**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: Báo cáo giá xe đạp**

**Sinh viên thực hiện:**

Cao Thành Nam - 21130448

**Mục lục**

[**Lời mở đầu** 2](#_Toc182338079)

[**LỜI CÁM ƠN** 3](#_Toc182338080)

[**NHẬN XÉT** 4](#_Toc182338081)

[**I.** **Datafeed Specification** 5](#_Toc182338082)

[**1.** **Datafeed Specification** 5](#_Toc182338083)

[**II.** **Cơ sở dữ liệu** 5](#_Toc182338084)

# **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**

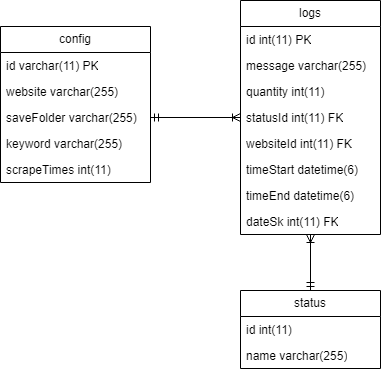
Ngày 02 tháng 06 năm 2024

(**Ký tên**)

1. **Datafeed Specification**

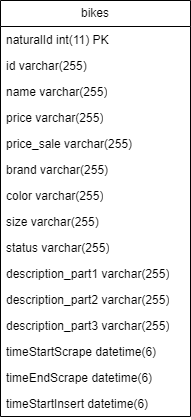
|  |  |  |  |
| --- | --- | --- | --- |
| **Trường** | **Kiểu dữ liệu** | **Mô tả** | **Bắt buộc** |
| id | String | Mã xe đạp | Không |
| name | String | Tên | Có |
| price | Integer | Giá bán | Không |
| price\_sale | Integer | Giá đã giảm | Có |
| brand | String | Thương hiệu | Không |
| color | String | Màu | Không |
| size | Integer | Kích thước | Không |
| status | String | Tình trạng | Có |
| description\_part1 | String | Các thông số kỹ thuật | Không |
| description\_part2 | String | Các thông số kỹ thuật | Không |
| description\_part3 | String | Các thông số kỹ thuật | Không |
| timeStartScrape | LocalDateTime | Thời gian bắt đầu lấy dữ liệu | Có |
| timeEndScrape | LocalDateTime | Thời gian kết thúc lấy dữ liệu | Có |
| timeStartInsert | LocalDateTime | Thời gian thêm vào cơ sở dữ liệu | Có |

1. **Datafeed Specification**
2. **Cơ sở dữ liệu**
3. **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, WAITING, FAILED, COMPLETE)
  + 1. **Cấu trúc**
* id: duy nhất, để định dang từng trạng thái
* name: tên trạng thái

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 dang từng xe đạp
* id: mã xe đạp
* 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