**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NÔNG LÂM TPHCM**

**KHOA CÔNG NGHỆ THÔNG TIN**



**Môn: Data Warehouse**

**GVHD: Th.S Nguyễn Đức Công Song**

**Đề tài: Hệ Thống Báo Giá Xe Đạp**

**Sinh viên thực hiện:**

Cao Thành Nam - 21130448

**Mục lục**

[**Lời mở đầu** 2](#_Toc184670586)

[**LỜI CÁM ƠN** 3](#_Toc184670587)

[**NHẬN XÉT** 4](#_Toc184670588)

[**Revision History** 5](#_Toc184670589)

[**I.** **Datafeed Specification** 5](#_Toc184670590)

[**II.** **Cơ sở dữ liệu** 6](#_Toc184670591)

[**1.** **control** 6](#_Toc184670592)

[**1.1.** **config** 6](#_Toc184670593)

[**1.2.** **logs** 7](#_Toc184670594)

[**1.3.** **status** 7](#_Toc184670595)

[**1.4.** **dateDim** 7](#_Toc184670596)

[**1.5.** **monthDim** 9](#_Toc184670597)

[2. **staging** 9](#_Toc184670598)

[**2.1.** **bikes** 10](#_Toc184670599)

[**III.** **Workflow** 11](#_Toc184670600)

[**1.** **Lấy dữ liệu từ web, lưu xuống database và tạo ra file** 11](#_Toc184670601)

[**IV.** **Exception handling** 12](#_Toc184670602)

[**1.** **Lỗi không truy cập được trang web** 12](#_Toc184670603)

[**2.** **Lỗi không lấy được mô tả** 14](#_Toc184670604)

[**3.** **Lỗi số lượng sản phẩm bằng 0** 14](#_Toc184670605)

[**4.** **Lỗi không thể lưu vào CSDL** 14](#_Toc184670606)

# **Lời mở đầu**

Trong thời đại công nghệ số ngày nay, việc quản lý và phân tích dữ liệu đóng vai trò quan trọng trong việc ra quyết định chiến lược cho các doanh nghiệp. Đặc biệt trong ngành bán lẻ, dữ liệu về sản phẩm, giá cả, và hành vi người tiêu dùng trở thành nguồn tài nguyên quý giá giúp nâng cao hiệu quả kinh doanh. Đề tài “Báo cáo giá xe đạp” ra đời nhằm giải quyết nhu cầu lưu trữ, quản lý, và phân tích dữ liệu giá cả từ các nguồn khác nhau, hỗ trợ doanh nghiệp trong việc cung cấp thông tin chính xác và kịp thời. Kho dữ liệu (Data Warehouse) không chỉ giúp lưu trữ dữ liệu từ nhiều nguồn mà còn cung cấp các công cụ phân tích mạnh mẽ để dự đoán xu hướng giá, quản lý hàng tồn kho, và tối ưu hóa chiến lược kinh doanh. Bằng cách tổng hợp dữ liệu giá từ các website bán xe đạp, hệ thống sẽ cho phép doanh nghiệp theo dõi biến động giá cả, so sánh giữa các thương hiệu, và phát hiện những cơ hội kinh doanh tiềm năng. Báo cáo này sẽ trình bày quá trình xây dựng hệ thống Data Warehouse cho dữ liệu giá xe đạp, bao gồm các bước thu thập dữ liệu, thiết kế hệ thống kho dữ liệu, và ứng dụng phân tích nhằm hỗ trợ doanh nghiệp ra quyết định dựa trên dữ liệu. Thông qua nghiên cứu này, chúng tôi kỳ vọng sẽ giúp các doanh nghiệp bán lẻ xe đạp không chỉ nâng cao năng lực quản lý dữ liệu mà còn phát triển chiến lược cạnh tranh hiệu quả trong thị trường.

**LỜI CÁM ƠN**

Chúng em xin chân thành cảm ơn Thầy đã tận tình hướng dẫn và truyền đạt những kiến thức quý báu trong suốt khóa học Data Warehouse. Nhờ sự chỉ dẫn tận tâm của Thầy, chúng em đã có cơ hội tiếp cận với những khái niệm và kỹ thuật mới mẻ, cũng như thực hành và phát triển các dự án thực tế. Khóa học này đã trang bị cho chúng em không chỉ kiến thức chuyên môn mà còn cả những kỹ năng quan trọng trong công việc và cuộc sống. Những bài học và kinh nghiệm quý báu mà chúng em đã nhận được sẽ là hành trang vững chắc trên con đường phát triển sự nghiệp trong lĩnh vực công nghệ thông tin. Song cũng không tránh khỏi những thiếu sót, mong thầy sẽ tận tình chỉ bảo thêm cho chúng em.

**NHẬN XÉT**

Ngày 10 tháng 12 năm 2024

(**Ký tên**)

**Revision History**

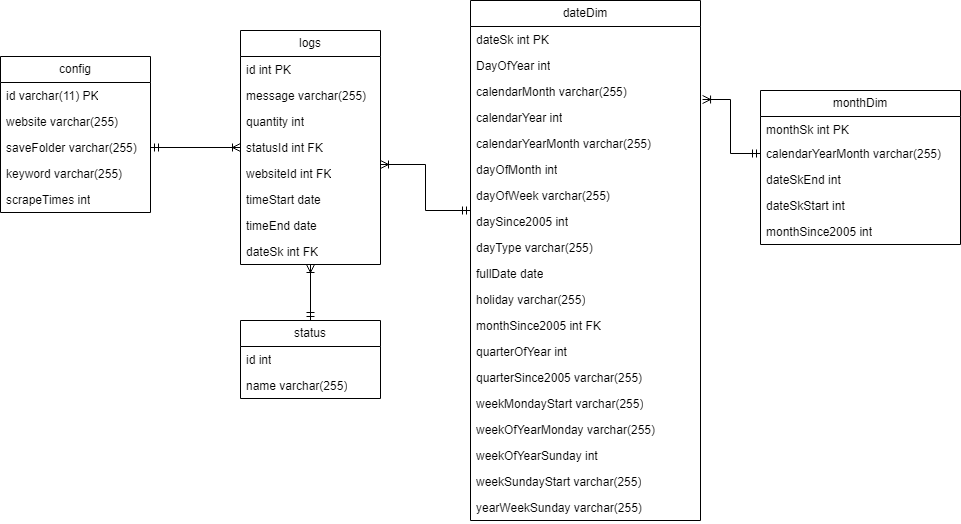
|  |  |  |
| --- | --- | --- |
| **Ngày** | **Tác giả** | **Nội dung** |
| 12/10/2024 | C.T.Nam | Sửa lại datafeed\_specification, thêm các file excel là các dữ liệu được scrape |
| 03/11/2024 | C.T.Nam | Cập nhập datafeed specification |
| 13/11/2024 | C.T.Nam | Bổ sung bảng dateDim và monthDim |
| 09/12/2024 | C.T.Nam | Thêm exception |
| 09/12/2024 | C.T.Nam | Sửa loại dữ liệu trong table logs thành date cho timeStart và timeEnd |

1. **Datafeed Specification**

|  |  |  |  |
| --- | --- | --- | --- |
| **Trường** | **Kiểu dữ liệu** | **Mô tả** | **Bắt buộc** |
| naturalId | number | Mã xe đạp duy nhất | có |
| id | text | Mã xe đạp từ trang web | Không |
| name | text | Tên | Không |
| price | text | Giá bán | Không |
| price\_sale | text | Giá đã giảm | Không |
| brand | text | Thương hiệu | Không |
| color | text | Màu | Không |
| size | text | Kích thước | Không |
| status | text | Tình trạng | Không |
| descriptionPart1 | text | Các thông số kỹ thuật | Không |
| descriptionPart2 | text | Các thông số kỹ thuật | Không |
| descriptionPart3 | text | Các thông số kỹ thuật | Không |
| timeStartScrape | datetime | Thời gian bắt đầu lấy dữ liệu | Có |

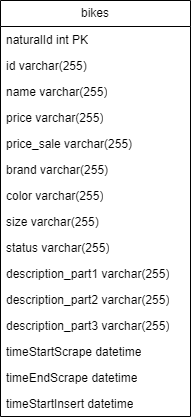
|  |  |  |  |
| --- | --- | --- | --- |
| timeEndScrape | datetime | Thời gian kết thúc lấy dữ liệu | Có |
| timeStartInsert | datetime | Thời gian thêm vào bảng | Có |

1. **Cơ sở dữ liệu**
2. **control**



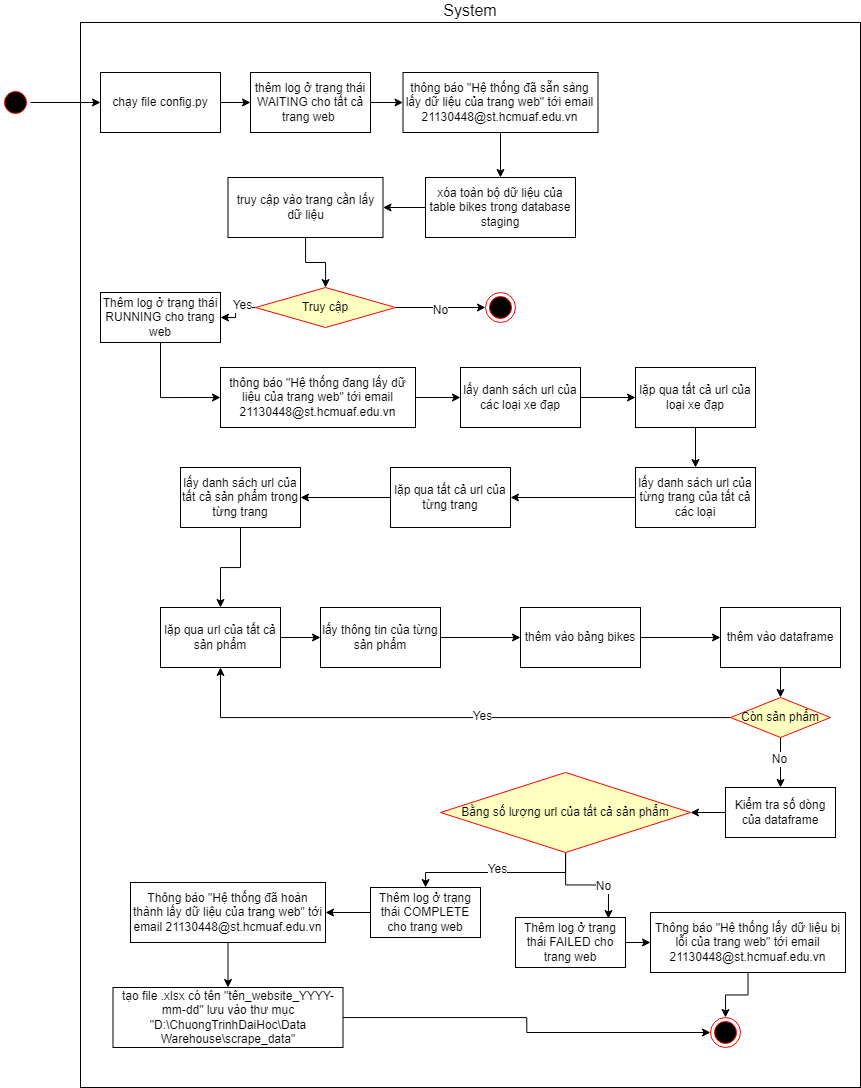
* 1. **config**
     1. **Mô tả**
* Chứa các thông tin của các trang web muốn lấy dữ liệu
  + 1. **Cấu trúc**
* id: duy nhất, để định dang từng trang web
* website: tên của trang web
* saveFolder: thư mục lưu trữ sau khi lấy dữ liệu
* keyword: duy nhất, là từ khóa của trang web
* scrapeTimes: số lần lấy dữ liệu
  1. **logs**
     1. **Mô tả**
* Dùng để lưu lại quá trình lấy dữ liệu của trang web
  + 1. **Cấu trúc**
* id: duy nhất, để định dang từng log
* message: ghi chú cho log
* quantity: số lượng dữ liệu đã lấy
* statusId: trạng thái của log
* websiteId: id của website
* timeStart: thời gian bắt đầu vào trang web
* timeEnd: thời gian thoát khỏi trang web
* dateSk: đây là id của bảng dateDim giúp truy xuất thông tin về thời gian nhanh chóng
  1. **status**
     1. **Mô tả**
* Lưu trữ trạng thái của log
* RUNNING: hệ thống đang trong quá trình lấy dữ liệu
* WAITING: thể hiện trong ngày hôm nay website này chưa được lấy dữ liệu và đang chờ để lấy dữ liệu
* FAILED: thể hiện nếu số lượng dòng lấy dữ liệu bằng không
* COMPLETE: đã hoàn thành lấy dữ liệu thành công
  + 1. **Cấu trúc**
* id: duy nhất, để định dang từng trạng thái
* name: tên trạng thái
  1. **dateDim**
     1. **Mô tả**
* Lưu trữ các thông tin về các ngày trong quá khứ và tương lai. Thể hiện ở dạng chữ cũng như dạng số giúp truy xuất nhanh chóng
  + 1. **Cấu trúc**
* dateSk: khóa chính, duy nhất để định danh từng ngày
* DayOfYear: ngày thứ i(1, 2, 3, 4,…) trong năm
* calendarMonth: tháng được lưu trữ ở dạng chữ
* calendarYear: năm(2005, 2006, 2007,…)
* calendarYearMonth: năm và tháng(2005-jan, 2005-Feb,…)
* dayOfMonth: ngày thứ i(1, 2, 3, 4,…) trong tháng
* dayOfWeek: thứ trong tuần được lưu trữ ở dạng text(monday, tuesday, wednesday,… sunday)
* daySince2005: ngày thứ i được tính từ ngày 01/01/2005
* dayType: lưu trữ ngày đó là ngày trong tuần hay cuối tuần(weekday, weekend)
* fullDate: ngày tháng năm(YYYY/mm/dd)
* holiday: ngày lễ(non-holiday, holiday)
* monthSince2005: tháng thứ i tính từ ngày 01/01/2005
* quarterOfYear: quý của năm(tháng 1, 2, 3 là quý 1, tháng 4, 5, 6 là quý 2,…) ở dạng số
* quarterSince2005: quý của năm(2005-Q01, 2005-Q02,…) ở dạng chữ
* weekMondayStart: fullDate(dd/mm/YYYY) của ngày thứ 2 trong tuần
* weekOfYearMonday: tuần thứ i(1…53) của năm, đầu tuần tính từ thứ hai
* weekOfYearSunday: tuần thứ i(1…52) của năm, đầu tuần tính từ chúa nhật
* weekSundayStart: fullDate(dd/mm/YYYY) của ngày chúa nhật trong tuần
* yearWeekSunday: năm và tuần(2005-W01, 2005-W02) lưu ở dạng chữ
  1. **monthDim**
     1. **Mô tả**
* Lưu trữ thông tin về tháng của ngày nào đó giúp truy xuất thông tin nhanh chóng
  + 1. **Cấu trúc**
* monthSk: khóa ngoại liên kết đến cột “monthSince2005” trong bảng dateDim
* calendarYearMonth: năm và tháng(2005-Jan, 2005-Feb)
* dateSkEnd: khóa ngoại liên kết đến cột “dateSk” trong bảng dateDim, thể hiện tháng này kết thúc ở ngày nào
* dateSkStart: khóa ngoại liên kết đến cột “dateSk” trong bảng dateDim, thể hiện tháng này bắt đầu từ ngày nào
* monthSince2005: tháng thứ i(1, 2, 3, 4,…) tính từ năm 2005

1. **staging**



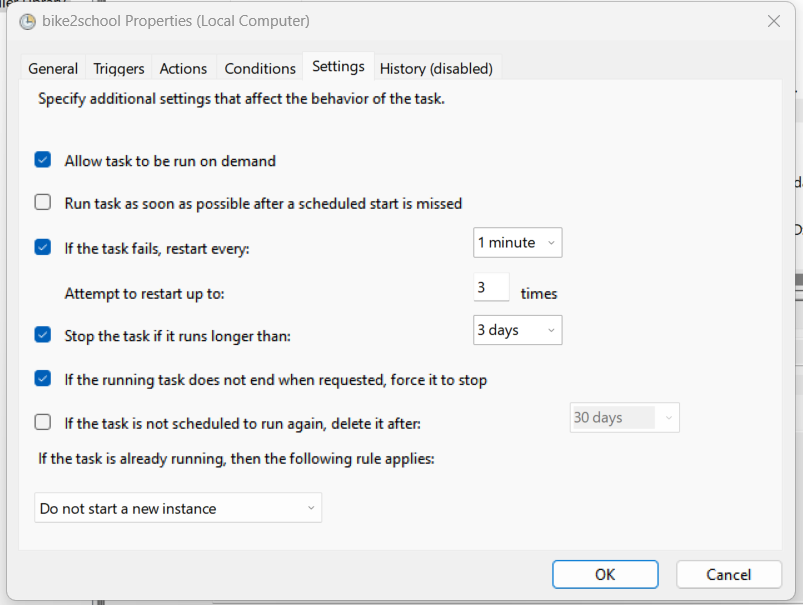
* 1. **bikes**
     1. **Mô tả**
* Dùng để lưu trữ các xe đạp đã lấy được từ các trang web khác nhau
  + 1. **Cấu trúc**
* naturalId: duy nhất, để định danh từng xe đạp
* id: mã xe đạp trên trang web
* name: tên
* price: giá bán
* price\_sale: giá đã giảm
* brand: thương hiệu
* color: màu
* size: kích thước
* status: tình trạng
* description\_part1: các thông số kỹ thuật
* description\_part2: các thông số kỹ thuật
* description\_part3: các thông số kỹ thuật
* timeStartScrape: thời gian bắt đầu lấy dữ liệu
* timeEndScrape: thời gian kết thúc lấy dữ liệu
* timeStartInsert: thời gian thêm vào cơ sở dữ liệu

1. **Workflow**
2. **Lấy dữ liệu từ web, lưu xuống database và tạo ra file**

****

1. **Exception handling**
2. **Lỗi không truy cập được trang web**
   1. **Nguyên nhân**

* Trang web block IP
* Kết nối không ổn định
  1. **Xử lý**
     1. **Cách 1**
* Truy cập vào trang web bằng trình duyệt
* Nếu không vào được thì đợi khoảng 30p sau đó sẽ chạy chương trình thủ công
  + 1. **Cách 2**
* Vào Task Scheduler > chọn thư mục “Datawarehouse”>chọn dòng chạy chương trình > qua tab Settings > check vào ô "If the task fails, restart every: ”
* Có thể cấu hình theo mong muốn

****

* Chương trình sẽ tự động chạy lại sau mỗi 1p nếu như lỗi. Tối đa 3 lần

1. **Lỗi không lấy được mô tả**
   1. **Nguyên nhân**

* Các phần mô tả không tổ chức đồng nhất
* Selenium không thể tìm thấy
  1. **Xử lý**
     1. **Cách 1**
* Không lấy mô tả của sản phẩm đóvà tiếp tục chạy

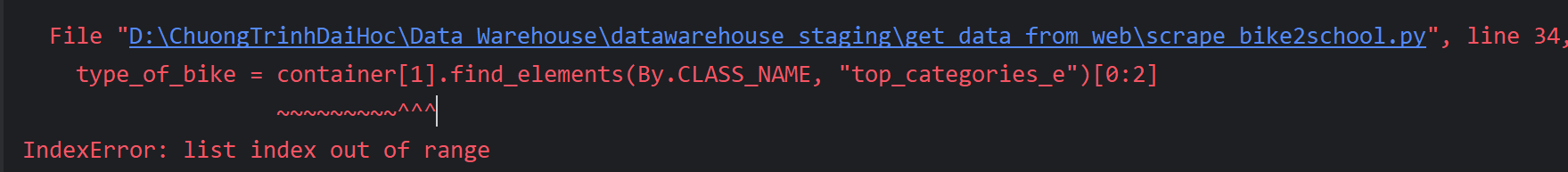
1. **Lỗi số lượng sản phẩm bằng 0**
   1. **Nguyên nhân**

* Không thể truy cập trang web(lỗi 1)
* Không thể truy cập sản phẩm
  1. **Xử lý**
     1. **Cách 1**
* Bỏ qua sản phẩm đó và tiếp tục chạy
  + 1. **Cách 2**
* Xử lý giống lỗi 1

1. **Lỗi không thể lưu vào CSDL**
   1. **Nguyên nhân**

* Do mô tả quá dài vượt quá khả năng lưu trữ của CSDL
  1. **Xử lý**
* Tạo 3 cột mô tả
* Chia độ dài của mô tả thành 3 phần bằng nhau

1. **Lỗi index out of**



* 1. **Nguyên nhân**
* Selenium chưa kịp xử lý nhưng trang web đã bị tắt dẫn tới container rỗng
  1. **Xử lý**
* Xử lý giống lỗi 1