Data Feed Specification

Nhóm 1

Môn DataWareHouse

Giảng Viên Hướng Dẫn: Nguyễn Đức Công Song

Mục Lục

1.0 Overview

1.1 Data Feed Specification

- Nguồn dữ liệu:

https://www.premierleague.com

Phương pháp lấy dữ liệu: script, nhập tay

Tài liệu này cung cấp mô tả chi tiết và cập nhật kết quả giải Premierleagua của Bóng Đá Ngoại Hạng Anh.

Phần đầu tiên của tài liệu cung cấp định nghĩa về dữ liệu được đưa vào nguồn cấp dữ liệu và các thông số kỹ thuật thuộc tính. Đặc tả thuộc tính bao gồm định nghĩa, định dạng và quy tắc xác thực

1.2 Contact Information

Thông tin các thành viên

MSSV	Tên	Liên Hệ	Chức vụ
19130091	Đặng Thái Kế	19130091@st.hcmuaf.edu.vn	admin
19130105	Lê Đăng Khoa	19130105@st.hcmuaf.edu.vn	admin
19130120	Cao Huy Tấn Lộc	19130120@st.hcmuaf.edu.vn	admin

1.3 References

Document	Author
https://www.premierleague.com	premierleague.com
https://www.premierleague.com/results	premierleague.com/results

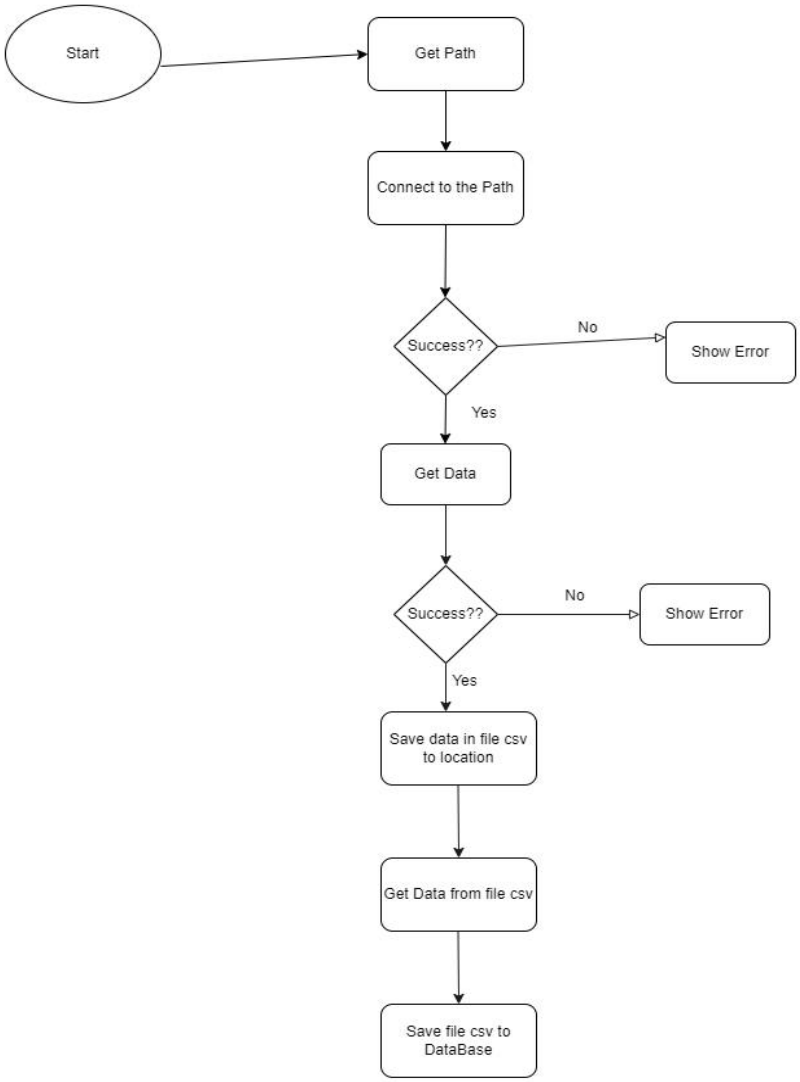
2.0 Transaction Definition

2.1 Overview

2.2 Data Feed Process Flow

Sơ đồ tiến trình công việc được thực hiện bằng phần mềm Diagrams.net

Sơ đồ tiền trình công việc :



2.3 Validation Guidelines

Đối với các trường bắt buộc, hai xác nhận sau sẽ được thực hiện.

2.4 Attribute Specification

Phần này cung cấp các định nghĩa ngắn gọn về từng thuộc tính, bao gồm định dạng trường và tiêu chí chỉnh sửa. Để biết thêm thông tin về các thuật ngữ được sử dụng để định nghĩa thuộc tính.

2.4.1 Gold access logs (wap-site_access_log.YYYY-MM-DD_appN-servername.tx t)

Attribute Specification

Field Name	Format	Mask	Edit Rules	Description	Example
LeagueName	Text	League		Tên giải đấu	Premier League
HomeTeam	Text	Team		Đội chủ nhà	Manchester Utd
AwayTeam	Text	Team		Đội khách	Liverpool
TimeStart	Time	Time		Thời gian diễn ra trận đấu	22:30
Data	Date	Date		Ngày diễn ra trận đấu	05.11.2022
Gwinner	Interger	Goal		Tỉ số đội nhà ghi bàn	1
GLoser	Interger	Goal		Tỉ số đội khách ghi bàn	3
Refreree	Text	Person		Trọng tài bắt chính	Tierney P. (Eng)
Venue	Text	Location		Sân	Old Tranfford
Attendance	Interger	Attendance		Số lượng khách	74000
Round	Text	Round		Vòng đấu	4
Status	Text	Status		Trạng thái trận đấu	Hoàn thành

Catalog

2.5 Code Values

Giá trị mã là chữ viết tắt của các giá trị tiêu chuẩn được sử dụng cho một thuộc tính. Chúng chỉ được sử dụng bất cứ khi nào một thuộc tính có miền giá trị đã biết. Nó cung cấp một kỹ thuật hiệu quả hơn để lưu trữ các tập hợp lớn các giá trị lặp lại. Chúng được sử dụng cho mọi thứ, từ các loại giao dịch và trạng thái đến các bộ giá trị lớn như tên nhà sản xuất

Đối với bất kỳ thuộc tính nào được xác định là LOV (Danh sách các giá trị) do nhà cung cấp cung cấp, danh sách đầy đủ các giá trị sẽ được cung cấp cho danh sách đó cùng với các mô tả được liên kết với giá trị mã đó.

Việc duy trì các giá trị mã có thể là một vấn đề - cả hai bên phải duy trì đồng bộ. Điều này được hỗ trợ bởi các thủ tục hoạt động và lập phiên bản nguồn cấp dữ liệu (bất kỳ thay đổi nào đối với các đặc tính của nguồn cấp dữ liệu sẽ làm tăng thuộc tính phiên bản nguồn cấp dữ liệu). List of Values (LOV)

2.6 Data Source Extraction and Exception Handling Procedures

2.6.1 Special Extract Procedure: Source => Import

Chọn các trường được liệt kê từ nhật ký máy chủ ứng dụng nguồn và tải vào Stagging. stg_wap_site_access_log là một kho lưu trữ cho tất cả các bản ghi và được phân vùng theo ngày. Trong Stagging, dữ liệu được tổng hợp cho báo cáo hàng tháng và được lưu giữ trong ba tháng.

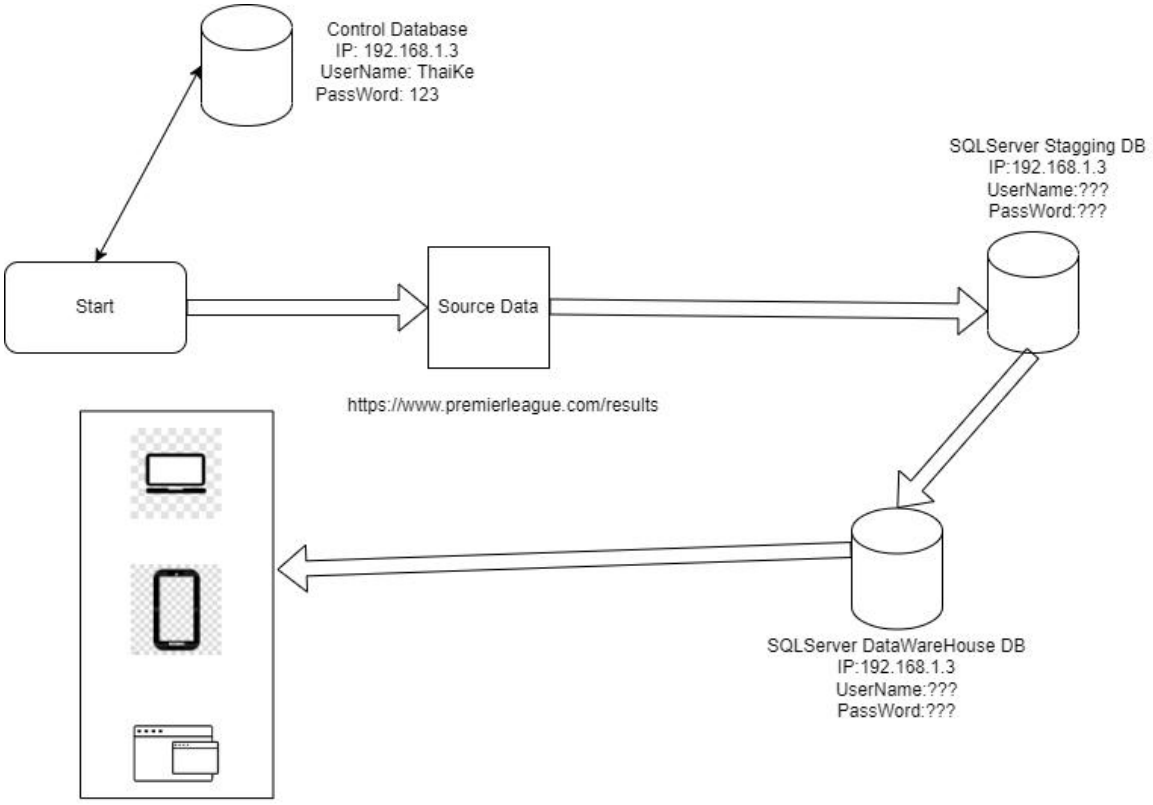
2.6.2 Error Code

Độ phân giải 1: Không đi qua

2.6.3 Exception Handling Procedure (Not Applicable)

2.6.4 Special Design Consideration (Not Applicable)

3.0 Feed Architecture



3.1 Overview

Các thành phần chính của kiến trúc nguồn cấp dữ liệu là:

Transport Mechanism:

Dữ liệu sẽ truyền trực tiếp từ tệp nhật ký truy cập của máy chủ ứng dụng trong ~ home / dataWarehouse_dd/mm/yy .csv đến STAGGING sử dụng SQLSERVER bằng cách gọi các câu lệnh truy vấn sql.

Feed Characteristics. Tệp nhật ký: wap-site_access_log.YYYY-MM-DD_app {N} - servername.txt.gz

Tần suất: Hàng tuần

Nội dung: Chứa dữ liệu từ ngày tuần trước

Kích thước bảng ước tính: ~ 500MB / tuần

Tiền xử lý: giải nén tệp zip (kích thước tệp được giải nén: ~ 1,4GB) Yêu cầu kích thước hàng tháng ước tính: 130 * 30 + 1400 = 5300MB / tháng

SOURCE =>> IMPORT(nguồn cung dữ liệu) Các thủ tục được sử dụng cho ba nhóm thành viên vận hành để quản lý các quy trình nguồn cấp dữ liệu. Các thủ tục thường sẽ xác định các mẫu cho các thủ tục xử lý, các thủ tục báo cáo cho các vấn đề và dữ liệu liên hệ được yêu cầu.

3.2 Transport Mechanism (cơ chế vận chuyển)

Cơ sở dữ liệu được thể hiện dưới dạng một bảng duy nhất mà tất cả các hồ sơ được lưu trữ như hàng duy nhất của dữ liệu, được phân cách bởi dấu phẩy

3.3 Feed Characteristics File Format:

Các định dạng tệp được chấp nhận hiện tại là ASCII CSV (các giá trị được phân tách bằng dấu phẩy) và phân cách bằng dấu sổ đứng ASCII. Lưu ý rằng các tệp trong "DOS" (dòng kết thúc bằng ký tự xuống dòng & dòng cấp dữ liệu) hoặc "Unix" (kết thúc bởi dòng cấp dữ liệu) đều được chấp nhận miễn là nguồn cấp dữ liệu nhất quán là một hoặc khác. Cũng lưu ý rằng định dạng tệp không được thay đổi trong quá trình chuyển để tránh lỗi tổng kiểm tra. Mặc dù không được khuyến khích, nhưng nếu tệp dữ liệu rất lớn, chúng có thể bị nén. Các định dạng nén được chấp nhận là z (nén), gzip, zip và jar. Khi tệp được nén, phần mở rộng tiêu chuẩn cho phần mềm nén sẽ được sử dụng và một mục nhập nén được yêu cầu trong tệp điều khiển.

Data Size: Khoảng: ~ 500MB / tuần

Data Location:

~ home / dataWarehouse / logs - trong DW1

Data Frequency:(tần suất dữ liệu) hằng tuần

Delivery Location: Quá trình này là tệp phẳng được chuyển dữ liệu cơ sở dữ liệu từ nhật ký máy chủ ứng dụng sang STAGING schema: stg_wap_site_access_log table

No.	Field Name	Format	Length	Mask	Edit Rules	Description(sample, value)
1.	LeagueName	text				Premier League
2.	HomeTeam	text				Man Utd
3.	AwayTeam	text				Liverpool
4.	TimeStart	time				22:30
5.	Date	data				05.11.2022
6.	Gwinner	int				1
7.	GLoser	int				3
8.	Refreree	text				Tierney P. (Eng)
9.	Venue	text				Old Trafford
10.	Attendance	int				74000
11.	Round	text				4
12.	Status	text				Finished

Naming Convention: :(Quy ước đặt tên) Quy ước đặt tên cho nguồn cấp dữ liệu

wap-site_access_log.YYYY-MM-DD_app{N}-username.txt.gz

Ở đâu:

YYYY-MM-DD là ngày của tệp.

ứng dụng {N} là tên ứng dụng: giải đấu Premier League

Tên máy chủ: https://www.premierleague.com

Delivery Schedule: : (lịch trình truyền tải data) Hàng tuần

3.4 Data/Process Controls

Kiểm soát quy trình và dữ liệu được thực hiện để đảm bảo rằng nguồn cấp dữ liệu là chính xác, đầy đủ và kịp thời.

Các điều khiển chính là

Control	Description	Implementation
ER	Trích xuất sẵn sàng	Tệp trích xuất công việc
AR	Trạng thái transform thành công	Tệp trích xuất công việc
SC	Hoàn tất quá trình load vào datwarehouse thành công	Tệp trích xuất công việc
ERR	Trích xuất không thành công	Tệp trích xuất công việc

3.5 Checksum Definition (Not Applicable)

Vì dữ liệu được thu thập thông qua truy cập trực tiếp vào hệ thống, định nghĩa tổng kiểm tra không áp dụng cho nguồn cấp dữ liệu này.

3.6 Control File Format (Not Applicable)

Vì dữ liệu được thu thập thông qua quyền truy cập trực tiếp vào hệ thống, nên các tệp điều khiển không thể áp dụng cho nguồn cấp dữ liệu này.

3.7 FTP Setup (Not Applicable)

Vì dữ liệu được thu thập thông qua quyền truy cập trực tiếp vào hệ thống nên Thiết lập FTP không thể áp dụng cho nguồn cấp dữ liệu này.

3.8 Operations Interface (Not Applicable)

Định nghĩa về các thủ tục hoạt động để quản lý nguồn cấp dữ liệu, bao gồm cả việc xử lý ngoại lệ và báo cáo vấn đề.

Định nghĩa quy trình này sẽ xác định rõ ràng các hành động được thực hiện bởi từng nhóm nhân viên vận hành đối với các thành phần thủ công của quy trình.

Nó cũng sẽ xác định các yêu cầu giám sát đối với cả hai nhóm nhân viên.

Phần giải quyết vấn đề sẽ xác định tất cả các điều kiện lỗi có thể xảy ra và cung cấp một quá trình hành động thích hợp.

Phần báo cáo vấn đề xác định các tình huống trong đó vấn đề phải được báo cáo và các hành động cần thực hiện để leo thang.

Một danh sách liên hệ sẽ được phát triển cho mỗi tổ chức. Điều này sẽ cung cấp tên, số điện thoại và lĩnh vực phụ trách

4.0 SLA Negotiation (Not Applicable)

Vì dữ liệu được thu thập thông qua truy cập trực tiếp từ hệ thống nội bộ, SLA không được áp dụng cho nguồn cấp dữ liệu này.

Appendix A – Attribute Specification

Các giá trị sau được sử dụng để định nghĩa định dạng thuộc tính

Attribute Format	Description	Example
date	Dữ liệu ngày tháng. Độ dài và định dạng sẽ phụ thuộc vào mặt nạ được sử dụng. Mặt nạ mặc định là yyyy-MM-dd_HH-mm-ss.	28-12-2022_10-20-39
load	values tại cột "file_name" trong bảng config	"resultScriptToBongDa"
paths	values của cột "paths" tại bảng file_log	D:\file_csv
paths_dim	đường dẫn của file date_dim	D:\Date_dim
ERROR.txt	tên file chứa lỗi file lỗi nếu chạy không thành công	"result_football\\Error.txt"
sql	câu lệnh query update trạng thái tại bảng file_log	"update file_log set log_status =? where paths =?;"
pathFile.csv	đường dẫn nơi lưu file csv chứa dữ liệu lấy từ website về local	D:\file_csv là đường dẫn lưu file
PathFileFileExcel	đường dẫn nơi lưu file excel chứa dữ liệu lấy từ website về local	D:\file_excel là đường dẫn lưu file

Để định nghĩa cho thuộc tính ngày tháng, các giá trị sau được sử dụng:

Mask Character	Description	Example
MM	Month in numeric format	01 is January
DD	Ngày	23 là ngày thứ 23 trong tháng
YY	Năm không bao gồm thế kỷ	02

Mask Character	Description	Example
HH	Giờ ở định dạng 24 giờ	23 là 11 giờ tối
SS	Giây sau phút	05 là 05 giây
/	Dấu phân cách giá trị	01/02/02
-	Dấu phân cách giá trị	01-02-02

Để định nghĩa cho thuộc tính chuỗi, các giá trị sau được sử dụng:

Mask Character	Description	Example
NONE	Không có sở thích	Mr Jones home address
UPPER	Tất cả chữ hoa	MR JONES HOME ADDRESS
LOWER	Tất cả chữ thường	mr jones home address
INITCAP	Tất cả các từ đều được viết hoa	Mr Jones Home Address