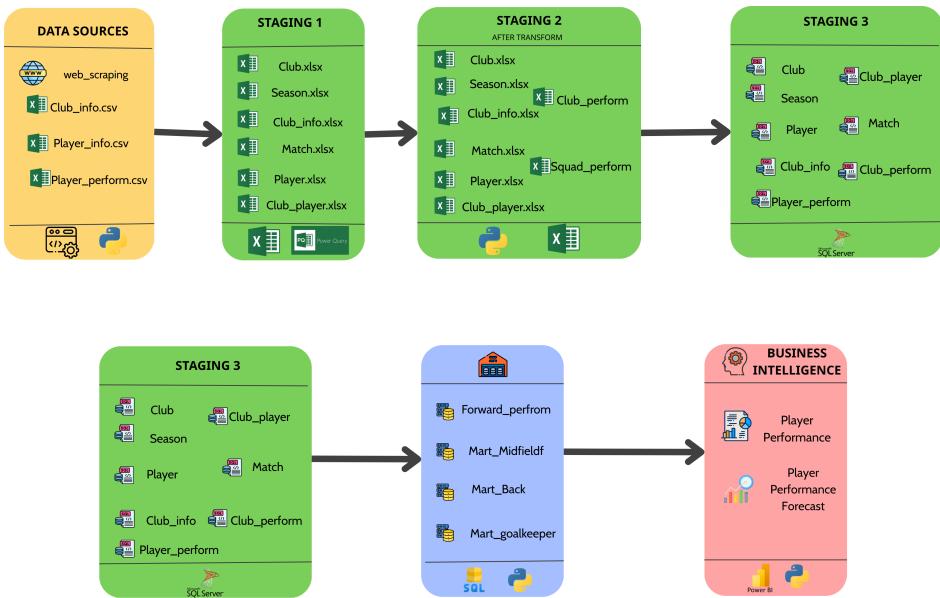


Hình 3.23: Kiến trúc kho dữ liệu

- **Data Sources:** Đây là nguồn cung cấp dữ liệu được sử dụng trong quá trình hoạt động, dữ liệu được cào từ trang web FbRef. Ta thu được các bảng gồm "club info", "player info", "player performance" như đã trình bày ở trên.
- **Data Staging:** Đây là vùng trung chuyển dữ liệu, là nơi mà dữ liệu được chuẩn hóa, làm sạch để được lưu trữ vào cơ sở dữ liệu, chuẩn bị thiết lập các bảng Dimension, fact trong Data Warehouse. Quá trình được thực hiện thông qua quy trình ETL, công nghệ thường được sử dụng là Python.
- **Data Warehouse:** Đây là nơi lưu trữ dữ liệu đã được làm sạch sau quá trình ETL, được tổ chức theo các chủ đề (bảng Fact) và các khía cạnh (Dimension) khác nhau. Cụ thể trong đề tài này có fact là "player_perform" và các dim là
- **Business Intelligence:** Đây là lớp cuối cùng trong kiến trúc Data Warehouse, nơi dữ liệu được trình bày cho người dùng cuối là các đối tác thông qua các công cụ BI và phân tích. Cụ thể là phân tích hiệu suất của các vị trí như tiền đạo, tiền vệ, hậu vệ và thủ môn.

3.6 Quy trình ETL

3.6.1 Data Pipeline



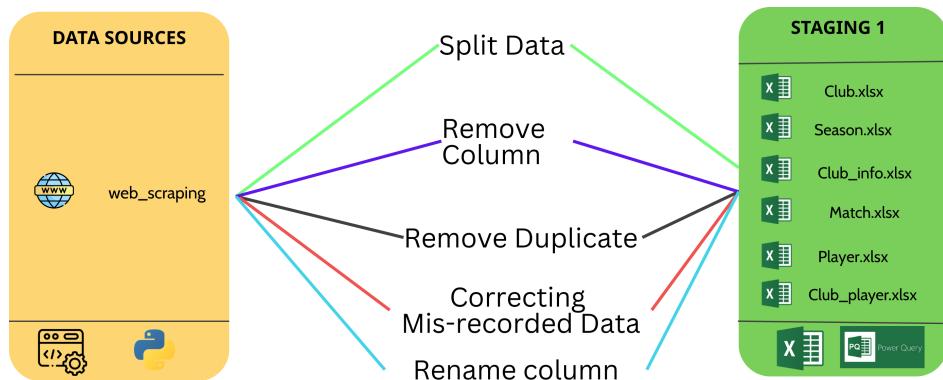
Hình 3.24: Data Pipeline trong quy trình ETL

Quy trình xử lý dữ liệu được thể hiện rõ ràng trong Data Pipeline ở trên, các công đoạn cụ thể như sau:

- **Data Source:** Nguồn dữ liệu được cào từ Web với các file csv "club_info", "player_info", "player_perform".
- **Staging 1:** Dữ liệu từ Data Source sẽ được trích xuất, bóc tách thành các bảng độc lập, chuẩn bị tiền để để thực hiện các thao tác phức tạp hơn.
- **Staging 2:** Ở đây dữ liệu tiếp tục được tách thành các bảng phức tạp hơn, gồm các liên kết với các bảng trước đó đã chuẩn bị tại Staging 1. Dữ liệu được làm sạch để chuẩn bị cho việc tải vào Data Warehouse.
- **Staging 3:** Dữ liệu sau khi được làm sạch, được tải vào hệ quản trị cơ sở dữ liệu SQL Server, dữ liệu được lưu trữ trong các bảng trong cơ sở dữ liệu.
- **Data Warehouse:** Dữ liệu được lưu trữ trong cơ sở dữ liệu tiếp tục được phân chia vào các bảng fact và dimension để tiện cho việc phân tích theo từng chủ đề cụ thể.

- **BI và analysis:** Dữ liệu từ Data Warehouse được sử dụng để tạo các báo cáo trực quan và biểu đồ phân tích. Các công cụ như Power BI được sử dụng để dễ dàng trực quan hóa dữ liệu và cung cấp thông tin hỗ trợ ra quyết định.

3.6.2 Quy trình ETL

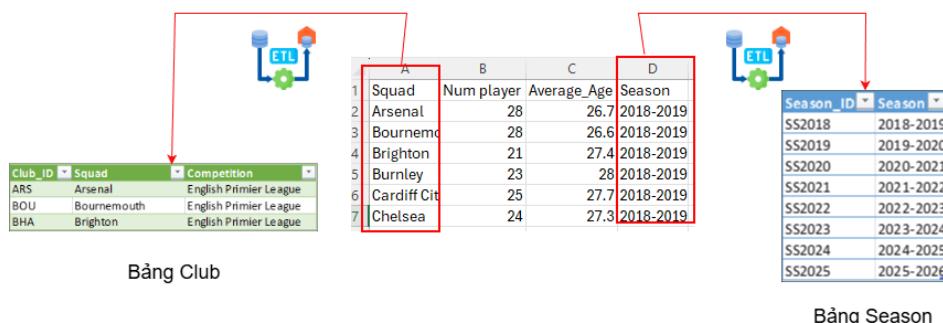


Hình 3.25: Quy trình ETL

Bây giờ ta sẽ tiến hành xử lý dữ liệu bằng cách phân chia các bảng gốc, xóa cột thừa, xóa dữ liệu trùng lặp.

1) Bảng Season và Club

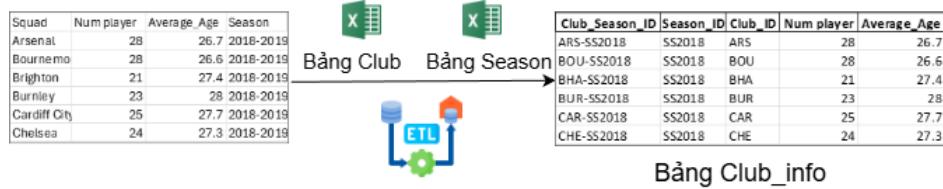
Tiến hành lọc lấy 2 cột Season và Club trong bảng "club_info", tạo ra 2 bảng mới để đánh ID theo cho mỗi mùa và cho mỗi câu lạc bộ. Ở bảng câu lạc bộ thêm một cột là "Competition", mô tả giải đấu các câu lạc bộ đó đang thi đấu.



Hình 3.26: Tiền xử lý bảng Club và bảng Season

2) Bảng club_info

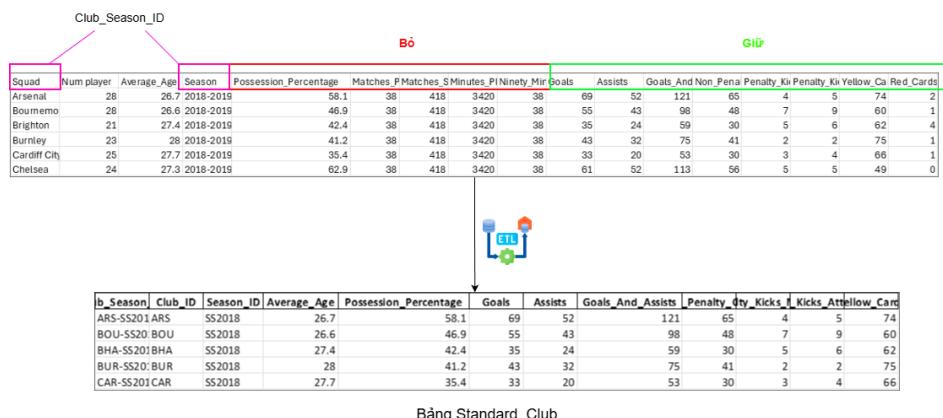
Các thông số của câu lạc bộ sẽ thay đổi theo từng mùa giải, cho nên ta tiến hành lấy các cột "Club", "Season", "Avg_age", "Num_player" trong bảng "club_info", sau đó ánh xạ qua hai bảng Club và Season đã tạo ở trên để lấy được các ID tương ứng.



Hình 3.27: Tiền xử lý bảng Club_info

3) Các bảng hiệu suất của các câu lạc bộ

Ta tiến hành xử lý các bảng hiệu suất (gồm 7 bảng lần lượt là: "standard", "goal and shot", "defend", "passing", "passtype", "possession" và bảng "goalkeeper"). Dùng 2 cột "Club" và "Season" để lấy được Club_info_ID từ bảng Club_info vừa tạo trước đó. Xóa các cột không cần thiết cho việc phân tích, giữ lại những cột hiệu suất của các đội qua mỗi mùa.



Hình 3.28: Tiền xử lý bảng club_perform

4) Bảng Player

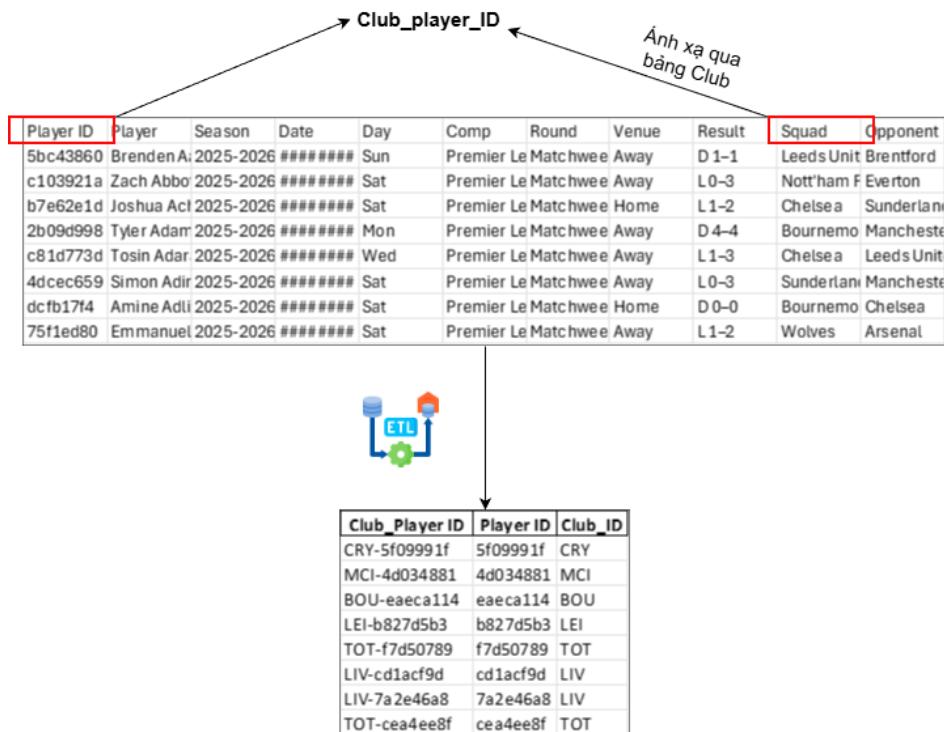
Với bảng player, ta đã cài đặt việc cào dữ liệu sạch sẽ, bao gồm các cột như :"Player_ID", "Player_name", "Born_date", "National", "Footed", "Height", "Weight",... Chúng ra chỉ cần định dạng lại kiểu dữ liệu cho cột "Born_date" là dữ liệu đã sạch sẽ.

Player ID	Name	Born Date	National Team	Footed	Height (cm)	Weight (kg)
77128f7a	Luke Amos	1997-02-23	England	Both	177	68
9f1893b5	Felipe Anderson	1993-04-15	Brazil	Right	175	69
dc008610	Florin Andone	1993-04-11	Romania	Right	182	78
ac05f970	Michail Antonio	1990-03-28	Jamaica	Right	180	82
97333cf5	Stuart Armstrong	1992-03-30	Scotland	Right	183	68
00459419	Marko Arnautović	1989-04-19	Austria	Right	192	83
28d596a0	Kepa Arrizabalaga	1994-10-03	Spain	Right	189	87

Hình 3.29: Tiền xử lý bảng Player

5) Club _ Player

Với đặc thù mỗi cầu thủ có thể thi đấu cho nhiều câu lạc bộ khác nhau, ta phải dựa vào lịch sử trận đấu của các cầu thủ để có thể tạo ra bảng "club _ Player". Lấy từ bảng "Match _ log" 2 cột là "Player _ ID" và "Squad", ánh xạ sang bảng "Club" để lấy được "Club _ ID". Sau đó xóa những dữ liệu trùng lặp, tạo ID cho mỗi cầu thủ nếu thi đấu ở các câu lạc bộ khác nhau bằng cách: "club _ ID" + " - " + "Player _ ID".

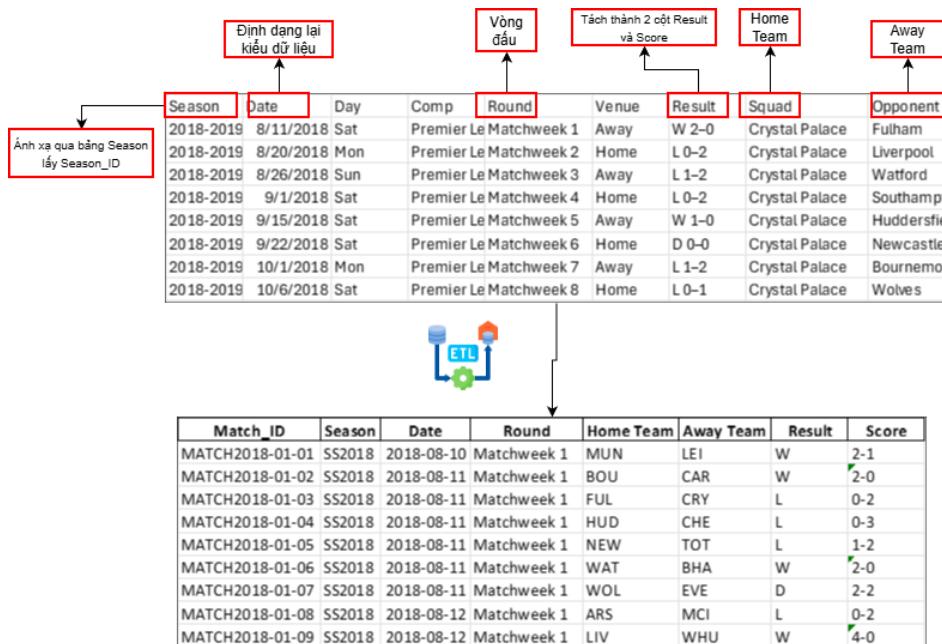


Hình 3.30: Tiền xử lý bảng Club _ player

6) Bảng Matches

Từ dữ liệu lịch sử trận đấu của các cầu thủ, ta tạo thêm một bảng dữ liệu chứa thông tin các trận đấu trong giải đấu. Các cột trong bảng được xử lý như sau:

- Lấy cột "Season", ánh xạ qua bảng "Season" để lấy "Season_ID".
- Định dạng lại dữ liệu thành Date ở cột "Date".
- Tạo cột "Home_Team" và "Away_Team" bằng cách: Nếu ở cột "Venue" có giá trị là "Home" thì "Home_Team" là giá trị nằm ở cột "Squad", còn nếu giá trị là "Away" thì "Home_Team" là giá trị nằm ở cột "Opponent", Sau đó ánh xạ qua bảng "CLub" để lấy được "Club_ID".
- Tách cột "Result" thành hai cột là "Result" và "Score", kết quả của hai cột này đều được ghi giá trị theo đội chủ nhà, tức là nếu cột "Venue" có giá trị là "Home" thì tỉ số sẽ giữ nguyên, còn nếu giá trị là "Away" thì sẽ đảo lại kết quả và tỉ số.
- Sắp xếp lại những trận có cùng vòng đấu và cùng mùa đấu.
- Cuối cùng là việc đánh ID cho mỗi trận đấu, đầu tiên là xóa dữ liệu trùng lặp, sau đó đánh ID theo nguyên tắc "MATCH" + "Season" + "Roundweek" + "số thứ tự", số thứ tự ở đây được đánh từ 1->10 do mỗi vòng đấu sẽ có 10 trận đấu.



Hình 3.31: Tiền xử lý bảng Matches

7) Bảng Match_Club_Player

Với việc mỗi cầu thủ thi đấu nhiều trận đấu, mỗi trận đấu cũng có

nhiều trận đấu, ta tạo thêm một bảng trung gian là bảng "Bảng Match_Club_Player" trách việc xảy ra mối quan hệ nhiều-nhiều giữa các bảng "Club_Player" - "Matches" và các bảng "Player_perform". Các thông tin trong bảng này được xử lý từ bảng "Player_perform", xử lý các trận đấu như bảng "Matches", ánh xạ qua bảng "Club_player" để lấy "Club_player_ID", sau đó tạo ID cho mỗi trận của mỗi cầu thủ, và thêm một số thông tin như "Venue", "Start", "Pos", "Min".

Club_player_matchlog_ID	Match_ID	Club_Player_ID	Venue	Start	Pos	Min
MATCH2018-01-03CRY-5f09991f	MATCH2018-01-03	CRY-5f09991f	Away	Y	LB	89
MATCH2018-02-10CRY-5f09991f	MATCH2018-02-10	CRY-5f09991f	Home	Y	LB	90
MATCH2018-03-09CRY-5f09991f	MATCH2018-03-09	CRY-5f09991f	Away	Y	LB	90
MATCH2018-04-03CRY-5f09991f	MATCH2018-04-03	CRY-5f09991f	Home	Y	LB	90
MATCH2018-05-03CRY-5f09991f	MATCH2018-05-03	CRY-5f09991f	Away	Y	LB	90
MATCH2018-06-04CRY-5f09991f	MATCH2018-06-04	CRY-5f09991f	Home	Y	LB	90
MATCH2018-07-10CRY-5f09991f	MATCH2018-07-10	CRY-5f09991f	Away	Y	LB	90

Hình 3.32: Tiền xử lý bảng Matches_Club_player

8) Các bảng chỉ số của cầu thủ Player_perform

Ta tiến hành xử lý các bảng hiệu suất của các cầu thủ (gồm 7 bảng lần lượt là: "standard", "goal and shot", "defend", "passing", "passtype", "possession" và bảng "goalkeeper"), lần lượt xử lý logic để lấy các ID như các bảng trước, ánh xạ qua bảng "Matches_Club_player" để lấy ID cuối cùng, sau đó chỉ giữ lại cột ID và các cột hiệu suất. Ta thu được các bảng "Player_performance".

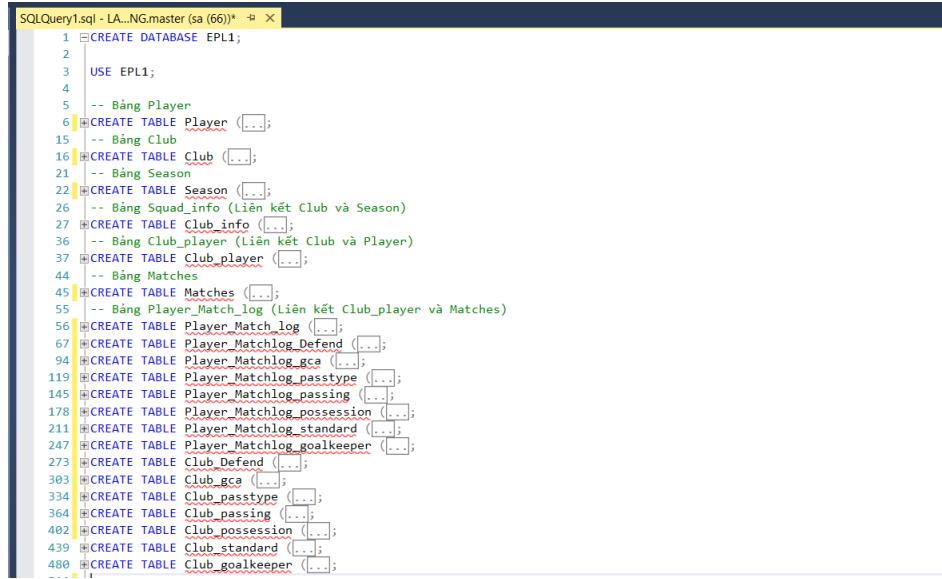
Club_player_matchlog_ID	Gls	Ast	PK	PKatt	Sh	SoT	CrdY	CrdR	Touches
MATCH2018-01-03CRY-5f09991f	0	1	0	0	1	1	0	0	57
MATCH2018-02-10CRY-5f09991f	0	0	0	0	1	0	1	0	61
MATCH2018-03-09CRY-5f09991f	0	0	0	0	1	0	0	0	62
MATCH2018-04-03CRY-5f09991f	0	0	0	0	3	1	0	0	63
MATCH2018-05-03CRY-5f09991f	0	0	0	0	0	0	0	0	56
MATCH2018-06-04CRY-5f09991f	0	0	0	0	2	0	0	0	77
MATCH2018-07-10CRY-5f09991f	1	0	0	0	1	1	0	0	77
MATCH2018-08-03CRY-5f09991f	0	0	0	0	1	0	1	0	110

Hình 3.33: Tiền xử lý bảng Player_performance

3.6.3 Tạo cơ sở dữ liệu và lưu trữ dữ liệu

Sau quá trình tiền xử lý dữ liệu, ta tiến hành đổ dữ liệu từ các file vào các bảng trong hệ quản trị cơ sở dữ liệu SQL Server, chuyển thành dữ liệu có cấu trúc, dễ dàng truy vấn và sử dụng trong các ứng dụng hoặc hệ thống phân tích dữ liệu. Các bước thực hiện việc import dữ liệu:

- Tạo cơ sở dữ liệu: Thực hiện tạo cơ sở dữ liệu trong hệ quản trị cơ sở dữ liệu SQL server với các bảng và các khóa ngoại tương ứng.



```

1 CREATE DATABASE EPL1;
2
3 USE EPL1;
4
5 -- Bảng Player
6 CREATE TABLE Player (...) ;
15 -- Bảng Club
16 CREATE TABLE Club (...) ;
21 -- Bảng Season
22 CREATE TABLE Season (...) ;
26 -- Bảng Squad_info (Liên kết Club và Season)
27 CREATE TABLE Club_info (...) ;
36 -- Bảng Club_player (Liên kết Club và Player)
37 CREATE TABLE Club_player (...) ;
44 -- Bảng Matches
45 CREATE TABLE Matches (...) ;
55 -- Bảng Player_Match_log (Liên kết Club_player và Matches)
56 CREATE TABLE Player_Match_log (...) ;
67 CREATE TABLE Player_Matchlog_Defend (...) ;
94 CREATE TABLE Player_Matchlog_gca (...) ;
119 CREATE TABLE Player_Matchlog_passtype (...) ;
145 CREATE TABLE Player_Matchlog_passing (...) ;
178 CREATE TABLE Player_Matchlog_possession (...) ;
211 CREATE TABLE Player_Matchlog_standard (...) ;
247 CREATE TABLE Player_Matchlog_goalkeeper (...) ;
273 CREATE TABLE Club_Defend (...) ;
303 CREATE TABLE Club_gca (...) ;
334 CREATE TABLE Club_passtype (...) ;
364 CREATE TABLE Club_passing (...) ;
402 CREATE TABLE Club_possession (...) ;
439 CREATE TABLE Club_standard (...) ;
480 CREATE TABLE Club_goalkeeper (...) ;
...

```

Hình 3.34: Tạo cơ sở dữ liệu

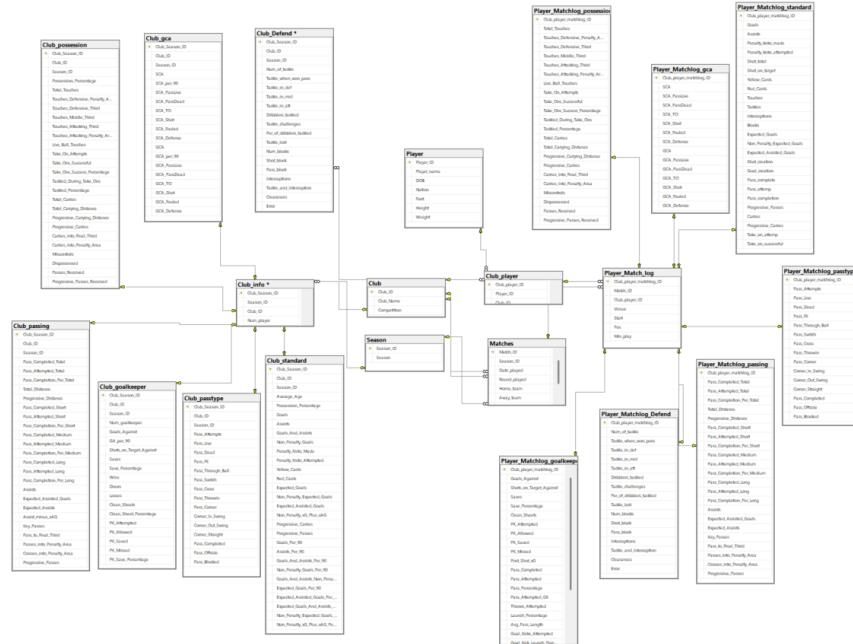
- Đọc dữ liệu từ file: Đọc dữ liệu từ các file đã xử lý vào các dataframe bằng thư viện pandas, sẵn sàng cho việc import
- Kiểm tra cấu trúc dữ liệu: Xác minh số lượng cột trong dữ liệu, tiến hành đổi tên để khớp với các cột trong cơ sở dữ liệu, nhằm đảm bảo dữ liệu được import đúng và đủ
- Kết nối với cơ sở dữ liệu và tải dữ liệu vào

```
1
2 # 1. Thông tin kết nối từ hình ảnh bạn cung cấp
3 server = 'LAP-CUA-HOANGGG\DONGHOANG'
4 database = 'EPL1'
5 username = 'sa'
6 password = 'donghoang1610' # <-- Thay mật khẩu của bạn vào đây
7 driver = '{ODBC Driver 17 for SQL Server}' # Hoặc Driver 13, 18 tùy máy bạn
8
9 # 2. Tạo kết nối tới SQL Server
10 params = urllib.parse.quote_plus(
11     f'DRIVER={driver};'
12     f'SERVER={server};'
13     f'DATABASE={database};'
14     f'UID={username};'
15     f'PWD={password};'
16 )
17 engine = create_engine(f"mssql+pyodbc:///?odbc_connect={params}")
18
```

Hình 3.35: Kết nối với cơ sở dữ liệu

- Xử lý ngoại lệ: Đảm bảo chương trình không bị dừng khi xảy ra lỗi, thay vào đó là hiển thị thông báo để người dùng biết và khắc phục.

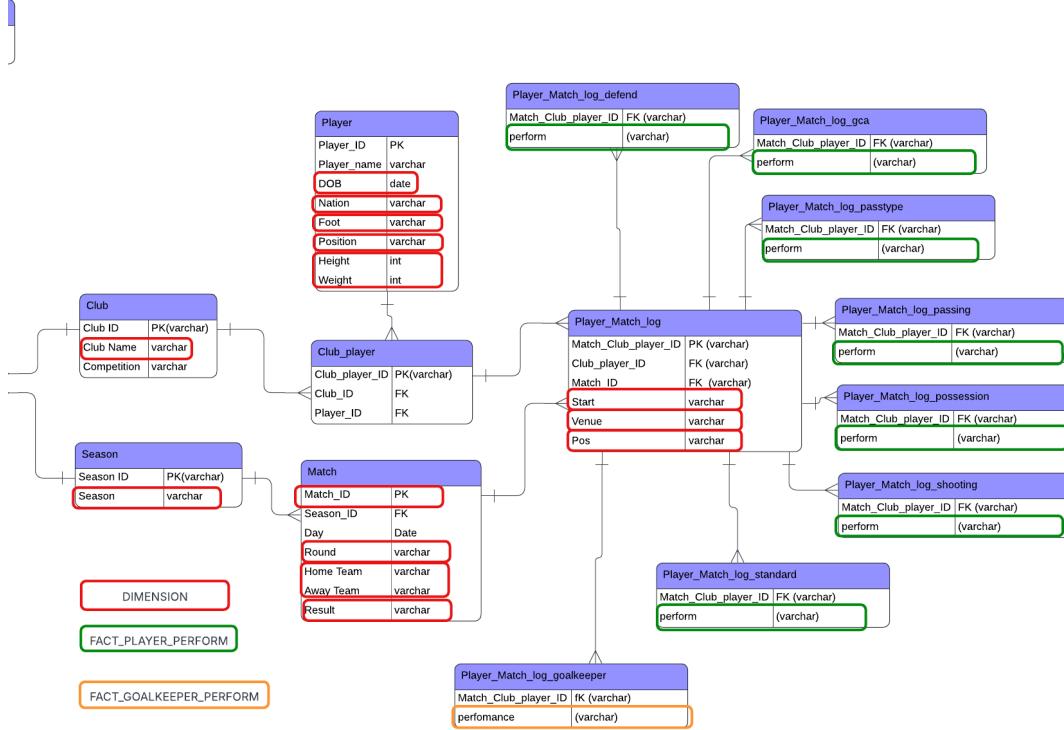
Sau quá trình tiền xử lý dữ liệu, ta thu được mô hình dữ liệu OLTP như sau:



Hình 3.36: Mô hình dữ liệu OLTP

3.6.4 ETL dữ liệu qua Data Warehouse

Sau khi dữ liệu đã được đổ vào database, ta thực hiện các câu truy vấn ở trong database để tạo ra các bảng fact và dim trong Data Warehouse

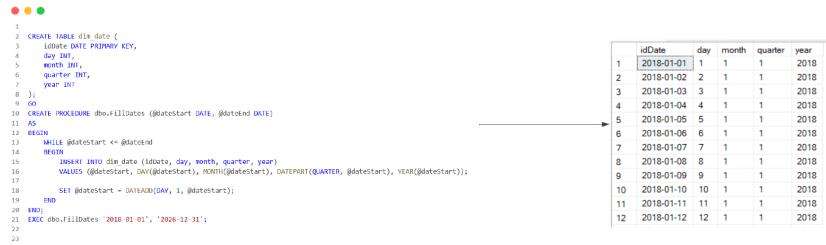


Hình 3.37: Quá trình mapping các cột để tạo bảng dimension và bảng fact

- Fact_Player_performance: Gồm các trường miêu tả chỉ số ở các bảng "Player_standard", "Player_defend", "Player_gca", "Player_passtype", "Player_passing", "Player_possession"
- Fact_Goalkeeper_performance: Gồm các trường miêu tả chỉ số ở bảng "Player_goalkeeper"
- Dim_Nation: bao gồm trường nation lấy từ bảng "Player"
- Dim_dominated_foot: bao gồm trường foot lấy từ bảng "Player"
- Dim_Height: bao gồm trường Height lấy từ bảng "Player"
- Dim_Weight: bao gồm trường Weight lấy từ bảng "Player"
- Dim_Club: bao gồm trường Club_name lấy từ bảng "Club"

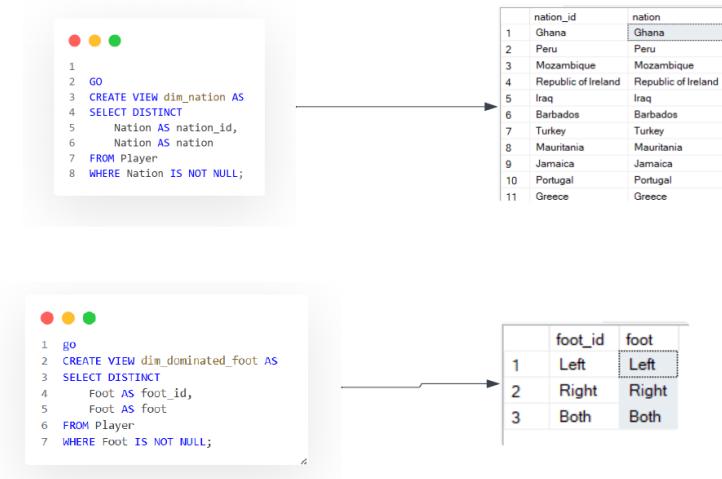
- Dim_Season: Bao gồm trường Season lấy từ bảng "Season"
- Dim_Match: Bao gồm trường Match lấy từ bảng "Matches"
- Dim_Round: Bao gồm trường Round lấy từ bảng "Matches"
- Dim_Result: Bao gồm trường Result lấy từ bảng "Matches"
- Dim_Start: Bao gồm trường Start lấy từ bảng "Player_Matchlog"
- Dim_Venue: Bao gồm trường Venue lấy từ bảng "Player_matchlog"
- Dim_position: bao gồm trường position lấy từ bảng "Player_matchlog"
- Dim_Date: được tạo bằng cách lấy dữ liệu từ năm 2018 đến năm 2026

Tạo procedure để đổ dữ liệu date, từ đó tạo ra bảng Dim_Date với thời gian từ năm 2018 đến năm 2026.

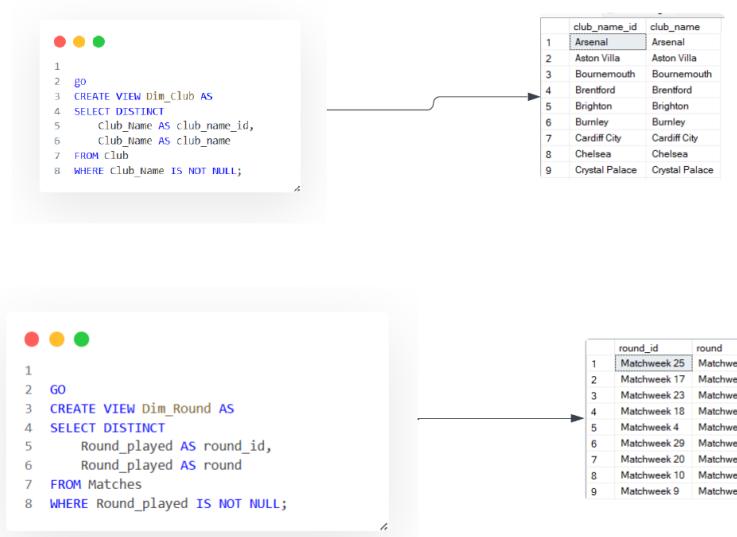


Hình 3.38: Bảng dim_date

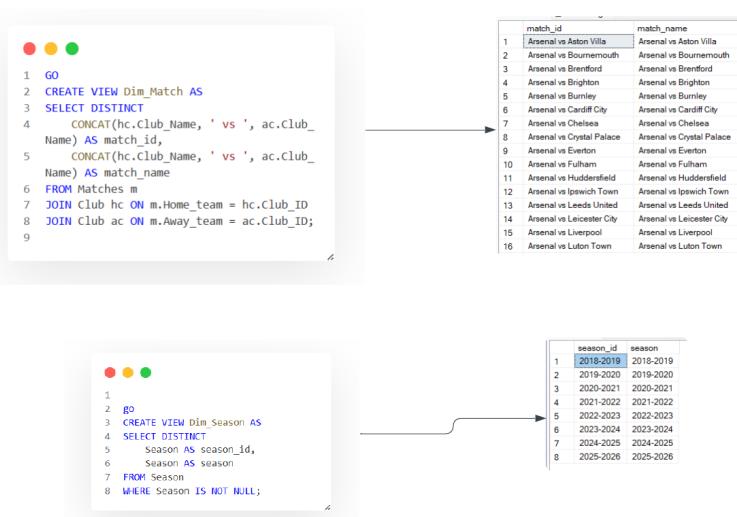
Để tạo ra các bảng dim, ta sử dụng câu lệnh SELECT DISTINCT từ các cột chứa các trường liên quan trong các bảng cơ sở dữ liệu



Hình 3.39: Ví dụ các bảng Dimesion

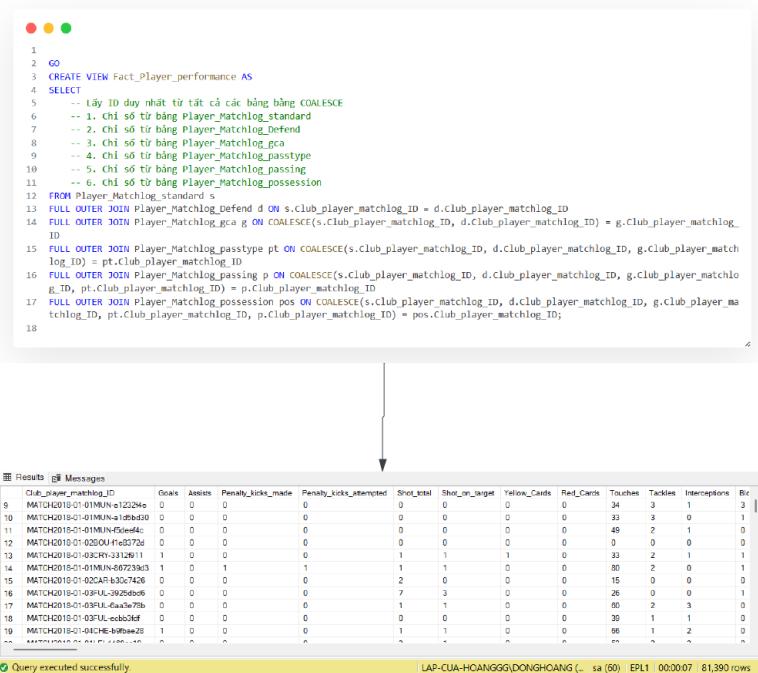


Hình 3.40: Ví dụ các bảng Dimension



Hình 3.41: Ví dụ các bảng Dimension

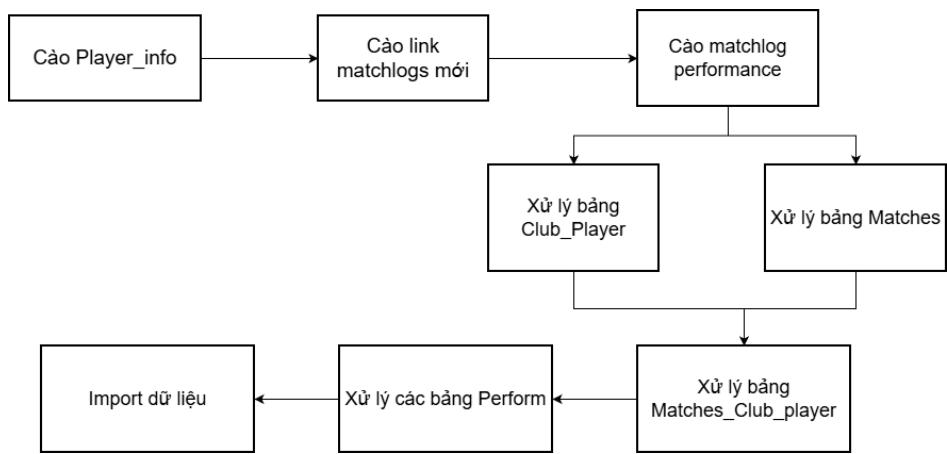
Sau đó ta tiến hành tạo bảng fact



Hình 3.42: Bảng fact Player Performance

3.7 Lập lịch quy trình tự động

Với việc dữ liệu của giải bóng đá được cập nhật liên tục hàng tuần trên web, ta sẽ tiến hành cài đặt quá trình cào, xử lý dữ liệu và tải dữ liệu vào cơ sở dữ liệu trở nên tự động vào mỗi tuần. Thực hiện đặt lệnh chạy các file python bằng **Window Task Scheduler**

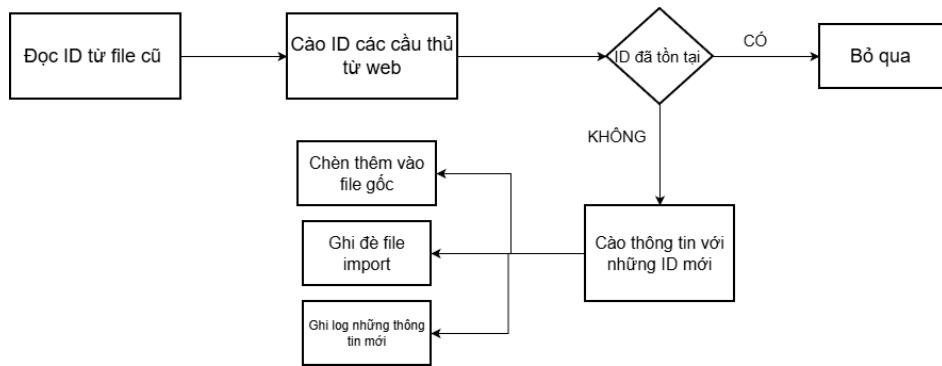


Hình 3.43: Lập lịch quy trình tự động

3.7.1 Quy trình ETL

1. Player Info

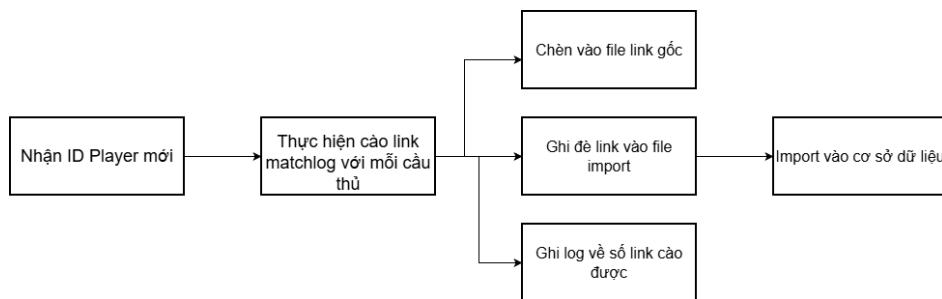
- **Bước 1: Thu thập dữ liệu đầu vào** - Đọc danh sách ID từ file cũ và thực hiện cào ID các cầu thủ từ website.
- **Bước 2: Kiểm tra trùng lặp** - So sánh ID mới cào được với ID đã tồn tại:
 - Nếu đã tồn tại: Bỏ qua.
 - Nếu chưa tồn tại: Chuyển sang bước xử lý tiếp theo.
- **Bước 3: Cào thông tin chi tiết** - Tiến hành cào dữ liệu chi tiết đối với những ID mới.
- **Bước 4: Lưu trữ và ghi nhật ký** - Thực hiện đồng thời các thao tác:
 - Chèn thêm dữ liệu mới vào file gốc.
 - Ghi đè thông tin vào file import.
 - Ghi log các thông tin mới vừa được cập nhật.



Hình 3.44: Xử lý thông tin cầu thủ mới

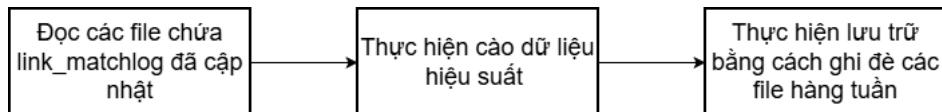
2. Cào link Matchlogs

- **Bước 1: Tiếp nhận đầu vào** - Nhận danh sách các ID Player mới cần xử lý.
- **Bước 2: Thu thập dữ liệu** - Thực hiện cào (scrape) các đường dẫn (link) matchlog tương ứng với mỗi cầu thủ.
- **Bước 3: Phân phối và lưu trữ** - Sau khi thu thập, hệ thống thực hiện đồng thời ba tác vụ:
 - Chèn các link vừa cào được vào file link gốc.
 - Ghi đè danh sách link vào file import để chuẩn bị cập nhật.
 - Ghi log báo cáo về số lượng link đã thu thập được.
- **Bước 4: Cập nhật hệ thống** - Thực hiện import dữ liệu từ file import vào cơ sở dữ liệu (database).



Hình 3.45: Xử lý cào link matchlog mới

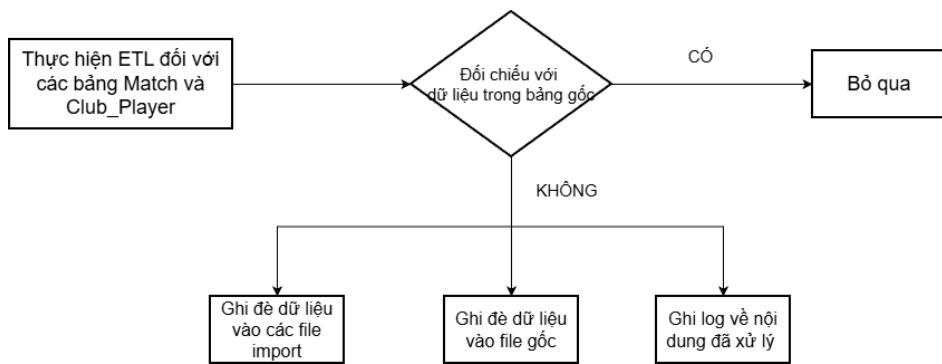
3. Cào dữ liệu hiệu suất



Hình 3.46: Xử lý cào dữ liệu hiệu suất mới

4. Xử lý dữ liệu bảng Matches và bảng Club_Player

- **Bước 1: Thực hiện ETL** - Tiến hành quá trình trích xuất, chuyển đổi và nạp dữ liệu (ETL) đối với các bảng *Match* và *Club_Player*.
- **Bước 2: Đổi chiều dữ liệu** - So sánh dữ liệu sau khi ETL với dữ liệu hiện có trong bảng gốc:
 - **Trường hợp đã tồn tại (CÓ)**: Hệ thống thực hiện lệnh *Bỏ qua*.
 - **Trường hợp chưa tồn tại (KHÔNG)**: Chuyển sang bước cập nhật dữ liệu mới.
- **Bước 3: Lưu trữ và ghi nhật ký** - Thực hiện đồng thời ba thao tác đối với dữ liệu mới:
 - Ghi đè dữ liệu vào các file import.
 - Ghi đè dữ liệu vào file gốc để cập nhật hệ thống.
 - Ghi log chi tiết về các nội dung đã được xử lý.

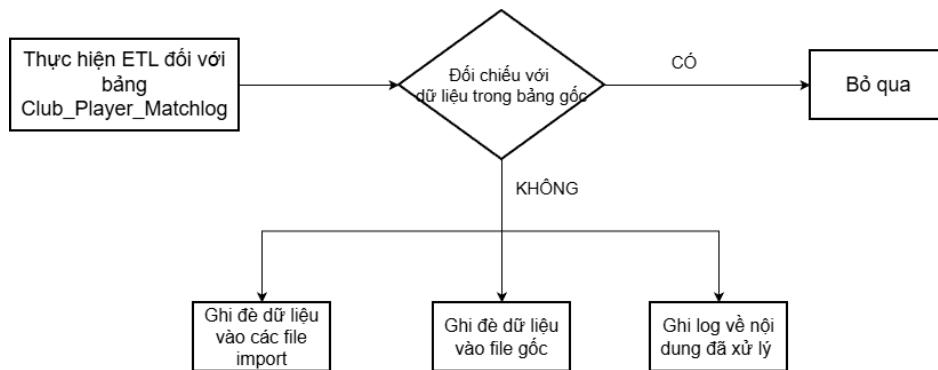


Hình 3.47: Xử lý dữ liệu các bảng Matches và Club_Player

5. Xử lý dữ liệu bảng Club_Player_Matchlogs

- **Bước 1: Thực hiện ETL** - Tiến hành quá trình trích xuất, chuyển đổi và nạp dữ liệu (ETL) đối với các bảng *perform*.

- **Bước 2: Đổi chiều dữ liệu** - So sánh dữ liệu vừa xử lý với dữ liệu hiện có trong bảng gốc:
 - **Trường hợp đã tồn tại (CÓ):** Hệ thống thực hiện lệnh *Bỏ qua*.
 - **Trường hợp chưa tồn tại (KHÔNG):** Chuyển sang bước cập nhật dữ liệu mới.
- **Bước 3: Lưu trữ và ghi nhật ký** - Thực hiện đồng thời ba thao tác đối với dữ liệu mới:
 - Ghi đè dữ liệu vào các file import.
 - Ghi đè dữ liệu vào file gốc.
 - Ghi log chi tiết về các nội dung đã được xử lý.

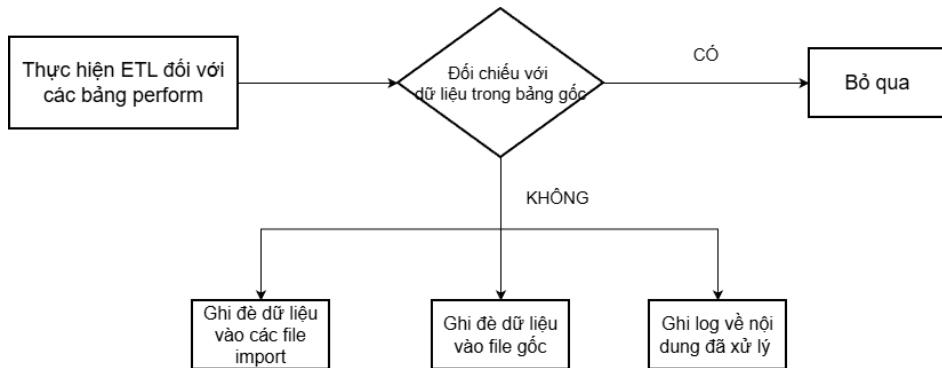


Hình 3.48: Xử lý dữ liệu bảng Club_Player_Matchlogs

6. Xử lý dữ liệu các bảng hiệu suất

- **Bước 1: Thực hiện ETL** - Tiến hành quá trình trích xuất, chuyển đổi và nạp dữ liệu (ETL) đối với các bảng *perform*.
- **Bước 2: Đổi chiều dữ liệu** - So sánh dữ liệu vừa xử lý với dữ liệu hiện có trong bảng gốc:
 - **Trường hợp đã tồn tại (CÓ):** Hệ thống thực hiện lệnh *Bỏ qua*.
 - **Trường hợp chưa tồn tại (KHÔNG):** Chuyển sang bước cập nhật dữ liệu mới.
- **Bước 3: Lưu trữ và ghi nhật ký** - Thực hiện đồng thời ba thao tác đối với dữ liệu mới:

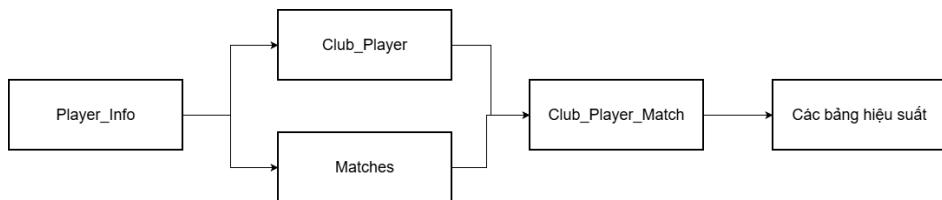
- Ghi đè dữ liệu vào các file import.
- Ghi đè dữ liệu vào file gốc.
- Ghi log chi tiết về các nội dung đã được xử lý.



Hình 3.49: Xử lý dữ liệu các bảng hiệu suất

3.7.2 Thực hiện đẩy thêm dữ liệu mới vào cơ sở dữ liệu

Sau khi đã có các bảng dữ liệu đã sẵn sàng để đẩy vào cơ sở dữ liệu, ta tiến hành import dữ liệu theo thứ tự để có thể đảm bảo các khóa ngoại hoạt động tốt nhất



Hình 3.50: Import dữ liệu vào cơ sở dữ liệu

3.7.3 Lên lịch tự động với Window Task Scheduler

Sau khi có các file python chạy hàng tuần, ta tiến hành điều phối thứ tự các file bằng 1 file khác, cho chạy lần lượt các file với yêu cầu xong file này rồi mới chạy file khác.



```

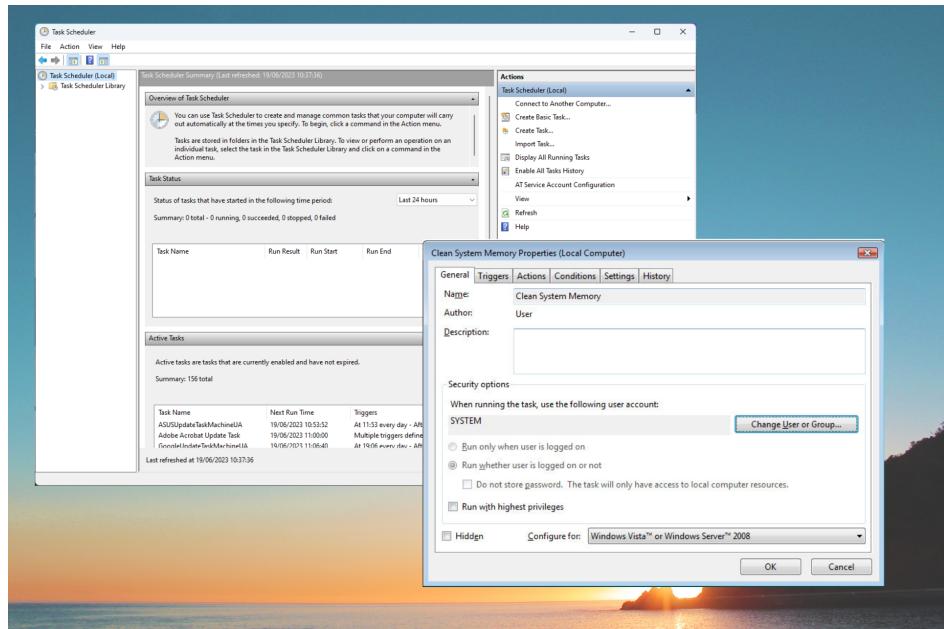
1
2 # Danh sách 14 file theo đúng thứ tự bạn cung cấp
3 scripts = [
4     r"E:\project 1\script_tuan_moi\player_info.py",
5     r"E:\project 1\script_tuan_moi\cao_link.py",
6     r"E:\project 1\script_tuan_moi\cao_gk.py",
7     r"E:\project 1\script_tuan_moi\cao_player_matchlog.py",
8     r"E:\project 1\script_tuan_moi\cao_goalkeeper_matchlog.py",
9     r"E:\project 1\script_tuan_moi\Club_player.py",
10    r"E:\project 1\script_tuan_moi\match.py",
11    r"E:\project 1\script_tuan_moi\Club_player_matchlog.py",
12    r"E:\project 1\script_tuan_moi\xuly_match_log.py",
13    r"E:\project 1\script_tuan_moi\import_player.py",
14    r"E:\project 1\script_tuan_moi\import_Club_player.py",
15    r"E:\project 1\script_tuan_moi\import_match.py",
16    r"E:\project 1\script_tuan_moi\import_club_player_matchlog.py",
17    r"E:\project 1\script_tuan_moi\import_performance.py"
18 ]
19

```

Hình 3.51: Điều phối thứ tự xử lý

Cuối cùng với file điều phối, tiến hành cài đặt Task trong Window Task Scheduler để chạy hàng tuần với:

- Thời gian: 11PM every Webnesday
- Program/script: "C:\Users\HP\myenv\Scripts\python.exe"
- Add arguments: E:\project 1\script_tuan_moi\run_all.py



Hình 3.52: Window Task Scheduler

Chương 4

Xây dựng dashboard

4.1 Yêu cầu phân tích

Trong kỷ nguyên “Big Data” của thể thao hiện đại, dữ liệu không chỉ là những con số thống kê khô khan mà là tài sản vô giá giúp các câu lạc bộ tuyển trạch, huấn luyện viên xây dựng chiến thuật và người hâm mộ đánh giá cầu thủ. Giải Ngoại hạng Anh (English Premier League - EPL) là giải đấu hấp dẫn nhất hành tinh với khối lượng dữ liệu khổng lồ sinh ra sau mỗi vòng đấu.

Từ đó, nhiệm vụ của công ty là tiến hành phân tích hiệu suất cầu thủ dựa trên các thông số thu thập được, đánh giá các vị trí theo khu vực thi đấu như Tiền đạo, Tiền vệ, Hậu vệ và Thủ môn. Từ đó đưa ra được các sơ đồ đội hình hợp lý, các chiến thuật cũng như cách sắp xếp đối với mỗi đội thủ

4.2 Hệ thống Dimension

Sau quá trình xây dựng hệ thống ta thu thập được hệ thống các chiều khái niệm như sau:

MONTH	QUARTER	YEAR	DIM_Result	DIM_Venue
1	1	2018	Loss	Home
2	2	2019	Win	AWAY
...	3	2020	Draw	
11	4	2021		2 giá trị
12		2022		3 giá trị
4 giá trị		2023		
12 giá trị		2024		
9 giá trị		2025		
9 giá trị		2026		

Hình 4.1: Ví dụ các bảng Dimesion

DIM_Nation	DIM_dominated_foot	DIM_Height	DIM_Weight	DIM_Club	DIM_Season
Ghana	Left	160-165	50-55	Asenal	2018-2019
Peru	Right	165-170	55-60	Manchester City	2019-2020
Portugal	Both	Liverpool	2020-2021
England		195-200	90-95		2021-2022
Spain	3 giá trị	200-205	95-100	Brenford	2022-2023
.....		9 giá trị	10 giá trị	Astonvilla	2023-2024
97 giá trị				29 giá trị	2024-2025
					2025-2026

Hình 4.2: Ví dụ các bảng Dimension

DIM_Match	DIM_Round	DIM_Start	DIM_Position
Arsenal vs Aston Villa	Round 1	No	Goalkeeper
Bournemouth vs Chelsea	Round 2	Yes	Center Back
Chelsea vs Brighton	Round 3	Yes and captain	Forward
.....	...	3 giá trị	Center Midfield
Everton vs Liverpool	Round 37		Defend Midfield
Leicester City vs West Ham	Round 38		Left Winger
750 giá trị	38 giá trị		Right Winger
			Attack Midfield
			Right Back
			Left Back
			Left Midfield
			Right Midfield
			Wing Back
			13 giá trị

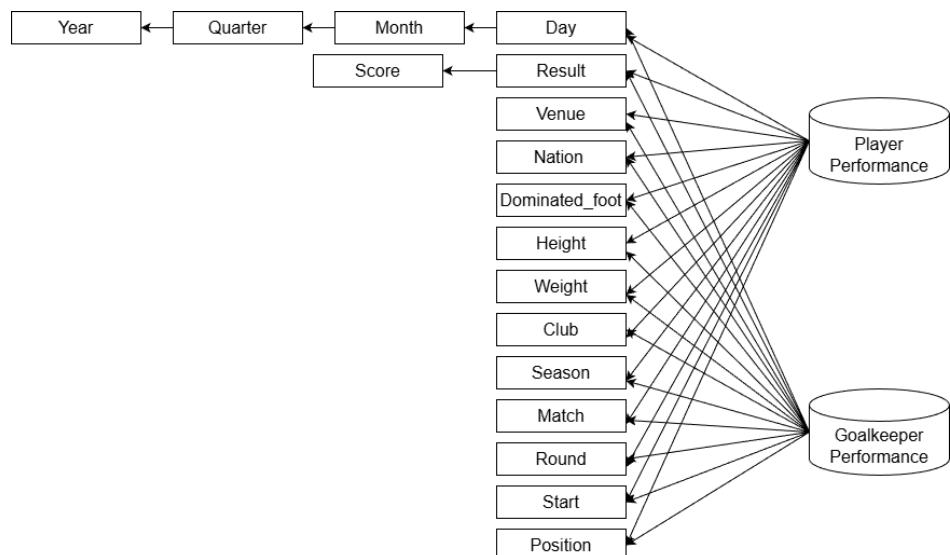
Hình 4.3: Ví dụ các bảng Dimension

- Dim_date: Chứa thông tin về khía cạnh thời gian.
 - Dim_Result: Chứa thông tin về khía cạnh kết quả trận đấu.

- Dim_Venue: Chứa thông tin về khía cạnh thi đấu sân nhà hay sân khách.
- Dim_Nation: Chứa thông tin về khía cạnh quốc tịch của cầu thủ.
- Dim_Dominated_foot: Chứa thông tin về khía cạnh chân thuận của cầu thủ.
- Dim_Height: Chứa thông tin về khía cạnh chiều cao của cầu thủ
- Dim_Weight: Chứa thông tin về khía cạnh cân nặng của cầu thủ.
- Dim_Club: Chứa thông tin về khía cạnh câu lạc bộ.
- Dim_Season: Chứa thông tin về khía cạnh mùa giải.
- Dim_Match: Chứa thông tin về khía cạnh trận đấu.
- Dim_Round: Chứa thông tin về khía cạnh vòng đấu.
- Dim_Start: Chứa thông tin về khía cạnh cầu thủ có xuất phát hay không.
- Dim_Position: Chứa thông tin về khía cạnh vị trí thi đấu.

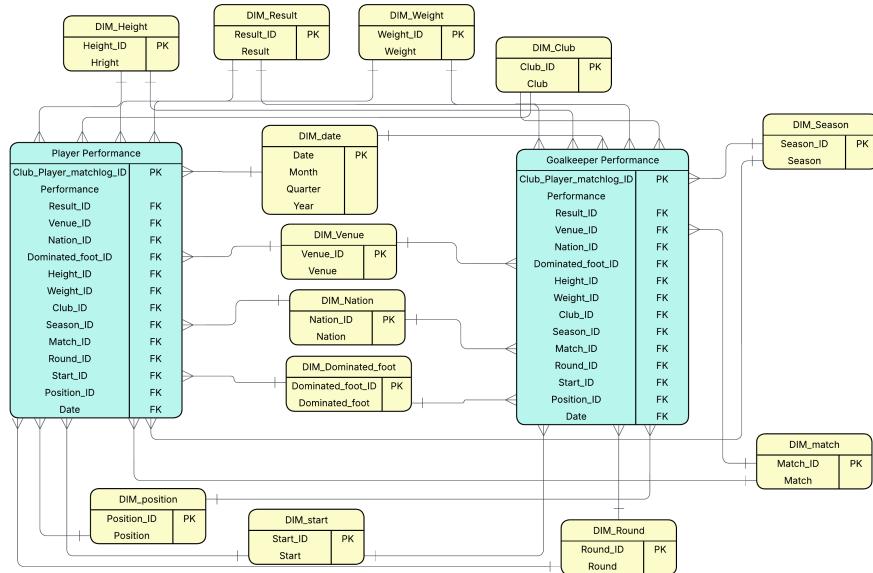
4.3 Data Model Logic

Từ các yêu cầu phân tích, ta tiến hành xây dựng Data model logic theo 2 chủ điểm chính là Player_Performance và Goalkeeper_Performance



Hình 4.4: Data Model Logic

Sau khi có được Data Model logic, ta đưa dữ liệu về mô hình vật lý để lưu trữ trong cơ sở dữ liệu và sử dụng để phân tích. Mô hình dữ liệu vật lý như sau:



Hình 4.5: Data Physical Model

4.4 Dashboard

4.4.1 Dashboard tổng quan về giải đấu



Hình 4.6: Dashboard tổng quan về giải đấu

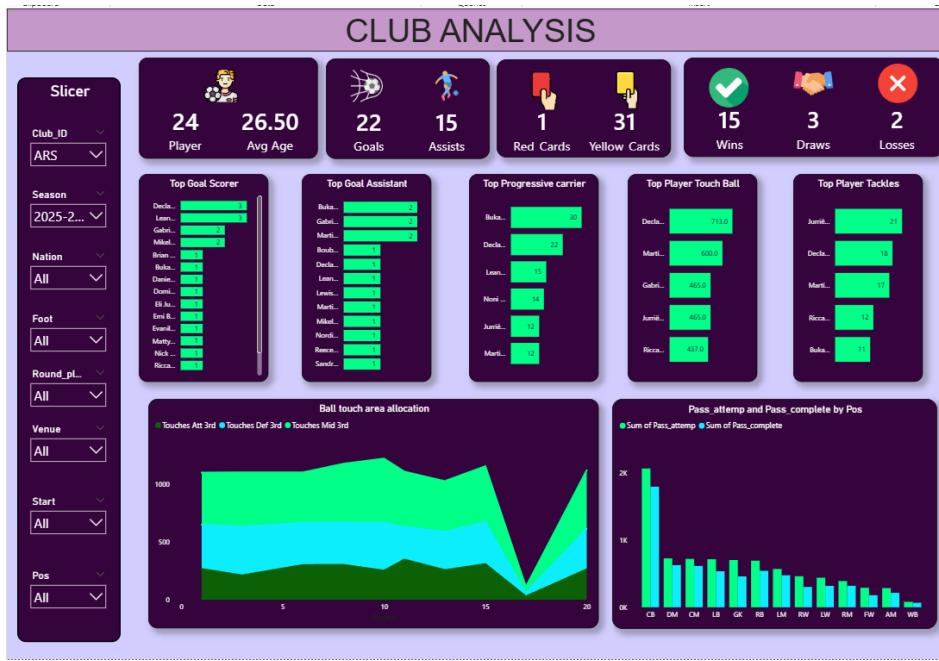
1. Đối tượng người dùng

- **Chuyên gia phân tích chiến thuật:** Sử dụng dữ liệu để đánh giá sự chuyển dịch sức mạnh giữa các câu lạc bộ qua các mùa giải.
- **Cơ quan truyền thông và báo chí:** Khai thác các con số thống kê nhanh như tổng bàn thắng (7,856) hay các cầu thủ dẫn đầu danh sách ghi bàn để làm tư liệu bài viết.
- **Người hâm mộ bóng đá:** Theo dõi bảng xếp hạng tổng thể và hiệu suất của đội bóng yêu thích thông qua bảng "Statistical results".

2. Mục đích

- **Tổng hợp chỉ số cốt lõi (KPIs):** Hiển thị quy mô giải đấu thông qua số lượng câu lạc bộ (30), cầu thủ (1,593) và tính chất quyết liệt của giải đấu qua số thẻ phạt.
- **Theo dõi biến động thứ hạng:** Biểu đồ đường *Ranking Over Season* cho phép so sánh trực quan sự ổn định của Manchester City (MCI) so với các đội thủ khác trong nhóm Big Six.
- **Dánh giá hiệu suất thi đấu:**
 - Tỷ lệ thắng sân nhà chiếm **44.23%**, cho thấy lợi thế sân bãi vẫn là yếu tố then chốt.
 - Bảng thống kê chi tiết cung cấp *Win Rate* và *Total Points* tích lũy, khẳng định vị thế dẫn đầu của MCI với 71.33% tỷ lệ thắng.
- **Phân tích phân phối bàn thắng:** Biểu đồ cột *Goal / Round Played* giúp xác định các vòng đấu có hiệu suất ghi bàn cao nhất, hỗ trợ việc dự báo xu hướng trận đấu.

4.4.2 Dashboard tổng quan cho câu lạc bộ



Hình 4.7: Dashboard tổng quan cho câu lạc bộ

1. Thông tin chung (Key Performance Indicators) Dựa trên các thẻ chỉ số (Cards) ở hàng đầu tiên của Dashboard:

- Nhân sự:** Đội hình đăng ký 24 cầu thủ, độ tuổi trung bình 26.50 cho thấy sự chín muồi về kinh nghiệm.
- Kết quả:** Đội bóng đạt phong độ cực cao với 15 trận thắng (Wins) và chỉ 2 trận thua (Losses).
- Kỷ luật:** Chỉ số thẻ phạt (1 Red, 31 Yellow) nằm trong mức kiểm soát tốt đối với một đội bóng chơi áp sát.

2. Phân tích đóng góp cá nhân (Top Performers) Danh sách các cầu thủ dẫn đầu cho thấy sự phân hóa vai trò rõ rệt:

- Khả năng đột biến:** Buka... là nhân tố chủ chốt trong việc tinh tiến bóng với 30 lần *Progressive carries*.
- Điều tiết lối chơi:** Decla... đóng vai trò "trạm trung chuyển" với 713 lần chạm bóng, cao nhất đội hình.
- Hệ thống phòng ngự:** Jurrie... (21 tắc bóng) và Decla... (18 tắc bóng) tạo thành lá chắn vững chắc ở khu trung tuyến.

3. Phân tích lối chơi và cấu trúc chiến thuật

Sử dụng dữ liệu từ các biểu đồ xu hướng và phân bổ vị trí:

- a) *Phân bổ khu vực chạm bóng (Ball Touch Area Allocation)*: Biểu đồ vùng cho thấy sự ổn định của **Mid 3rd** (Khu vực giữa sân). Điều này khẳng định đội bóng ưu tiên kiểm soát bóng (Possession-based) thay vì đá trực diện hoàn toàn.
- b) *Hiệu suất theo vị trí (Pass by Position)*: Số lượng đường chuyền tập trung cực lớn vào nhóm **CB** và **DM**. Công thức tính hiệu suất triển khai bóng:

$$\text{Pass Accuracy} = \frac{\sum \text{Pass Complete}}{\sum \text{Pass Attempt}} \times 100\%$$

Tỷ lệ này đạt mức tối ưu ở các vị trí lùi sâu, hỗ trợ quá trình thoát pressing từ sân nhà.

4. Nhận xét về xu hướng theo thời gian

Biểu đồ *Ball touch area* qua các vòng đấu (Round) cho thấy:

- Đội bóng duy trì khống đội hình dâng cao ổn định từ vòng 1 đến vòng 15.
- Có sự biến động mạnh về khu vực chạm bóng ở các vòng gần nhất (vòng 17-20), cần đổi chiều với lịch thi đấu để xem đây có phải là các trận gặp đối thủ mạnh (Big Six) hay không.

5. Kết luận và Đề xuất cải thiện

- ✓ **Kết luận:** Đội bóng đang vận hành theo lối chơi kiểm soát chủ động, dựa trên khả năng chuyền bóng của hàng thủ và khả năng kéo bóng của các tiền đạo cánh.
- ✓ **Đề xuất:** Cần bổ sung thêm biểu đồ *Scatter Plot* giữa "Tỷ lệ kiểm soát" và "Tốc độ tấn công" để làm rõ hơn bản sắc phản công khi cần thiết.

Kết luận

Kết quả đồ án

Tổng thể về đồ án của em đã đạt được một số kết quả sau:

- Lấy được dữ liệu từ trang web FBRef bằng cách sử dụng Web Scraping với Selenium và Webdriver
- Xây dựng được hệ thống Data Warehouse và thiết kế được quy trình ETL tự động
- Xây dựng được dashboard bằng ứng dụng Power BI
- Tổng hợp được quá trình làm và viết thành báo cáo hoàn chỉnh

Kỹ năng đạt được

- Biết sử dụng python để cào dữ liệu từ các trang web
- Biết sử dụng SQL Server để tổ chức sắp xếp và lưu trữ dữ liệu
- Biết sử dụng ngôn ngữ DAX để xây dựng các báo cáo trực quan trong Power BI
- Biết viết báo cáo bằng Latex
- Biết sử dụng các hệ thống AI chatbot trong quá trình làm.

Hướng phát triển tương lai

Hiện tại đồ án vẫn còn một số hạn chế, trong thời gian tới em sẽ tiếp tục nghiên cứu và phát triển đồ án theo một số hướng như:

- Sử dụng các công cụ chuyên biệt cho việc thu thập, xử lý và lưu trữ dữ liệu: Google Cloud, Apache Airflow

- Xây dựng các mô hình dự đoán hiệu suất cầu thủ
- Xây dựng website để public kết quả phân tích cũng như kết quả dự đoán của các mô hình
- Sử dụng Docker để đóng gói sản phẩm

Tài liệu tham khảo

- [1] ThS. Nguyễn Danh Tú (2023). *Slide bài giảng Kho dữ liệu và Kinh doanh thông minh*. Khoa Toán - Tin, Đại học Bách khoa Hà Nội.
- [2] ThS. Nguyễn Danh Tú (2025). *Giáo trình Kho dữ liệu và kinh doanh thông minh*.