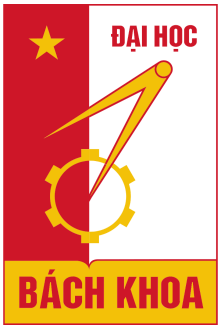
NGUYỄN PHÚ VINH

ĐẠI HỌC BÁCH KHOA HÀ NỘI
KHOA TOÁN - TIN



NGUYỄN PHÚ VINH

XÂY DỰNG KHO DỮ LIỆU CHO PHÂN TÍCH BÓNG ĐÁ

ĐỒ ÁN II

Chuyên ngành: TOÁN TIN

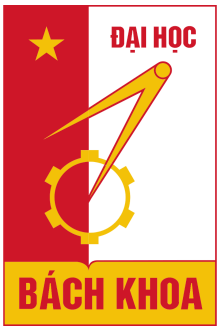
HÀ NỘI - 2025

NGUYỄN PHÚ VINH

HÀ NỘI - 2025

ĐẠI HỌC BÁCH KHOA HÀ NỘI

KHOA TOÁN - TIN



XÂY DỰNG KHO DỮ LIỆU CHO PHÂN TÍCH BÓNG ĐÁ

ĐỒ ÁN II

Chuyên ngành: TOÁN TIN

Giảng viên hướng dẫn: TS. Nguyễn Đình Hân _____
Chữ kí của GVHD

Sinh viên thực hiện: Nguyễn Phú Vinh

MSSV: 20227169

Lớp: Toán-Tin 01 – K67

HÀ NỘI - 2025

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. Mục tiêu và nội dung của đồ án

.....

.....

.....

.....

.....

.....

2. Kết quả đạt được

.....

.....

.....

.....

.....

.....

3. Ý thức làm việc của sinh viên

.....

.....

.....

.....

.....

.....

Hà Nội, ngày ... tháng ... năm 2025

Giảng viên hướng dẫn

Mục lục

Danh sách hình vẽ

Bảng ký hiệu và chữ viết tắt

xG	Xác suất một cú sút thành bàn (Expected Goals)
$G - xG$	Hiệu số giữa tổng số bàn thắng thực tế và tổng xG của cầu thủ hoặc đội bóng (Goals minus Expected Goals)
xA	Xác suất một đường chuyền trở thành kiến tạo (Expected Assists)
$PPDA$	Số đường chuyền trung bình của đội B trong khu vực 2/3 sân cuối cùng (có tọa độ $x \geq 40$ trên sân có kích cỡ 120×80) trước khi đội A thực hiện một hành động phòng ngự (Passes Per Defensive Action)
PRR	Tỷ lệ đội bóng giành lại quyền kiểm soát bóng trong vòng vài giây sau khi thực hiện một hành động gây áp lực (Pressure Regain Rate)
$TiB/90$	Số lần chạm bóng trong vòng cấm của đối phương, được chuẩn hóa theo 90 phút thi đấu (Touches in Box/90)
$PAdjI/90$	Số lần cắt bóng đã điều chỉnh theo quyền kiểm soát bóng (Possession-Adjusted Interceptions/90)
TSR	Tỷ lệ tắc bóng thành công (Tackles Success Rate)
DW	Kho dữ liệu (Data Warehouse)
OLTP	Hệ thống xử lý giao dịch trực tuyến (Online Transaction Processing)
OLAP	Hệ thống phân tích trực tuyến (Online Analytical Processing)
BI	Kinh doanh thông minh (Business Intelligence)
EDW	Kho dữ liệu doanh nghiệp (Enterprise Data Warehouse)

Lời mở đầu

Trong thời đại số hóa, dữ liệu đã trở thành một trong những yếu tố then chốt trong việc đưa ra quyết định và định hướng chiến lược ở mọi lĩnh vực, bao gồm cả thể thao.

Trong bóng đá, việc phân tích hiệu suất cầu thủ, tối ưu hóa chiến thuật, điều chỉnh giáo án tập luyện và đánh giá trận đấu đều phụ thuộc vào khả năng thu thập, xử lý và phân tích lượng dữ liệu khổng lồ (bao gồm dữ liệu sự kiện, dữ liệu theo dõi vị trí, dữ liệu vật lý, ...). Để khai thác tối đa giá trị của nguồn dữ liệu này, việc xây dựng một kho dữ liệu (Data Warehouse) hiệu quả và đầy đủ là rất cần thiết.

Xuất phát từ lý do trên, em quyết định lựa chọn đề tài "Xây dựng kho dữ liệu cho phân tích bóng đá". Đồ án mong muốn xây dựng một kho dữ liệu giúp giải quyết các bài toán phân tích, dự báo, và cung cấp những góc nhìn đa chiều về dữ liệu bóng đá, hỗ trợ công tác huấn luyện và quản lý bóng đá hiệu quả hơn.

Ngoài phần Mở đầu và Kết luận, đồ án của em sẽ bao gồm 4 chương chính:

- Chương I: Cơ sở lý thuyết.
- Chương II: Điều tra, khảo sát hệ thống.
- Chương III: Phân tích và thiết kế hệ thống.
- Chương IV: Cài đặt hệ thống.

Hà Nội, tháng 10 năm 2025

Sinh viên

Nguyễn Phú Vinh

Lời cảm ơn

Em xin gửi lời cảm ơn chân thành đến thầy Nguyễn Đình Hân, người đã tận tình hướng dẫn và đồng hành cùng em trong suốt quá trình thực hiện đồ án này. Sự chỉ bảo tận tâm cùng những ý kiến đóng góp quý giá của thầy đã giúp em xác định hướng đi đúng đắn và vượt qua nhiều thử thách trong quá trình phát triển phần mềm. Em cũng xin gửi lời cảm ơn chân thành đến các thầy cô Khoa Toán-Tin, Đại học Bách khoa Hà Nội. Sự tận tâm giảng dạy và kiến thức mà các thầy cô truyền đạt đã giúp em tự tin áp dụng lý thuyết vào thực tiễn, góp phần quan trọng vào việc hoàn thiện đồ án này.

Dù đã nỗ lực hết mình để thực hiện và hoàn thành đồ án, em nhận thấy sản phẩm của mình vẫn còn những thiếu sót. Vì vậy, em rất mong nhận được những ý kiến nhận xét quý báu từ thầy cô để có thể cải thiện đồ án tốt hơn, đồng thời tích lũy thêm kinh nghiệm thực tế cho bản thân.

Em xin chân thành cảm ơn!

Chương 1

Cơ sở lý thuyết

1.1 Giới thiệu về phân tích dữ liệu bóng đá

Bóng đá (hay còn gọi là túc cầu, đá bóng, đá banh) là một môn thể thao đồng đội được chơi với quả bóng hình cầu giữa hai đội gồm 11 cầu thủ mỗi bên. Môn thể thao này có khoảng hơn 250 triệu người chơi ở hơn 200 quốc gia và vùng lãnh thổ, khiến nó trở thành môn thể thao phổ biến nhất trên thế giới. Môn này chơi trên một mặt sân hình chữ nhật với một khung thành ở mỗi đầu. Mục tiêu là ghi bàn vào khung thành đối phương. Đội nào có số bàn thắng nhiều hơn sẽ giành chiến thắng. [1].

Trong bóng đá hiện đại, các kỹ thuật, công nghệ hỗ trợ cho việc phân tích, đánh giá ngày càng trở nên phổ biến hơn vì những lợi ích mà chúng mang lại. Rất nhiều đội bóng trên toàn thế giới, đặc biệt là các đội bóng giàu thành tích tại các giải đấu hàng đầu châu Âu, có thể sẵn sàng chi những số tiền rất lớn để đầu tư vào những công nghệ này nhằm cải thiện thành tích cho đội bóng, nâng cao hiệu quả trong công tác huấn luyện, thi đấu, đào tạo các cầu thủ trẻ tài năng hay thậm chí là để có thể mang về những bản hợp đồng chất lượng, đáng tiền trong mỗi kì chuyển nhượng căng thẳng. Nhờ vậy, dữ liệu được tổng hợp từ các trận đấu lại trở thành nguồn tài nguyên vô cùng quý giá đối với họ, điều này đã phần nào phản ánh tầm quan trọng của một kho dữ liệu lưu trữ nguồn tài nguyên này để phục vụ cho sự phân tích, đánh giá của các chuyên gia.

Trong một trận đấu, điều mà những cổ động viên cuồng nhiệt lưu tâm đến không chỉ là những bàn thắng. Đó còn là phong cách chơi bóng độc đáo của các

cầu thủ trên sân, những đường chuyền, đường kiến tạo đẹp mắt, những tình huống tranh chấp quyết liệt, những pha cản phá xuất thần của hậu vệ hoặc thủ môn hay những tình huống cố định, tình huống phản công,... Tất cả đều có thể được hiểu đơn giản là những sự kiện diễn ra trong một trận đấu. Nhưng ẩn sâu trong những dữ liệu sự kiện đó, các chuyên gia phân tích thường quan tâm đến các chỉ số sau:

- xG (Expected Goals): Xác suất một cú sút thành bàn, với giá trị dao động từ 0 đến 1, được tính dựa trên dữ liệu lịch sử của hàng nghìn cú sút có đặc điểm (vị trí tọa độ, góc sút, khoảng cách tới khung thành, bộ phận cơ thể, loại cơ hội,...) tương tự. Chỉ số này giúp đánh giá chất lượng cơ hội.

- $G - xG$ (Goals minus Expected Goals): Hiệu số giữa tổng số bàn thắng thực tế và tổng xG của cầu thủ hoặc đội bóng. Chỉ số này giúp đánh giá khả năng dứt điểm thành bàn của cầu thủ hoặc đội bóng.

- xA (Expected Assists): Xác suất một đường chuyền trở thành kiến tạo, được tính bằng cách lấy xG của cú sút ngay sau đường chuyền đó. Chỉ số này giúp đánh giá khả năng tạo cơ hội.

- $PPDA$ (Passes Per Defensive Action): Số đường chuyền trung bình của đội B trong khu vực 2/3 sân cuối cùng (có tọa độ $x \geq 40$ trên sân có kích cỡ 120×80) trước khi đội A thực hiện một hành động phòng ngự. Đây là chỉ số đo lường mức độ bị ép sân của đội A.

$$\textbf{PPDA}_{\textbf{A}} = \frac{\text{Số đường chuyền của B trong khu vực } x \geq 40}{\text{Số sự kiện phòng ngự của A trong khu vực } x \geq 40}$$

- PRR (Pressure Regain Rate): Tỷ lệ đội bóng giành lại quyền kiểm soát bóng trong vòng vài giây sau khi thực hiện một hành động gây áp lực. Chỉ số này giúp đo lường hiệu quả của chiến thuật gây áp lực (Pressing) và khả năng cầm bóng.

$$\textbf{PRR} = \frac{\text{Số pha giành lại bóng thành công sau khi gây áp lực}}{\text{Số lần gây áp lực}}$$

- $TiB/90$ (Touches in Box/90): Số lần chạm bóng trong vòng cấm của đối phương, được chuẩn hóa theo 90 phút thi đấu. Chỉ số này giúp đánh giá khả năng chọn vị trí và độ nguy hiểm khi tham gia tấn công của cầu thủ.

$$\textbf{TiB/90} = \frac{\text{Tổng số lần chạm bóng trong vòng cấm}}{\text{Tổng số phút đã chơi}} \times 90$$

- $PAdjI/90$(Possession-Adjusted Interceptions/90): Số lần cắt bóng đã điều chỉnh theo quyền kiểm soát bóng. Chỉ số này đo lường số lần một cầu thủ cắt đường chuyền của đối phương trong 90 phút, sau đó điều chỉnh bằng một hệ số dựa trên thời gian đội đó không kiểm soát bóng.

$$\textbf{PAdjI/90} = \frac{\text{Tổng số lần cắt bóng}}{\text{Tổng số phút đã chơi}} \times 90 \times \frac{\text{Tỷ lệ \% kiểm soát bóng đội bạn}}{\text{Tỷ lệ \% kiểm soát bóng đội nhà}}$$

- TSR (Tackles Success Rate): Tỷ lệ tắc bóng thành công. Chỉ số này đánh giá khả năng tắc bóng chính xác, sự quyết đoán trong khâu phòng ngự của cầu thủ.

$$\textbf{TSR} = \frac{\text{Tổng số lần tắc bóng thành công}}{\text{Tổng số lần tắc bóng}} \times 100\%$$

Sử dụng những chỉ số như vậy, các chuyên gia bóng đá có thể đưa ra các phân tích, đánh giá cơ bản và chuyên sâu trong một số khía cạnh:

- Phân tích, đánh giá phong độ của cầu thủ, tìm kiếm và phát hiện tài năng ẩn trong phong cách chơi của cầu thủ.
- Điều chỉnh giáo án tập luyện, chiến thuật, vị trí thi đấu hợp lý, phù hợp với từng đối thủ, từng giải đấu hoặc từng cầu thủ.
- Phân tích, đánh giá điểm mạnh và rủi ro trong hệ thống vận hành của đội bóng.
- Tư vấn chuyển nhượng và tuyển dụng.
- Dự đoán kết quả thi đấu và phong độ của cầu thủ trong tương lai.

1.2 Tổng quan về Kho dữ liệu (Data Warehouse) và Phân tích xử lý trực tuyến (OLAP)

1.2.1 Kho dữ liệu (Data Warehouse)

Khái niệm

Kho dữ liệu (Data Warehouse - DW) là một cơ sở dữ liệu lớn, tập trung, lưu trữ dữ liệu lịch sử từ nhiều nguồn khác nhau đã được tích hợp và cấu trúc hóa riêng biệt cho mục đích phân tích.

Khái niệm này phân biệt rõ ràng DW với các hệ thống tác nghiệp (Online Transaction Processing - OLTP):

- **Mục tiêu**: Trong khi các hệ thống OLTP được tối ưu cho việc vận hành kinh doanh hàng ngày (xử lý các giao dịch nhỏ, nhanh), DW được tối ưu cho việc phân tích và ra quyết định (Online Analytical Processing - OLAP).
- **Chức năng**: DW hoạt động như trái tim của kinh doanh thông minh (Business Intelligence - BI). Nó giúp hợp nhất dữ liệu từ nhiều nguồn, đảm bảo tính nhất quán và cung cấp cái nhìn toàn cảnh.
- **Triết lý thiết kế**: DW thường sử dụng mô hình phi chuẩn hóa, phổ biến nhất là mô hình đa chiều (Dimensional Model), như lược đồ sao (Star Schema). Triết lý này ưu tiên tốc độ truy vấn phân tích bằng cách giảm số lượng phép JOIN, vốn là vấn đề của các cơ sở dữ liệu chuẩn hóa cao (3NF) trong OLTP.

Tính chất

- **Hướng chủ đề**: Dữ liệu trong DW được tổ chức xoay quanh các chủ đề kinh doanh chính thay vì theo các quy trình nghiệp vụ của từng phòng ban như hệ thống OLTP, giúp cung cấp một cái nhìn toàn diện về một chủ đề cụ thể, hợp nhất dữ liệu liên quan từ nhiều hệ thống nguồn khác nhau.
- **Tích hợp**: Dữ liệu từ các nguồn khác nhau phải được tổng hợp và nhất quán hóa. Sự tích hợp này thể hiện ở việc áp dụng các quy ước đặt tên chung, đơn vị đo lường thống nhất và định dạng dữ liệu chuẩn trên toàn bộ kho dữ liệu.

- **Bất biến**: Dữ liệu một khi đã được nạp vào DW thì gần như không bao giờ bị xóa hay cập nhật. DW là một kho lưu trữ lịch sử. Khi có sự thay đổi trong hệ thống nguồn (ví dụ: khách hàng đổi địa chỉ), DW không ghi đè lên địa chỉ cũ mà sẽ thêm một bản ghi mới để ghi nhận sự thay đổi đó theo thời gian. Đặc tính này là đặc biệt quan trọng để phân tích xu hướng và so sánh lịch sử.
- **Tính thời gian**: Mọi dữ liệu trong DW đều được gắn với một yếu tố thời gian (ngày, quý, năm). Kiến trúc của DW luôn được thiết kế để cho phép phân tích theo dòng thời gian (ví dụ: so sánh doanh thu quý này so với cùng kỳ năm ngoái). Đây là điểm khác biệt cốt lõi so với hệ thống OLTP, vốn chỉ quan tâm đến trạng thái hiện tại.

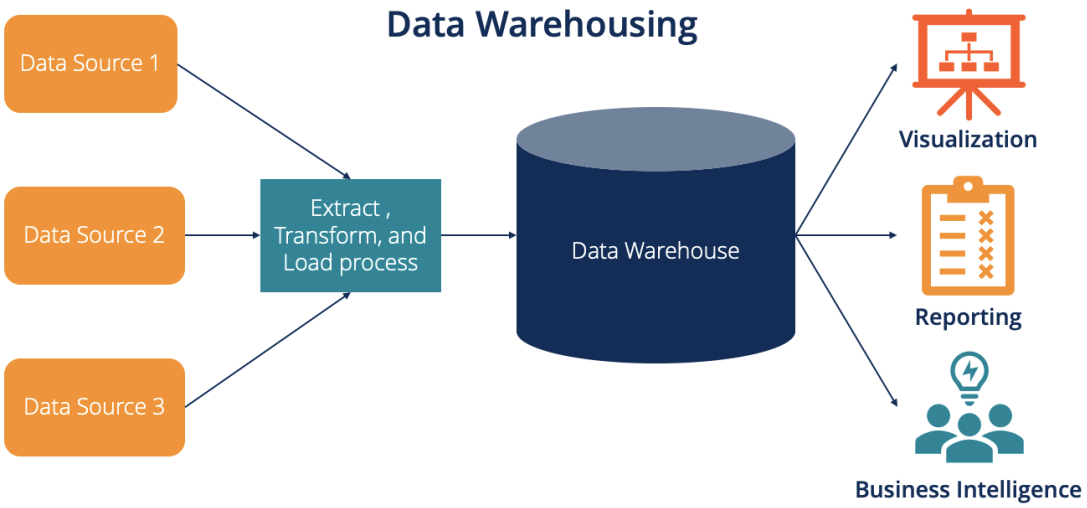
Ưu điểm

- **Hỗ trợ ra quyết định chiến lược**: DW/BI là công cụ mạnh mẽ thúc đẩy chuyển đổi dữ liệu thô thành trí tuệ để dẫn dắt chiến lược. Nó hỗ trợ ra quyết định ở cả ba cấp độ: chiến lược, chiến thuật và tác nghiệp.
- **Tính nhất quán và độ tin cậy**: DW giúp đảm bảo tính đúng đắn của dữ liệu.
- **Phân tích lịch sử**: Khả năng lưu trữ dữ liệu lịch sử chi tiết cho phép phân tích xu hướng dài hạn.
- **Hiệu năng phân tích cao**: Thiết kế theo mô hình đa chiều (phi chuẩn hóa) giúp tăng tốc độ truy vấn đáng kể, giảm thiểu các phép JOIN khi tổng hợp dữ liệu quy mô lớn.
- **Nền tảng cho Machine Learning/AI**: Dữ liệu sạch, tích hợp và có bối cảnh lịch sử trong DW giúp giảm thiểu thời gian chuẩn bị dữ liệu cho các mô hình học máy.

Nhược điểm

- **Độ trễ dữ liệu**: Dữ liệu trong DW thường được cập nhật theo lô, dẫn đến độ trễ nhất định so với dữ liệu thời gian thực.

- **Chi phí và thời gian triển khai ban đầu**: Việc xây dựng một DW truyền thống đòi hỏi chi phí đầu tư ban đầu lớn cho cơ sở hạ tầng, phần mềm, nhân lực, công cụ tiền xử lý và hỗ trợ kỹ thuật, có thể mất nhiều thời gian để thấy được giá trị.
- **Khả năng xử lý dữ liệu phi cấu trúc**: DW truyền thống được thiết kế tối ưu cho dữ liệu có cấu trúc. Nó gặp khó khăn khi xử lý dữ liệu bán cấu trúc và phi cấu trúc.
- **Tính cứng nhắc**: Mô hình phải được định nghĩa trước khi dữ liệu được nạp vào. Việc thay đổi cấu trúc DW sau này có thể phức tạp và tốn kém.



Hình 1.1: Xây dựng và ứng dụng kho dữ liệu

Kiến trúc kho dữ liệu cơ bản

Mô hình kiến trúc	Đặc điểm chính	Ưu điểm/Hạn chế
Một tầng	Hoạt động như một hệ thống ảo hoặc lớp trung gian để tổng hợp dữ liệu, nhằm giảm thiểu sự dư thừa dữ liệu trong quá trình lưu trữ.	Ít được sử dụng trong thực tế vì hạn chế về khả năng mở rộng và tích hợp dữ liệu, bao gồm việc hợp nhất dữ liệu và loại bỏ trùng lặp.
Hai tầng	Bao gồm Nguồn dữ liệu, Khu vực xử lý tạm thời (cho quy trình ETL), Lớp kho dữ liệu (lưu trữ dữ liệu đã xử lý, Data Marts, Metadata), và Lớp phân tích/báo cáo.	Đơn giản hơn, nhưng khả năng tích hợp dữ liệu có thể chưa tối ưu.
Ba tầng	Mô hình phổ biến nhất, đặc biệt trong các doanh nghiệp lớn, nhờ khả năng tổ chức dữ liệu khoa học và hiệu quả. Bao gồm 3 lớp: 1. Lớp nguồn dữ liệu (Source Layer) 2. Lớp xử lý trung gian (Intermediate Processing Layer) 3. Lớp kho dữ liệu (DW Layer).	Đòi hỏi không gian lưu trữ lớn cho lớp xử lý trung gian và có thể gặp hạn chế trong việc phân tích dữ liệu theo thời gian thực.

Kiến trúc BI tổng thể

Các Thành phần Cấu thành Cốt lõi

Tầng nguồn (Source Layer) và tích hợp Dữ liệu

- **Nguồn dữ liệu (Data Sources)**: Là điểm khởi đầu, bao gồm các hệ thống tác nghiệp (OLTP), hoặc các nguồn bên ngoài (file Excel, dữ liệu mạng xã

hội). Tầng này có thể chứa nhiều loại dữ liệu: có cấu trúc, bán cấu trúc (JSON, XML), và phi cấu trúc (video, log server).

- **Vùng trung chuyển (Staging Area)**: Là khu vực lưu trữ trung gian.
  - **Vai trò**: Dữ liệu thô sau khi trích xuất (Extract) sẽ được đưa vào đây. Mọi quá trình biến đổi (Transform) như làm sạch, chuẩn hóa, kết hợp và định hình lại dữ liệu sẽ diễn ra tại đây.
  - **Mục đích**: Cách ly quá trình xử lý nặng khỏi hệ thống nguồn để không làm chậm hệ thống tác nghiệp.

Tầng lưu trữ (Storage Layer) Đây là nơi dữ liệu đã được xử lý và tích hợp được lưu trữ.

- **Kho dữ liệu doanh nghiệp (Enterprise Data Warehouse - EDW)**:
  - Là một cơ sở dữ liệu tập trung, lớn, lưu trữ dữ liệu lịch sử đã được tích hợp và cấu trúc hóa cho mục đích phân tích.
- **Kho dữ liệu chủ đề (Data Marts)**:
  - Là các tập con nhỏ hơn, chuyên biệt, trích xuất từ kho dữ liệu doanh nghiệp hoặc được xây dựng riêng.
  - Được thiết kế để phục vụ nhu cầu phân tích của một phòng ban hoặc lĩnh vực nghiệp vụ cụ thể.
- **Kho siêu dữ liệu (Metadata Repository)**: Nơi lưu trữ thông tin về nguồn gốc, cấu trúc bảng, các phép biến đổi, và cách truy cập dữ liệu.

Tầng phân tích và trình bày (Analytics and Presentation Layer) Đây là nơi dữ liệu được chuyển hóa thành tri thức và giao tiếp đến người dùng cuối.

- **Động cơ phân tích / Khối OLAP**: Tầng xử lý các truy vấn phức tạp trên dữ liệu, thường sử dụng các kỹ thuật OLAP. Khối dữ liệu đa chiều cho phép người dùng thực hiện các thao tác phân tích như Drill-Down, Roll-Up, Slice, Dice, và Pivot.

- **Lớp ngữ nghĩa (Semantic Layer)**: Một lớp trừu tượng ánh xạ cấu trúc kỹ thuật (bảng, cột) sang các thuật ngữ kinh doanh dễ hiểu (ví dụ: "doanh thu", "lợi nhuận").
- **Công cụ BI và Báo cáo**: Giao diện người dùng cuối tương tác, bao gồm các báo cáo (Reports) và dashboard tương tác.

Luồng dữ liệu và kiến trúc Hiện đại (ELT/Lakehouse)

- **Quy trình ETL/ELT**: ETL (Extract, Transform, Load) là quy trình nơi biến đổi dữ liệu diễn ra ở máy chủ trung gian. ELT (Extract, Load, Transform) là mô hình hiện đại hơn, nơi dữ liệu thô được tải thẳng (Load) vào kho dữ liệu đám mây trước, sau đó mới dùng sức mạnh xử lý của chính kho dữ liệu để biến đổi (Transform).
- **Sự kết hợp Data Lake**: Một kiến trúc hiện đại thường bao gồm Hồ dữ liệu (Data Lake), nơi lưu trữ mọi loại dữ liệu (có cấu trúc, phi cấu trúc) ở định dạng thô.
- **Kiến trúc Lakehouse**: Là sự hợp nhất của Data Lake và Data Warehouse, nhằm phá bỏ sự phức tạp và dư thừa của kiến trúc hai tầng truyền thống. Data Lakehouse cung cấp một nền tảng duy nhất để phục vụ cho cả phân tích kinh doanh (BI) và khoa học dữ liệu (Machine Learning/AI).

1.2.2 Hệ thống phân tích xử lý trực tuyến (OLAP)

Trong kiến trúc kho dữ liệu, OLAP (Online Analytical Processing) là thành phần chủ đạo của tầng phân tích và trình bày. OLAP là cơ chế chuyển hóa dữ liệu lịch sử đã được tích hợp thành tri thức có thể hành động được.

Định nghĩa và vai trò của hệ thống OLAP

OLAP (Online Analytical Processing) là một giải pháp phân tích dữ liệu mạnh mẽ, được thiết kế để xử lý và khai thác thông tin từ nhiều góc độ khác nhau với hiệu suất cao, ngay cả khi làm việc với khối lượng dữ liệu khổng lồ. Các hệ thống

OLAP được sinh ra để giải quyết những câu hỏi phân tích phức tạp, hỗ trợ các truy vấn trên một khối lượng lớn dữ liệu lịch sử.

Vai trò cốt lõi của OLAP bao gồm:

- **Hỗ trợ phân tích và ra quyết định**: Hệ thống giúp tổ chức đưa ra các quyết định kinh doanh tốt hơn.
- **Phân tích đa chiều**: Cung cấp khả năng "nhìn" dữ liệu từ nhiều góc độ khác nhau (đa chiều) để tìm ra xu hướng, mẫu và tri thức ẩn.
- **Hỗ trợ dự báo và lập kế hoạch**: Với khả năng phân tích vượt trội, OLAP giúp doanh nghiệp xây dựng các kế hoạch chiến lược và dự báo các kịch bản trong tương lai.
- **Tối ưu hóa hoạt động**: Giúp nhận diện các vấn đề cần cải thiện trong quy trình kinh doanh, từ đó tăng hiệu quả vận hành.
- **Người dùng**: Nhà phân tích dữ liệu, nhà quản lý, lãnh đạo cấp cao, v.v.

Phân biệt với hệ thống xử lý giao dịch trực tuyến (OLTP)

- **Thiết kế cơ sở dữ liệu**: OLAP sử dụng mô hình dữ liệu phi chuẩn hóa, như lược đồ sao (Star Schema), để giảm số lượng phép JOIN và tăng tốc độ truy vấn phân tích. Ngược lại, OLTP yêu cầu chuẩn hóa cao (ví dụ: 3NF) để đảm bảo tính toàn vẹn dữ liệu.
- **Loại thao tác**: OLAP chủ yếu thực hiện các thao tác đọc và tổng hợp trên hàng triệu bản ghi. Các thao tác ghi (INSERT, UPDATE) rất hạn chế và thường diễn ra theo lô.

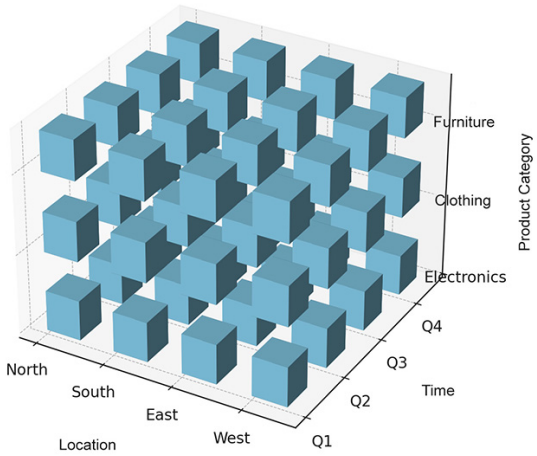
Mô hình khối dữ liệu đa chiều (OLAP Cube)

Khối OLAP (OLAP Cube) là một cấu trúc dữ liệu đa chiều được tối ưu hóa để truy vấn và phân tích nhanh, là hiện thực hóa của lớp ngữ nghĩa (Semantic Layer).

- **Lớp ngữ nghĩa**: Là lớp trừu tượng nằm giữa người dùng và cơ sở dữ liệu vật lý. Nó chuyển đổi các cấu trúc bảng phức tạp thành các thuật ngữ kinh

doanh dễ hiểu (ví dụ: "doanh thu", "lợi nhuận"). Điều này giải quyết vấn đề người dùng nghiệp vụ không quen thuộc với SQL hoặc cấu trúc bảng Fact/Dimension.

- **Tính toán trước**: Khối OLAP thường tính toán trước các giá trị tổng hợp ở nhiều cấp độ khác nhau để các truy vấn có thể được trả về gần như tức thời.



Hình 1.2: Khối OLAP

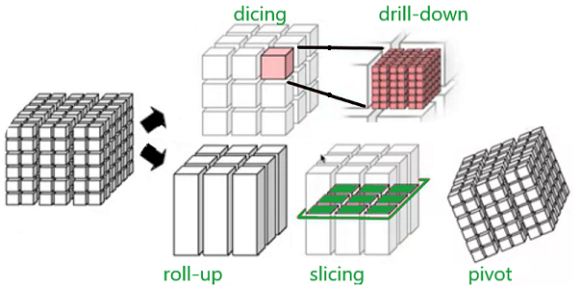
Thành phần khối dữ liệu

Khối dữ liệu OLAP được xây dựng dựa trên lược đồ sao, bao gồm hai thành phần chính:

1. **Chỉ số đo lường (Measures) / Bảng Fact**:
   - Là các cột Fact trong Bảng Fact.
   - Đây là các giá trị số mà ta muốn đo đếm và phân tích.
2. **Các chiều (Dimensions) / Bảng Dimension**:
   - Là các bảng Dimension.
   - Chúng trở thành các trục dùng để phân tích Measures, cung cấp bối cảnh cho các con số.
   - Các thuộc tính phân cấp (ví dụ: Năm, Quý, Tháng) tạo thành các hệ thống phân cấp (Hierarchies) bên trong chiều.

Các phép toán phân tích OLAP

1. **Drill-Down (Đào sâu)**:
   - Điều hướng từ một cấp độ tổng hợp cao hơn xuống một cấp độ chi tiết hơn trong một hệ thống phân cấp (ví dụ: từ xem Doanh thu theo Năm, xuống xem theo Quý).
2. **Roll-Up (Tổng hợp)**:
   - Tổng hợp dữ liệu từ một cấp độ chi tiết lên một cấp độ tổng quan hơn (ví dụ: từ xem theo Thành phố, lên xem theo Vùng).
3. **Slice (Cắt lát)**:
   - Lọc dữ liệu theo một chiều, chọn một giá trị duy nhất cho một chiều để xem một "lát cắt" của khối dữ liệu (ví dụ: đặt bộ lọc chỉ xem khu vực Miền Bắc).
4. **Dice (Cắt khối)**:
   - Lọc dữ liệu theo nhiều chiều, chọn các giá trị cụ thể trên hai hoặc nhiều chiều khác nhau để xem một "khối con" của dữ liệu (ví dụ: xem doanh thu của ngành hàng "Thời trang Nữ" tại "TP.HCM" trong "Quý 4").
5. **Pivot (Xoay)**:
   - Thay đổi cách dữ liệu được hiển thị mà không thay đổi giá trị của dữ liệu. Nó cho phép hoán đổi vị trí của các chiều giữa trục hàng và trục cột (ví dụ: thay vì xem doanh thu theo sản phẩm ở hàng và khu vực ở cột, người dùng chuyển sang khu vực ở hàng và sản phẩm ở cột).



Hình 1.3: Các phép toán trên khối OLAP

1.3 Quy trình tích hợp dữ liệu ETL và ELT

Quy trình tích hợp dữ liệu là xương sống vận hành của một kho dữ liệu. Nhiệm vụ của nó là di chuyển và chuẩn bị dữ liệu từ các hệ thống cơ sở dữ liệu tác nghiệp (OLTP) sang mô hình phân tích (OLAP) có trật tự và nhất quán. Quy trình này đảm bảo dữ liệu đi vào kho là sạch, đáng tin cậy và sẵn sàng cho việc phân tích kinh doanh (BI).

Có hai kiến trúc tích hợp dữ liệu chính được sử dụng: ETL truyền thống và ELT hiện đại.

1.3.1 Kiến trúc ETL Kinh điển

ETL (Extract - Trích xuất, Transform - Biến đổi, Load - Tải) là quy trình cốt lõi chịu trách nhiệm di chuyển và chuẩn bị dữ liệu cho kho dữ liệu. Một hệ thống ETL cấp doanh nghiệp là một tập hợp phức tạp của nhiều hệ thống con, quản lý mọi thứ từ làm sạch dữ liệu cho đến xử lý lỗi và lập lịch.

Kiến trúc ETL điển hình sử dụng một vùng trung chuyển (Staging Area), nơi các phép biến đổi phức tạp diễn ra. Việc này giúp giảm thiểu tác động lên hệ thống nguồn và đảm bảo kho dữ liệu đích chỉ nhận vào dữ liệu đã sạch.

Giai đoạn E - Trích xuất (Extract)

Mục tiêu: Đọc và lấy dữ liệu từ một hoặc nhiều hệ thống nguồn. Dữ liệu nguồn có thể từ cơ sở dữ liệu quan hệ, file phẳng (CSV, Excel) cho đến các API.

Các phương pháp Trích xuất cốt lõi:

- **Trích xuất Toàn bộ**: Sao chép toàn bộ bảng mỗi lần chạy, chỉ phù hợp với các bảng dữ liệu nhỏ, ít thay đổi.
- **Trích xuất tăng trưởng**: Chỉ trích xuất những dữ liệu đã thay đổi kể từ lần cuối cùng, tối ưu cho các bảng lớn (như bảng giao dịch bán hàng của sàn thương mại điện tử).
- **Trích xuất từ các nguồn phức tạp**: Đối với các nguồn như API hoặc file, quy trình trích xuất phải xử lý các thách thức như giới hạn số lần gọi, phân

trang và cấu trúc không nhất quán.

Giai đoạn T - Biến đổi (Transform)

Mục tiêu: Chuyển đổi dữ liệu thô, không nhất quán và phân mảnh thành một bộ dữ liệu sạch, tuân thủ các quy tắc nghiệp vụ và có cấu trúc phù hợp với mô hình lược đồ sao đã thiết kế. Các tác vụ chính diễn ra tại vùng trung chuyển bao gồm:

- **Làm sạch và Chuẩn hóa dữ liệu**:
  - **Phân tách cấu trúc**: Tách các cấu trúc phức tạp (như dòng log web hoặc JSON) thành các cột riêng biệt có ý nghĩa.
  - **Chuẩn hóa**: Đưa các giá trị khác nhau nhưng cùng ngữ nghĩa về một dạng chuẩn duy nhất (ví dụ: ánh xạ "HN" về "Hà Nội").
  - **Xử lý giá trị NULL**: Thay thế bằng giá trị mặc định hoặc loại bỏ bản ghi tùy theo chiến lược.
  - **Xác thực dữ liệu**: Kiểm tra xem dữ liệu có vi phạm các quy tắc nghiệp vụ không.
- **Tích hợp dữ liệu và Tạo khóa**:
  - **Loại bỏ Trùng lặp và Hợp nhất**: Định nghĩa các quy tắc để xác định và hợp nhất các bản ghi trùng lặp từ nhiều nguồn (ví dụ: cùng một khách hàng có mã khác nhau ở POS và CRM).
  - **Tạo khóa thay thế**: Quy trình ETL phải gán một khóa thay thế mới, đơn giản và có thứ tự cho mỗi giá trị của bảng Dimension, thay vì sử dụng khóa nghiệp vụ từ hệ thống nguồn.
  - **Hiện thực hóa Logic SCD**: Áp dụng logic SCD loại 1, loại 2, hoặc loại 3 để theo dõi lịch sử thay đổi của các thuộc tính Dimension (ví dụ: địa chỉ khách hàng thay đổi theo thời gian).
- **Biến đổi cho Bảng Fact**:
  - **Tra cứu và Thay thế Khóa**: Thay thế tất cả các khóa nghiệp vụ từ nguồn bằng các khóa thay thế tương ứng từ các bảng Dimension.

  - **Tính toán chỉ số đo lường**: Tính toán trước các chỉ số đo lường mới (ví dụ: Lợi nhuận = Doanh thu - Giá vốn) và lưu trữ chúng trong bảng Fact để tăng hiệu năng truy vấn.

Giai đoạn L - Tải (Load)

Mục tiêu: Di chuyển dữ liệu đã được biến đổi từ vùng trung chuyển vào các bảng Fact và Dimension trong kho dữ liệu đích một cách hiệu quả, an toàn.

Chiến lược tải và Tối ưu hóa:

- **Tải Ban đầu và Tăng trưởng**:
  - **Tải ban đầu**: Tải toàn bộ dữ liệu lịch sử lần đầu tiên.
  - **Tải tăng trưởng**: Chỉ tải các dữ liệu mới hoặc đã thay đổi kể từ lần tải cuối cùng, phải hoàn thành trong cửa sổ tải cho phép (thường là ban đêm).
- **Tối ưu hóa tải bảng Fact lớn**: Đối với các Bảng Fact khổng lồ (với hàng tỷ dòng giao dịch), kỹ thuật tối ưu hóa hiệu năng bao gồm: vô hiệu hóa hoặc xóa chỉ mục (indexes) trước khi bắt đầu tải, thực hiện tải dữ liệu hàng loạt, sau đó xây dựng lại chỉ mục.
- **Khả năng phục hồi**: Toàn bộ quá trình tải nên được bọc trong một giao dịch cơ sở dữ liệu duy nhất. Nếu có lỗi xảy ra, tất cả các thay đổi sẽ được quay trở lại trạng thái trước đó, đảm bảo tính toàn vẹn dữ liệu.

1.3.2 Sự chuyển dịch sang kiến trúc ELT hiện đại

Sự ra đời của các kho dữ liệu đám mây (Cloud Data Warehouses) như Google BigQuery hay Snowflake đã tạo ra sự chuyển dịch mạnh mẽ từ ETL sang ELT.

Kiến trúc ELT (Extract, Load, Transform)

Kiến trúc ELT đảo ngược thứ tự hai bước cuối cùng: Extract → Load → Transform.

1. **E (Extract)**: Tương tự như ETL, trích xuất dữ liệu từ nguồn.

2. **L (Load)**: Thay vì biến đổi trước, dữ liệu thô, kể cả dữ liệu bán cấu trúc được tải thẳng vào kho dữ liệu đám mây đích.

3. **T (Transform)**: Các phép biến đổi phức tạp được thực hiện ngay bên trong kho dữ liệu đám mây.

Động lực của ELT

Động lực chính thúc đẩy sự phổ biến của ELT là kiến trúc đột phá của các nền tảng đám mây:

- **Sức mạnh tính toán vô hạn**: Các nền tảng như BigQuery có khả năng xử lý song song khổng lồ, cho phép chạy các phép biến đổi phức tạp trên hàng tỷ dòng dữ liệu nhanh hơn nhiều so với một máy chủ ETL riêng biệt.

- **Chi phí linh hoạt**: Chi phí lưu trữ trên đám mây thấp và mô hình chi phí dựa trên nhu cầu khiến việc lưu trữ dữ liệu thô khả thi về mặt kinh tế.

- **Linh hoạt với dữ liệu thô**: ELT cho phép lưu trữ dữ liệu thô ở định dạng gốc, giúp các nhà khoa học dữ liệu dễ dàng khám phá và xây dựng mô hình.

Việc áp dụng kiến trúc ELT kết hợp với dbt cho phép xây dựng các pipeline dữ liệu mạnh mẽ, linh hoạt và dễ bảo trì hơn..

Chương 2

Điều tra, khảo sát hệ thống

2.1 Khảo sát nguồn dữ liệu

2.1.1 Phân loại dữ liệu bóng đá

Trong phân tích bóng đá hiện đại, dữ liệu thường được chia thành ba loại chính:

- **Dữ liệu sự kiện (Event Data)**: Đây là loại dữ liệu chi tiết nhất, ghi lại mọi hành động (sự kiện) diễn ra trên sân như chuyền bóng, sút, tắc bóng, rê bóng,... cùng với các thuộc tính của nó (cầu thủ thực hiện, thời gian, vị trí (x, y), kết quả, v.v).
- **Dữ liệu thống kê tổng hợp (Aggregated Stats Data)**: Đây là dữ liệu đã được xử lý và tổng hợp, thường ở cấp độ cầu thủ hoặc đội bóng theo từng trận hoặc từng mùa giải (ví dụ: tổng số bàn thắng, tổng xG, tỷ lệ chuyền bóng thành công, v.v).
- **Dữ liệu vị trí (Tracking Data)**: Dữ liệu này ghi lại vị trí tọa độ (x, y, z) của tất cả cầu thủ và quả bóng với tần suất cao. Đây là nguồn dữ liệu giàu thông tin nhất nhưng cũng phức tạp và khó thu thập nhất.

Trong phạm vi của đồ án này, cần tập trung vào hai loại dữ liệu đầu tiên.

2.1.2 Đánh giá các nguồn dữ liệu tiềm năng

Sau quá trình khảo sát, ba nguồn dữ liệu chính đã được xác định để sử dụng cho việc xây dựng kho dữ liệu:

1. **StatsBomb Open Data (GitHub)**

   - **Loại dữ liệu**: Dữ liệu sự kiện.
   - **Mô tả**: Đây là một trong những bộ dữ liệu sự kiện công khai, chi tiết và phổ biến nhất. StatsBomb cung cấp dữ liệu thô dưới dạng file JSON cho hàng nghìn trận đấu, bao gồm các giải đấu lớn như World Cup, Euro, Champions League, La Liga, v.v.
   - **Ưu điểm**:
     - **Độ chi tiết**: Cung cấp dữ liệu ở mức độ hành động, bao gồm tọa độ (x, y) trên sân, loại hành động, cầu thủ liên quan, và các ngữ cảnh chi tiết (ví dụ: một cú sút có dữ liệu về xG, vị trí thủ môn, v.v).
     - **Phù hợp để tính toán** các chỉ số nâng cao (như các chỉ số đã nêu ở Chương 1) và xây dựng các bảng Fact sự kiện chi tiết.
   - **Nhược điểm**:
     - **Độ phức tạp**: Dữ liệu có cấu trúc JSON lồng nhau, đòi hỏi quy trình trích xuất và chuyển đổi (ETL) phức tạp để làm phẳng và tải vào các bảng quan hệ trong kho dữ liệu.
     - **Phạm vi**: Dữ liệu mở chỉ gồm một số giải đấu và mùa giải nhất định.

2. **Football Players Stats 2024-2025 & 2025-2026 (Kaggle)**

   - **Loại dữ liệu**: Dữ liệu thống kê tổng hợp.
   - **Mô tả**: Các bộ dữ liệu này cung cấp các bảng dữ liệu dạng CSV chứa hàng trăm chỉ số thống kê đã được tính toán sẵn cho từng cầu thủ, thuộc các giải đấu hàng đầu châu Âu.
   - **Ưu điểm**:
     - **Tính sẵn sàng**: Dữ liệu sạch, có cấu trúc bảng, rất dễ sử dụng và tải vào kho dữ liệu.
     - **Độ bao phủ rộng**: Bao gồm nhiều chỉ số đã được tính toán cho nhiều cầu thủ và giải đấu.
     - **Cập nhật**: Các bộ dữ liệu được cập nhật cho các mùa giải mới nhất.

- **Nhược điểm**:
  - **Thiếu chi tiết**: Vì là dữ liệu tổng hợp nên không thể phân tích cách thức hoặc ngữ cảnh của các hành động.

2.1.3 Chiến lược sử dụng dữ liệu

Để xây dựng một kho dữ liệu vừa có chiều sâu phân tích, vừa có cái nhìn tổng quan, cần kết hợp cả hai nguồn dữ liệu trên:

- **StatsBomb Open Data**:
  - Được sử dụng làm nguồn dữ liệu chính để xây dựng các bảng Fact cốt lõi (ví dụ: Fact_Events, Fact_Shots).
  - Dữ liệu từ đây là cơ sở để thực hiện các quy trình Transform (T) trong ETL, nhằm tính toán các chỉ số phân tích phức tạp như $PPDA$, xG hoặc các chuỗi hành động.
- **Kaggle Datasets**:
  - Được sử dụng để xây dựng và bổ sung thông tin cho các bảng Dimension như Dim_Player, Dim_Team, Dim_Competition.
  - Ngoài ra, các bộ dữ liệu này còn được dùng để tạo ra một bảng Fact tổng hợp (ví dụ: Fact_PlayerSeasonStats), cho phép các nhà phân tích truy vấn nhanh các chỉ số tổng quan mà không cần chạy các phép toán nặng trên bảng sự kiện.

2.2 Khảo sát nhu cầu của các bên liên quan và các luồng nghiệp vụ phân tích bóng đá

2.2.1 Nhu cầu của các bên liên quan

Trong bối cảnh phân tích bóng đá chuyên nghiệp, các nhóm người dùng chính và nhu cầu đặc thù của họ được xác định như sau:

1. **Ban huấn luyện (Huấn luyện viên trưởng, Trợ lý, Giám đốc kỹ thuật, Bác sĩ)**

- **Nhu cầu về công cụ phân tích hiệu suất đội nhà khách quan và đa chiều**:
  - Đánh giá hiệu suất của từng cầu thủ sau mỗi trận đấu thông qua các chỉ số thống kê cơ bản và nâng cao (ví dụ: xG, xA, tỷ lệ chuyền bóng chính xác, số lần thu hồi bóng).
  - Xác định các chiến thuật, điểm mạnh, điểm yếu trong lối chơi chung của toàn đội (ví dụ: hiệu quả trong các pha chuyển đổi trạng thái, khả năng tận dụng tình huống cố định, các khu vực hoạt động kém hiệu quả trên sân).
  - Cần các báo cáo trực quan cho phép so sánh hiệu suất của cầu thủ và toàn đội qua nhiều trận đấu hoặc nhiều giai đoạn khác nhau của mùa giải.
- **Nhu cầu về hệ thống hỗ trợ phân tích đối thủ chuyên sâu**:
  - Nghiên cứu lối chơi, sơ đồ chiến thuật ưa thích, xu hướng tấn công/phòng ngự của đối thủ.
  - Xác định các cầu thủ then chốt, các mối đe dọa chính và các điểm yếu có thể khai thác của đối thủ.
  - Truy vấn được dữ liệu lịch sử đối đầu và hiệu suất của đối thủ khi chạm trán các đội có lối chơi tương tự.
- **Nhu cầu về dữ liệu để tối ưu hóa kế hoạch tập luyện**: Dữ liệu về thể chất và hiệu suất của cầu thủ cần được cung cấp để Ban huấn luyện có thể thiết kế, điều chỉnh các giáo án tập luyện, dinh dưỡng phù hợp, cá nhân hóa nhằm cải thiện điểm yếu và tránh quá tải. Ngoài ra còn giúp dự báo phòng tránh chấn thương, theo dõi quá trình hồi phục.

2. **Bộ phận Tuyển trạch và Quản lý thể thao (Tuyển trạch viên, Giám đốc thể thao)**
   - **Nhu cầu hỗ trợ quá trình tuyển trạch và tìm kiếm tài năng một cách khoa học**:

  - Sàng lọc cầu thủ từ một tập dữ liệu lớn (hàng nghìn cầu thủ từ nhiều giải đấu) dựa trên các tiêu chí cụ thể (ví dụ: độ tuổi, vị trí, quốc tịch, các chỉ số hiệu suất).
  - So sánh khách quan các cầu thủ tiềm năng ở cùng một vị trí để tìm ra lựa chọn tối ưu nhất.
  - Phát hiện các cầu thủ có chỉ số thống kê ấn tượng nhưng chưa được thị trường chuyển nhượng chú ý.
- **Nhu cầu về việc xây dựng hồ sơ dữ liệu đa chiều về cầu thủ**: Cho phép tạo ra một cái nhìn toàn diện về một cầu thủ, bao gồm lịch sử thi đấu, sự tiến bộ qua các mùa giải, phong cách chơi, sự phù hợp với triết lý của câu lạc bộ. Hệ thống phải cho phép thực hiện các truy vấn phức tạp.

3. **Ban Lãnh đạo Câu lạc bộ (Chủ tịch, Giám đốc điều hành)**

- **Nhu cầu về cái nhìn tổng quan, mang tính chiến lược**: Cung cấp các báo cáo cấp cao về hiệu suất tổng thể của đội bóng, sự phát triển của các tài năng trẻ, và hiệu quả hoạt động của Ban huấn luyện.
- **Nhu cầu đánh giá hiệu quả đầu tư**: Hệ thống cần cung cấp dữ liệu để đánh giá mức độ thành công của các thương vụ chuyển nhượng, so sánh giữa chi phí bỏ ra và đóng góp chuyên môn của cầu thủ.
- **Nhu cầu hỗ trợ việc ra quyết định dài hạn**: Dữ liệu từ kho phải là một nguồn tham khảo quan trọng cho các quyết định chiến lược như gia hạn hợp đồng với cầu thủ, đầu tư vào học viện đào tạo trẻ, hay định hướng phát triển chuyên môn của câu lạc bộ trong 3-5 năm tới.

4. **Các cầu thủ chuyên nghiệp**

- **Nhu cầu tự đánh giá và phát triển cá nhân**: Cầu thủ cần truy cập vào dashboard cá nhân để xem lại hiệu suất của mình sau mỗi trận đấu. Dữ liệu khách quan giúp họ nhận ra điểm mạnh cần phát huy và điểm yếu cần khắc phục.
- **Nhu cầu so sánh và đặt mục tiêu**: Hệ thống cần cho phép cầu thủ so sánh các chỉ số của mình với chính họ trong quá khứ, hoặc với những cầu

thủ hàng đầu khác ở cùng vị trí, từ đó đặt ra các mục tiêu phát triển cụ thể.

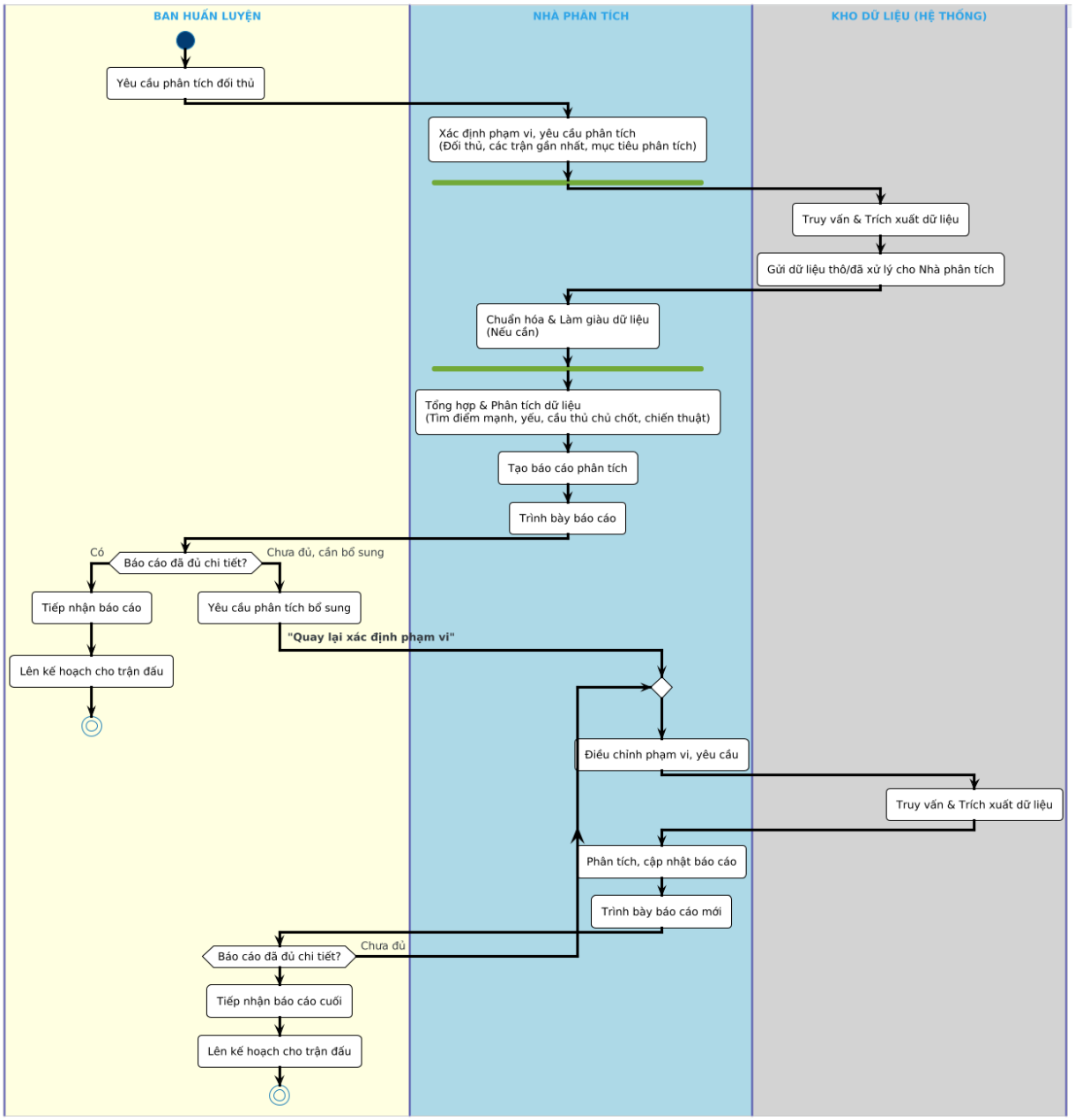
- **Nhu cầu về dữ liệu trong đàm phán hợp đồng**: Các số liệu thống kê về hiệu suất là bằng chứng thuyết phục để cầu thủ (và người đại diện) sử dụng trong các cuộc đàm phán về lương thưởng hoặc gia hạn hợp đồng.

5. **Bộ phận Truyền thông và Marketing** Bộ phận này có nhiệm vụ xây dựng hình ảnh câu lạc bộ, kết nối với người hâm mộ và tối đa hóa các cơ hội thương mại. Dữ liệu là nguồn tài nguyên quý giá để họ sáng tạo nội dung.

   - **Nhu cầu tìm kiếm các câu chuyện và thống kê thú vị**: Hệ thống cần cho phép truy vấn dễ dàng để tìm ra các cột mốc, các kỷ lục hoặc các chỉ số thống kê đặc biệt (ví dụ: "Cầu thủ X sắp có bàn thắng thứ 100 cho câu lạc bộ", "Đây là lần đầu tiên sau 5 năm đội bóng giữ sạch lưới 3 trận liên tiếp").
   - **Nhu cầu sản xuất nội dung số hấp dẫn**: Dữ liệu là nền tảng để tạo ra các sản phẩm đồ họa thông tin, video phân tích ngắn cho các nền tảng mạng xã hội, giúp tăng tương tác với cộng đồng người hâm mộ.
   - **Nhu cầu cá nhân hóa trải nghiệm người hâm mộ**: Phân tích dữ liệu về các cầu thủ được yêu thích có thể giúp bộ phận Marketing đưa ra các chiến dịch quảng bá sản phẩm (áo đấu, vật phẩm lưu niệm,...) hiệu quả hơn.

2.2.2 Luồng nghiệp vụ phân tích bóng đá

Luồng nghiệp vụ phân tích đối thủ



Hình 2.1: Luồng nghiệp vụ phân tích đối thủ

1. **Mục tiêu nghiệp vụ**: Thu thập, tổng hợp và phân tích dữ liệu về các đối thủ sắp tới nhằm xác định các đặc điểm về chiến thuật, điểm mạnh, điểm yếu và các nhân sự chủ chốt. Kết quả của nghiệp vụ này là các báo cáo chuyên sâu, phục vụ trực tiếp cho Ban huấn luyện trong việc xây dựng chiến lược và kế hoạch chuẩn bị cho trận đấu.

2. **Các bên liên quan**

   - **Ban huấn luyện**: Là người đưa ra yêu cầu phân tích và là người sử dụng cuối cùng của các sản phẩm phân tích (báo cáo, thống kê). Họ dựa vào thông tin này để ra quyết định về chiến thuật, nhân sự và phương án thi đấu.
   - **Nhà phân tích**: Là người chịu trách nhiệm chính trong việc thực thi nghiệp vụ. Họ trực tiếp làm việc với dữ liệu, sử dụng các công cụ để khai thác thông tin và chuyển hóa dữ liệu thô thành các báo cáo có ý nghĩa.
   - **Kho dữ liệu (Hệ thống)**: Đóng vai trò là nguồn cung cấp dữ liệu tập trung, chứa đựng thông tin lịch sử về các trận đấu, cầu thủ, đội bóng... Đây là nền tảng tài nguyên cho mọi hoạt động phân tích.

3. **Mô tả quy trình nghiệp vụ** Quy trình nghiệp vụ phân tích đối thủ diễn ra theo các bước tuần tự và có tính lặp lại như sau:

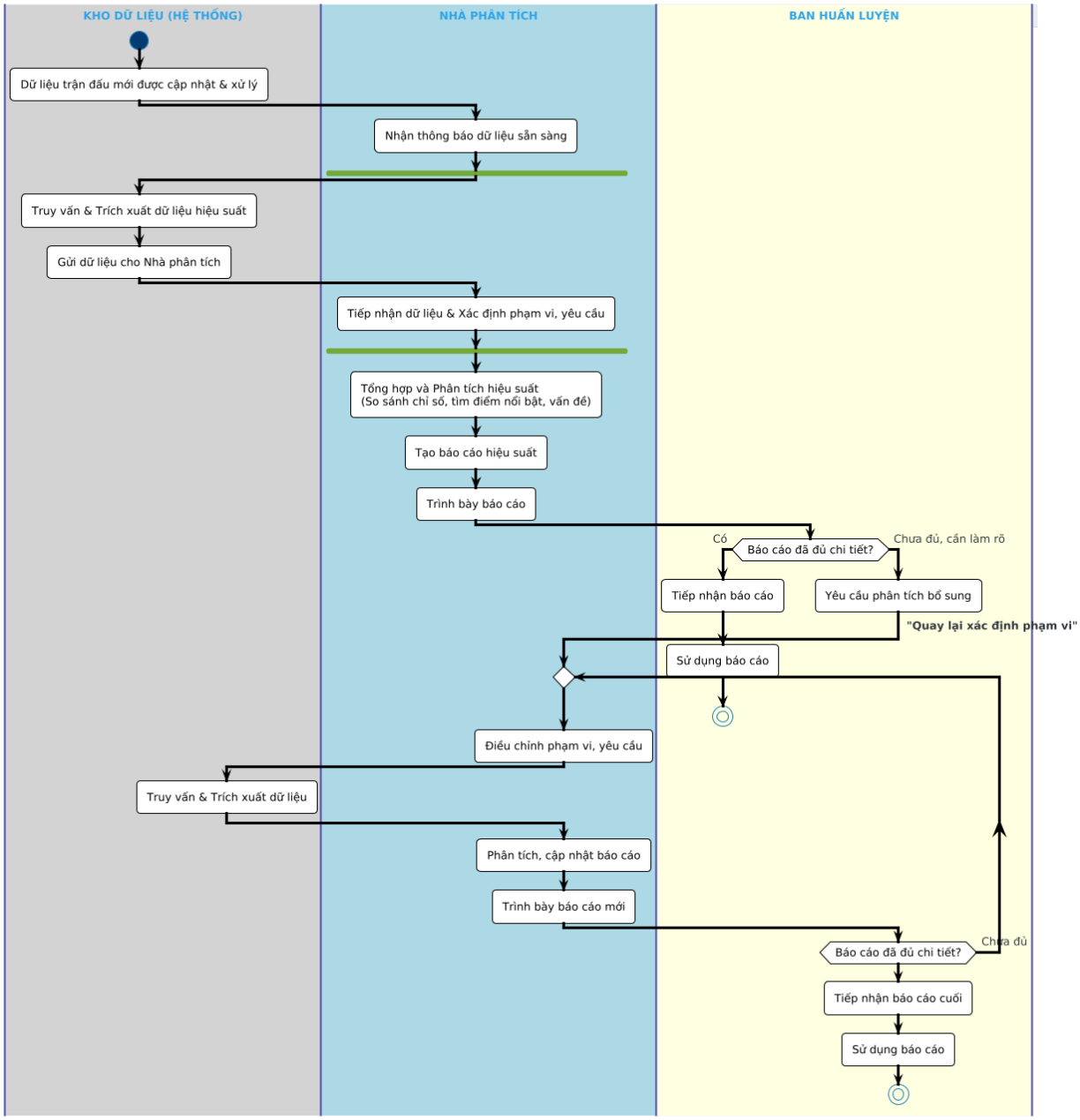
   - **Bước 1: Khởi tạo yêu cầu**: Khi có lịch thi đấu, Ban huấn luyện sẽ gửi yêu cầu cho Nhà phân tích để bắt đầu tìm hiểu về đối thủ cụ thể. Yêu cầu này khởi động toàn bộ quy trình.
   - **Bước 2: Xác định phạm vi và Thu thập dữ liệu**: Nhà phân tích làm việc với Ban huấn luyện để xác định rõ phạm vi cần phân tích (ví dụ: phân tích 5 trận gần nhất, tập trung vào các tình huống cố định). Dựa trên phạm vi đó, Nhà phân tích sẽ thực hiện truy vấn và trích xuất dữ liệu thô cần thiết từ Kho dữ liệu.
   - **Bước 3: Tổng hợp, xử lý và phân tích**: Dữ liệu thô được trích xuất sẽ được làm sạch, chuẩn hóa và tổng hợp. Nhà phân tích sử dụng các kỹ thuật thống kê để tìm ra các xu hướng, quy luật trong lối chơi, hiệu suất của cầu thủ và điểm mạnh/yếu của đối thủ.
   - **Bước 4: Tạo và trình bày Báo cáo**: Kết quả phân tích được trình bày dưới dạng một báo cáo hoàn chỉnh, bao gồm các số liệu, biểu đồ trực quan và nhận định chuyên môn, sau đó được gửi đến cho Ban huấn luyện.

- **Bước 5: Phản hồi và Hiệu chỉnh**: Ban huấn luyện xem xét báo cáo. Nếu cần thêm thông tin, họ sẽ yêu cầu bổ sung. Quy trình sẽ quay lại bước 2 hoặc 3 để Nhà phân tích thực hiện phân tích sâu hơn và cập nhật lại báo cáo.
- **Bước 6: Hoàn tất và ứng dụng**: Khi báo cáo cuối cùng được phê duyệt, các thông tin trong đó sẽ được Ban huấn luyện sử dụng để lên kế hoạch cho các buổi tập và xây dựng chiến thuật cho trận đấu. Nghiệp vụ kết thúc.

4. **Luồng dữ liệu và các thực thể chính**:

- **Dữ liệu đầu vào**:
  - **Trận đấu**: Dữ liệu về các trận đấu đã diễn ra của đối thủ (kết quả, sân nhà/sân khách, đội hình ra sân, các sự kiện chính trong trận).
  - **Cầu thủ**: Dữ liệu chi tiết về hiệu suất của từng cầu thủ đối phương (số bàn thắng, kiến tạo, số lần tắc bóng, tỷ lệ chuyền chính xác, bản đồ nhiệt hoạt động).
  - **Chiến thuật**: Dữ liệu về các sơ đồ chiến thuật đã sử dụng, các mẫu tấn công/phòng ngự phổ biến.
- **Dữ liệu đầu ra**:
  - **Báo cáo phân tích**: Sản phẩm thông tin tổng hợp, bao gồm các chỉ số hiệu suất chính, các biểu đồ, và nhận định chuyên môn.
  - **Các chỉ số tổng hợp**: Tỷ lệ kiểm soát bóng trung bình, số cú sút trung bình mỗi trận, tỷ lệ thành công trong các pha không chiến, v.v.
- **Mô tả luồng dữ liệu**: Dữ liệu được trích xuất từ Kho dữ liệu → Nhà phân tích biến đổi và xử lý thành các thông tin có ý nghĩa → Tải vào các báo cáo để Ban huấn luyện sử dụng.

Luồng nghiệp vụ phân tích hiệu suất đội nhà



Hình 2.2: Luồng nghiệp vụ phân tích hiệu suất đội nhà

1. **Mục tiêu nghiệp vụ**: Đánh giá hiệu suất thi đấu của đội nhà sau mỗi trận đấu, xác định các điểm mạnh cần phát huy và các vấn đề cần khắc phục. Kết quả phân tích cung cấp dữ liệu khách quan để Ban huấn luyện điều chỉnh chiến thuật, giáo án tập luyện và chuẩn bị cho các trận đấu trong tương lai.

2. **Các bên liên quan**

   - **Ban huấn luyện**: Là người sử dụng cuối cùng của các báo cáo hiệu suất,

đưa ra các yêu cầu phân tích chuyên sâu và áp dụng kết quả phân tích vào công tác huấn luyện.

- **Nhà phân tích**: Là người trực tiếp thực hiện quy trình phân tích, từ việc trích xuất dữ liệu, xử lý, tìm kiếm thông tin chuyên sâu và tạo ra các báo cáo trực quan.
- **Kho dữ liệu (Hệ thống)**: Tự động cập nhật, xử lý dữ liệu từ các trận đấu mới nhất và cung cấp nguồn dữ liệu hiệu suất sẵn sàng cho Nhà phân tích khai thác.

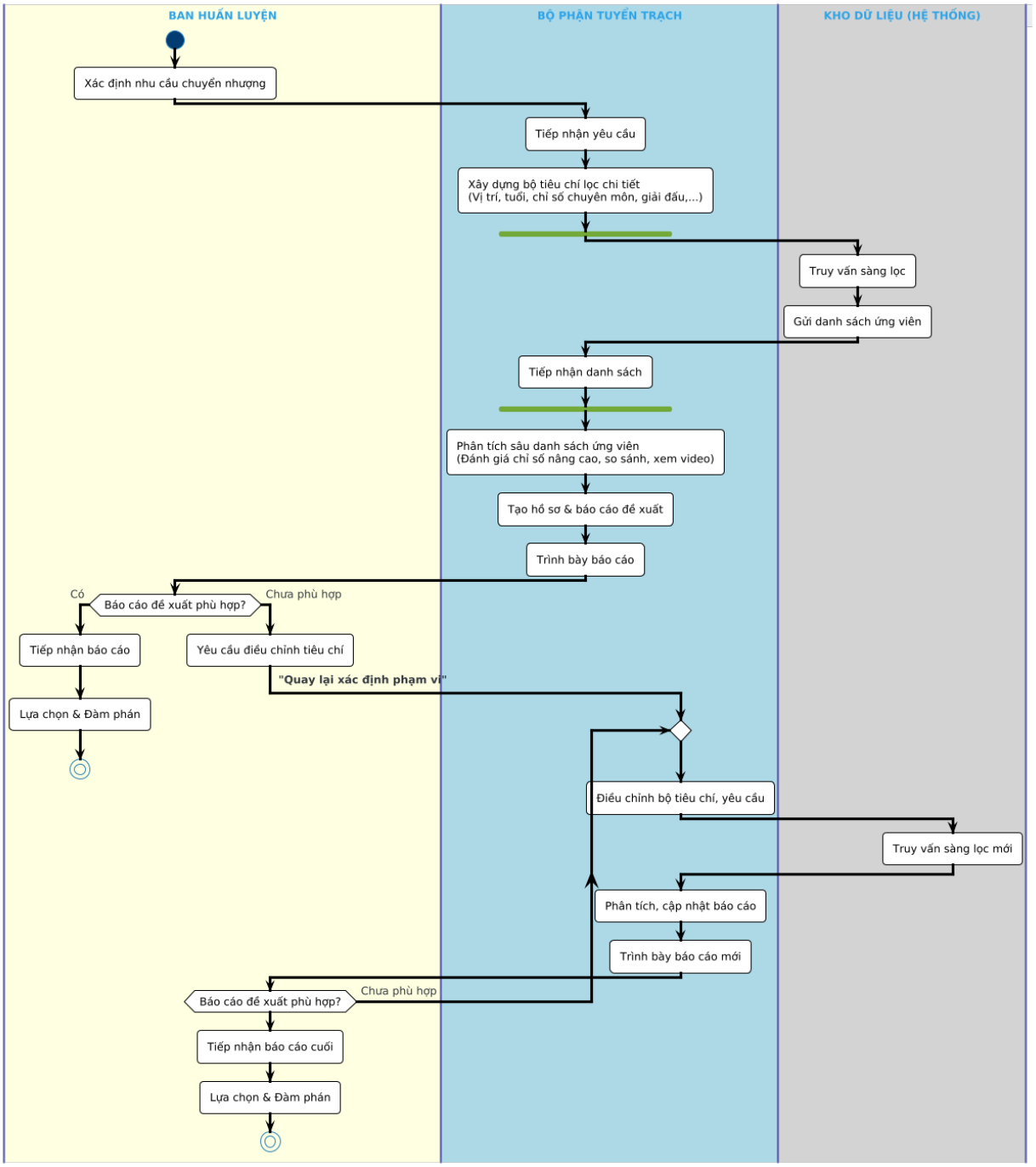
3. **Mô tả quy trình nghiệp vụ**

- **Bước 1: Cập nhật dữ liệu sau trận đấu**: Sau khi một trận đấu kết thúc, Kho dữ liệu tự động cập nhật và xử lý dữ liệu liên quan, sau đó gửi thông báo cho Nhà phân tích rằng dữ liệu đã sẵn sàng.
- **Bước 2: Khởi tạo phân tích và thu thập dữ liệu**: Nhà phân tích tiếp nhận yêu cầu, xác định phạm vi phân tích ban đầu và tiến hành truy vấn, trích xuất dữ liệu hiệu suất chi tiết của đội nhà từ Kho dữ liệu.
- **Bước 3: Phân tích và tạo báo cáo**: Nhà phân tích tổng hợp dữ liệu, thực hiện so sánh các chỉ số, tìm ra những điểm tích cực, tiêu cực và các vấn đề tồn đọng, từ đó tạo ra báo cáo hiệu suất.
- **Bước 4: Trình bày và xem xét**: Báo cáo được trình bày cho Ban huấn luyện. Ban huấn luyện sẽ đánh giá xem báo cáo đã đủ chi tiết và đáp ứng yêu cầu chuyên môn hay chưa.
- **Bước 5: Phản hồi và hiệu chỉnh**: Nếu báo cáo chưa đạt, Ban huấn luyện sẽ yêu cầu phân tích bổ sung, làm rõ các vấn đề cụ thể. Nhà phân tích sẽ điều chỉnh phạm vi, tiếp tục khai thác dữ liệu và cập nhật lại báo cáo. Quá trình này có thể lặp lại cho đến khi báo cáo đáp ứng đầy đủ yêu cầu.
- **Bước 6: Hoàn tất và ứng dụng**: Khi báo cáo cuối cùng được phê duyệt, Ban huấn luyện sẽ tiếp nhận và sử dụng báo cáo để phục vụ cho công tác chuyên môn. Nghiệp vụ kết thúc.

4. **Xác định luồng dữ liệu và các thực thể chính**:

- **Dữ liệu đầu vào**:
  - **Trận đấu**: Dữ liệu về trận đấu vừa diễn ra của đội nhà (tỷ số, đội hình, sơ đồ chiến thuật, các sự kiện chính).
  - **Cầu thủ đội nhà**: Dữ liệu hiệu suất chi tiết của từng cá nhân (số phút thi đấu, bàn thắng, kiến tạo, tỷ lệ chuyền bóng, số lần thu hồi bóng, các chỉ số thể chất như quãng đường di chuyển).
  - **Sự kiện trong trận**: Dữ liệu chi tiết đến từng hành động trên sân (tọa độ các cú sút, đường chuyền, các pha tắc bóng,...) được thu thập từ các hệ thống theo dõi.
- **Dữ liệu đầu ra**:
  - **Báo cáo hiệu suất**: Sản phẩm thông tin tổng hợp, bao gồm các biểu đồ, bản đồ nhiệt, video minh họa và nhận định từ Nhà phân tích.
  - **Các chỉ số hiệu suất**: Các chỉ số được tính toán để đánh giá hiệu quả của đội bóng.
- **Mô tả luồng dữ liệu**: Dữ liệu được trích xuất từ Kho dữ liệu → Nhà phân tích biến đổi và xử lý thành các thông tin có ý nghĩa → Tải vào các báo cáo để Ban huấn luyện sử dụng.

Luồng nghiệp vụ tuyển trạch cầu thủ



Hình 2.3: Luồng nghiệp vụ tuyển trạch cầu thủ

1. **Mục tiêu nghiệp vụ**: Tìm kiếm và xác định các cầu thủ tiềm năng phù hợp với chiến thuật và định hướng phát triển của đội bóng. Cung cấp cho Ban huấn luyện một danh sách các ứng viên chất lượng đã qua sàng lọc, làm cơ sở cho các quyết định chuyển nhượng.

2. **Các bên liên quan**

- **Ban huấn luyện**: Là người xác định nhu cầu chuyển nhượng ban đầu (vị trí cần bổ sung, phong cách cầu thủ) và là người ra quyết định cuối cùng trong việc lựa chọn và đàm phán.
- **Bộ phận tuyển trạch**: Là đơn vị thực thi chính, chịu trách nhiệm xây dựng tiêu chí chi tiết, phân tích dữ liệu, đánh giá chuyên môn và tạo báo cáo đề xuất các ứng viên tiềm năng.
- **Kho dữ liệu (Hệ thống)**: Là công cụ cốt lõi, cung cấp một cơ sở dữ liệu lớn về cầu thủ và các công cụ để sàng lọc, truy vấn dữ liệu theo các tiêu chí phức tạp do bộ phận tuyển trạch xây dựng.

3. **Mô tả quy trình nghiệp vụ**

- **Bước 1: Xác định nhu cầu**: Ban huấn luyện xác định nhu cầu nhân sự và gửi yêu cầu đến Bộ phận tuyển trạch.
- **Bước 2: Xây dựng tiêu chí và Sàng lọc lần đầu**: Bộ phận tuyển trạch tiếp nhận yêu cầu và cụ thể hóa thành một bộ tiêu chí lọc chi tiết (ví dụ: vị trí, độ tuổi, chỉ số chuyên môn, giải đấu đang thi đấu,...). Dựa trên bộ tiêu chí này, họ thực hiện truy vấn trên Kho dữ liệu để có được danh sách ứng viên sơ bộ.
- **Bước 3: Phân tích chuyên sâu và Tạo báo cáo đề xuất**: Bộ phận tuyển trạch tiến hành phân tích sâu danh sách ứng viên, bao gồm việc đánh giá các chỉ số nâng cao, so sánh giữa các cầu thủ, xem lại video thi đấu để tạo ra một hồ sơ và báo cáo đề xuất ban đầu.
- **Bước 4: Trình bày và Xem xét**: Báo cáo được trình bày cho Ban huấn luyện để đánh giá mức độ phù hợp của các ứng viên được đề xuất.
- **Bước 5: Phản hồi và Hiệu chỉnh**: Nếu Ban huấn luyện cho rằng báo cáo chưa phù hợp hoặc cần tìm kiếm thêm các phương án khác, họ sẽ yêu cầu điều chỉnh lại các tiêu chí. Bộ phận tuyển trạch sẽ cập nhật lại bộ lọc, thực hiện truy vấn mới và lặp lại quá trình phân tích để tổng hợp lại báo cáo mới.

- **Bước 6: Hoàn tất và Lựa chọn**: Khi báo cáo cuối cùng đã đáp ứng được yêu cầu, Ban huấn luyện sẽ tiếp nhận, lựa chọn ứng viên phù hợp và bắt đầu quá trình đàm phán chuyển nhượng. Nghiệp vụ kết thúc.

4. **Luồng dữ liệu và các thực thể chính**:

   - **Dữ liệu đầu vào**:
     - **Cầu thủ**: Dữ liệu về cầu thủ từ nhiều giải đấu, bao gồm thông tin cá nhân (tuổi, quốc tịch, chiều cao), thông tin hợp đồng (câu lạc bộ chủ quản, thời hạn), và bộ chỉ số hiệu suất chi tiết theo từng vị trí.
     - **Giải đấu**: Dữ liệu về mức độ cạnh tranh, phong cách chơi của các giải đấu khác nhau để đánh giá khả năng hòa nhập của cầu thủ.
     - **Thị trường chuyển nhượng**: Dữ liệu về định giá cầu thủ, lịch sử chuyển nhượng.
   - **Dữ liệu đầu ra**:
     - **Báo cáo tuyển trạch**: Thông tin tổng hợp về một hoặc nhiều cầu thủ, phân tích điểm mạnh/yếu dựa trên dữ liệu, so sánh với các cầu thủ hiện tại của đội, nhận định chuyên môn của tuyển trạch viên.
     - **Danh sách rút gọn**: Một danh sách các ứng viên tiềm năng nhất đã được xếp hạng theo mức độ ưu tiên.
   - **Mô tả luồng dữ liệu**: Dữ liệu được trích xuất từ Kho dữ liệu → Bộ phận tuyển trạch biến đổi và xử lý thành các thông tin có ý nghĩa → Tải vào các báo cáo để Ban huấn luyện sử dụng.

2.3 Đặc tả yêu cầu hệ thống

Dựa trên việc khảo sát và phân tích các luồng nghiệp vụ cốt lõi, các yêu cầu chức năng, phi chức năng và các ràng buộc thiết kế của hệ thống kho dữ liệu được xác định như sau:

2.3.1 Yêu cầu chức năng

Hệ thống phải cung cấp đầy đủ các chức năng để hỗ trợ toàn bộ vòng đời của dữ liệu, từ thu thập, xử lý cho đến khai thác.

- **Thu thập dữ liệu**:
  - Hệ thống phải có khả năng thu thập dữ liệu tự động từ nhiều nguồn khác nhau (ví dụ: API từ các nhà cung cấp dữ liệu thể thao, file CSV/JSON, hoặc kết quả từ web scraping).
  - Hệ thống phải hỗ trợ thu thập dữ liệu theo lịch trình định sẵn (ví dụ: hàng ngày, sau mỗi vòng đấu) hoặc theo sự kiện.
- **Xử lý và Tích hợp dữ liệu (Data Processing & Integration)**:
  - Hệ thống phải có khả năng thực hiện các quy trình ETL/ELT để làm sạch, chuẩn hóa, và biến đổi dữ liệu thô. Các tác vụ bao gồm: xử lý giá trị thiếu, đồng bộ hóa định dạng (ngày tháng, tên cầu thủ), và tạo ra các trường dữ liệu mới (ví dụ: tính toán các chỉ số phát sinh).
  - Hệ thống phải có khả năng tích hợp dữ liệu từ các nguồn khác nhau để tạo ra một bộ dữ liệu thống nhất và toàn diện.
- **Lưu trữ dữ liệu (Data Storage)**:
  - Hệ thống phải cung cấp một lớp lưu trữ cho dữ liệu thô và dữ liệu trung gian (Data Lake), cho phép lưu trữ linh hoạt nhiều định dạng dữ liệu khác nhau.
  - Hệ thống phải cung cấp một lớp lưu trữ cho dữ liệu đã qua xử lý, có cấu trúc và được tối ưu hóa cho việc truy vấn phân tích (Data Warehouse).
- **Điều phối và Tự động hóa luồng dữ liệu (Pipeline Orchestration)**:
  - Hệ thống phải cho phép định nghĩa, lập lịch và tự động hóa các luồng xử lý dữ liệu.
  - Hệ thống phải cung cấp giao diện để theo dõi trạng thái (thành công, thất bại), quản lý và gỡ lỗi các luồng dữ liệu.

- **Cung cấp khả năng truy vấn và phân tích (Data Serving & Analysis)**:
  - Hệ thống phải cho phép các công cụ phân tích (BI tools) kết nối và truy vấn dữ liệu từ kho dữ liệu một cách hiệu quả.
  - Hệ thống phải hỗ trợ các truy vấn phức tạp, tổng hợp dữ liệu từ nhiều chiều khác nhau để phục vụ cho các nghiệp vụ phân tích đối thủ, hiệu suất đội nhà và tuyển trạch.
- **Trực quan hóa dữ liệu và báo cáo (Visualization & Reporting)**:
  - Hệ thống phải hỗ trợ việc xây dựng các báo cáo và dashboard tương tác.
  - Cung cấp các dashboard chuyên biệt cho từng nghiệp vụ:
    * Dashboard phân tích đối thủ (so sánh chỉ số, sơ đồ chiến thuật).
    * Dashboard phân tích hiệu suất đội nhà (đánh giá sau trận đấu, theo dõi hiệu suất cá nhân).
    * Dashboard tuyển trạch (sàng lọc, so sánh và đánh giá cầu thủ).

2.3.2 Yêu cầu phi chức năng

Các yêu cầu phi chức năng xác định các tiêu chí về chất lượng và hiệu quả hoạt động của hệ thống.

- **Hiệu năng**:
  - **Độ trễ dữ liệu**: Dữ liệu của một trận đấu phải được cập nhật sớm và sẵn sàng để phân tích sau khi trận đấu kết thúc.
  - **Thời gian phản hồi truy vấn**: Các truy vấn trên dashboard PowerBI cho các báo cáo tiêu chuẩn phải có thời gian phản hồi nhanh.
- **Tính sẵn sàng**: Hệ thống phải đảm bảo độ sẵn sàng cao, đặc biệt là trong các giai đoạn cao điểm như kỳ chuyển nhượng hoặc các vòng đấu quan trọng.
- **Tính mở rộng**:
  - Hệ thống phải có khả năng mở rộng theo chiều ngang để xử lý khối lượng dữ liệu ngày càng tăng (thêm mùa giải mới, thêm các giải đấu mới).

  - Kiến trúc hệ thống phải cho phép bổ sung các nguồn dữ liệu mới hoặc các luồng xử lý mới mà không ảnh hưởng lớn đến các thành phần hiện có.

- **Tính tin cậy và Toàn vẹn**:
  - Các luồng dữ liệu phải có cơ chế xử lý lỗi và tự động thử lại khi có sự cố tạm thời.
  - Hệ thống phải đảm bảo tính nhất quán và toàn vẹn của dữ liệu trong kho dữ liệu thông qua các ràng buộc và quy tắc kiểm tra chất lượng dữ liệu.

- **Tính bảo mật**: Dữ liệu trong hệ thống phải được phân quyền truy cập. Chỉ những người dùng có vai trò phù hợp (Nhà phân tích, Ban huấn luyện) mới có quyền truy cập vào các dữ liệu và báo cáo tương ứng.

- **NFR6: Tính dễ bảo trì**: Toàn bộ hệ thống phải được đóng gói thành các container để dễ dàng triển khai, nâng cấp và bảo trì một cách độc lập.

2.3.3 Ràng buộc thiết kế

- **Công nghệ sử dụng**:
  - **Docker**: Được sử dụng để container hóa toàn bộ các thành phần của hệ thống, đảm bảo tính nhất quán giữa các môi trường.
  - **Apache Airflow**: Được sử dụng làm công cụ điều phối, lập lịch và giám sát các luồng xử lý dữ liệu.
  - **MinIO**: Được sử dụng làm Data Lake (lưu trữ đối tượng) để chứa dữ liệu thô và dữ liệu trung gian.
  - **Apache Spark**: Được sử dụng làm công cụ xử lý dữ liệu phân tán, chịu trách nhiệm cho các tác vụ biến đổi dữ liệu phức tạp và quy mô lớn.
  - **PostgreSQL**: Được sử dụng làm Data Warehouse, lưu trữ dữ liệu có cấu trúc đã được làm sạch và sẵn sàng cho việc truy vấn phân tích.

  - **Microsoft Power BI**: Được sử dụng làm công cụ BI để kết nối tới PostgreSQL, xây dựng các mô hình dữ liệu và tạo các báo cáo, dashboard trực quan.

- **Nguồn dữ liệu**: Hệ thống sẽ tập trung vào các nguồn dữ liệu có cấu trúc và bán cấu trúc từ các API công khai hoặc các tệp dữ liệu được cung cấp.
- **Môi trường triển khai**: Hệ thống được thiết kế để có thể triển khai trên hạ tầng máy chủ cục bộ hoặc trên một nền tảng đám mây riêng.

Chương 3

Phân tích & thiết kế hệ thống

Chương 4

Cài đặt hệ thống

Kết luận

Tài liệu tham khảo

[1] Wikipedia. *Bóng đá*. `https://vi.wikipedia.org/wiki/Football`. 2025.

[2] TS. Phạm Huyền Linh. *Bài giảng Phân tích và thiết kế hệ thống*. 2025.

[3] TS. Lê Hải Hà. *Bài giảng Phân tích và thiết kế hệ thống*. 2024.

[4] ThS. Nguyễn Danh Tú. *Giáo trình Kho dữ liệu và Kinh doanh thông minh*. 2025.