|  |  |
| --- | --- |
|  | **ĐẠI HỌC CÔNG NGHIỆP THÀNH PHỐ HỒ CHÍ MINH** |
|  | **KHOA CÔNG NGHỆ THÔNG TIN** |

**BÁOA logo with a red and blue letter and a black background

Description automatically generated CÁO – XÂY DỰNG LUỒNG THU THẬP DỮ LIỆU EDUMALL**

**Thành viên:**

|  |  |  |
| --- | --- | --- |
| **STT** | **Họ và tên** | **MSSV** |
| 1 | Lê Thị Kim Oanh | 22632471 |
| 2 | Nguyễn Duy | 22708301 |
| 3 | Lê Đặng Phương Trinh | 22636871 |
| 4 | Nguyễn Huy Khang | 22710921 |
| 5 | Đào Trung Kiên | 22635961 |
| 6 | Trần Thị Mỹ Uyên | 22718671 |

**Lớp học phần:** DHHTTT18BTT

**Ngày bắt đầu:** 10/11/2024

**Ngày kết thúc:** 22/11/2024

**Giảng viên hướng dẫn:** Thầy Huỳnh Nam

**Ngày nộp bài:** 22/11/2024

**MỤC LỤC**

[**I.** **THU THẬP DỮ LIỆU VÀ LƯU TRỮ DỮ LIỆU VÀO MONGODB** 3](#_Toc183098448)

[**II.** **XỬ LÝ DỮ LIỆU VỚI PYSPARK** 10](#_Toc183098449)

[**III.** **MÔ HÌNH HÓA CSDL** 12](#_Toc183098450)

[**IV.** **THIẾT KẾ BÀI TOÁN PHÂN TÍCH DỮ LIỆU** 15](#_Toc183098451)

[**V.** **TỐI ƯU MÔ HÌNH VỚI DATA WAREHOUSE** 16](#_Toc183098452)

[**VI.** **SẢN PHẨM CUỐI CÙNG** 20](#_Toc183098453)

[**1.** **Lược đồ kho dữ liệu** 20](#_Toc183098454)

[**2.** **Bảng điều khiển và phân tích** 22](#_Toc183098455)

1. **THU THẬP DỮ LIỆU VÀ LƯU TRỮ DỮ LIỆU VÀO MONGODB**

* Thiết lập Selenium WebDrive: Khởi động trình duyệt Chrome qua ChromeDriver để tự động truy cập danh sách các khóa học trong trang web Edumall.vn và lấy các thông tin chi tiết về khóa học. Sử dụng Selenium để tránh việc ngăn chặn thu thâp dữ liệu từ quản trị viên trang web.
* Truy cập vào danh sách các khóa học để thu thập các liên kết khóa học và lưu vào tệp list\_links\_full2.txt thông qua phương thức “find\_elements”. Để từ đó có thể trích xuất tất cả thông tin cần thiết.
* Tạo một vòng lặp “for” khác để lặp qua tất cả các trang của khóa học để thu thập dữ liệu. Do đó mỗi trang được hiển thị đầy đủ với các nguồn html và tất cả XPath của các phần tử có sẵn để thu thập dữ liệu.
* Tạo vòng lặp “for” để lặp qua tất cả danh sách các khóa học và thu thập dữ liệu trong mỗi vòng lặp.
* Sử dụng hàm WebDriverWait để đảm bảo dữ liệu được tải đầy đủ trước khi trích xuất, tránh các lỗi khi các phần tử chưa tải xong.
* Đây là nội dung file của mã nguồn “edumall\_selenium.py”:

A computer screen with text on it

Description automatically generated

A screen shot of a computer program

Description automatically generated

A screen shot of a computer program

Description automatically generated

A screen shot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screen shot of a computer program

Description automatically generated



A computer screen shot of text

Description automatically generated

* Đối với mỗi dữ liệu được thu thập từ trang web, bằng cách sử dụng PyMongo để kết nối với MongoDB, dữ liệu sẽ được chèn vào MongoDB dưới dạng JSON với hàm insert\_many () để lưu trữ nhanh và hiệu quả hơn, nếu không thành công sẽ tiếp tục thu thập hàng tiếp theo với khối “try\_except”.

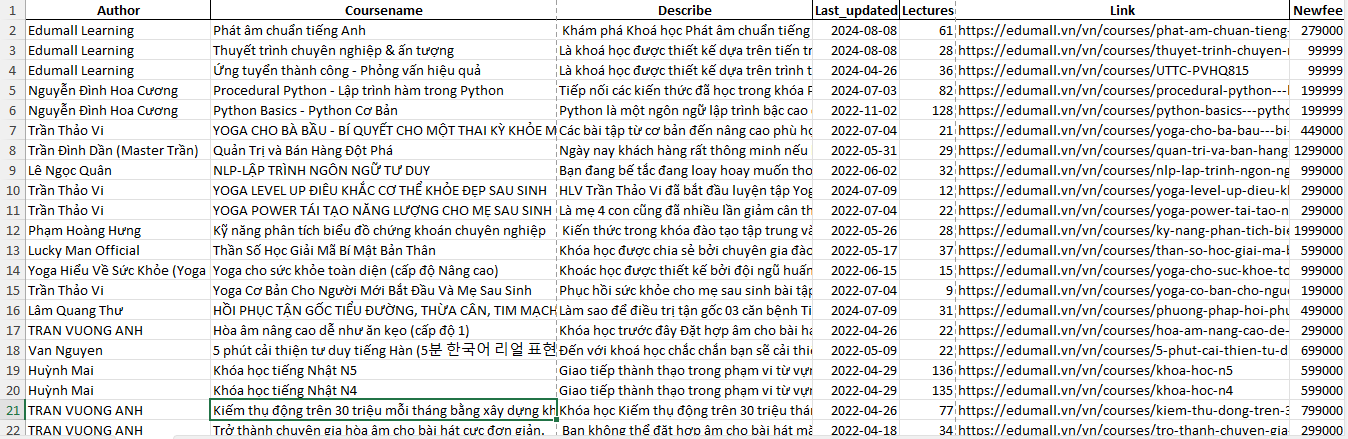
A computer screen shot of a black background

Description automatically generated with medium confidence

A screen shot of a computer code

Description automatically generated

* Kết quả thu được các khóa học với các thông tin: Tên khóa học, giảng viên, chủ đề, mô tả, những điều học được, giá cũ, giá mới, ngày cập nhật gần đây, số chương, số bài học, thời gian học, link từ các danh sách liên kết khóa học của từng trang trong khóa học dưới file csv và json. Đồng thời lưu trữ dữ liệu vào MongoDB để dễ dàng truy xuất dữ liệu linh hoạt và mở rộng trong quy mô   
  lưu trữ.
* *File CSV:*



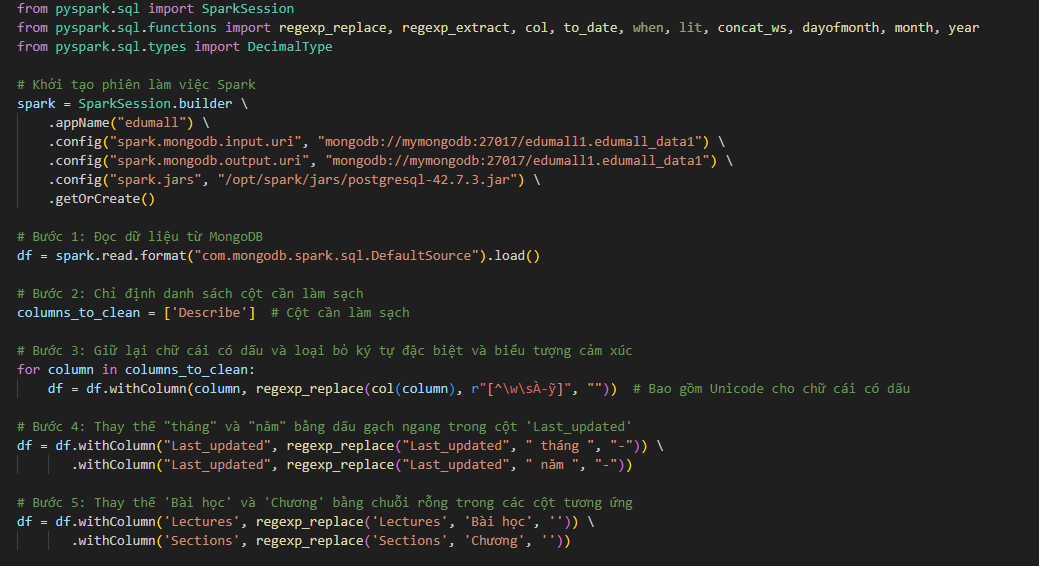
* *File JSON:*

A screenshot of a computer program

Description automatically generated

1. **XỬ LÝ DỮ LIỆU VỚI PYSPARK**

* Khởi tạo phiên làm việc Spark.
* Kết nối để đọc dữ liệu từ MongoDB.
* Loại bỏ các ký tự không mong muốn bằng cách sử dụng lệnh regexp\_replace. Áp dụng biểu thức chính quy để tìm và thay thế các ký tự đặc biệt giúp dữ liệu trở nên sạch hơn mà không làm thay đổi kiểu dữ liệu của cột.
* *Mã nguồn file “edumall\_cleaning.py”:*



A screen shot of a computer program

Description automatically generated



* *Dockerfile để build Spark container:*

A screenshot of a computer program

Description automatically generated

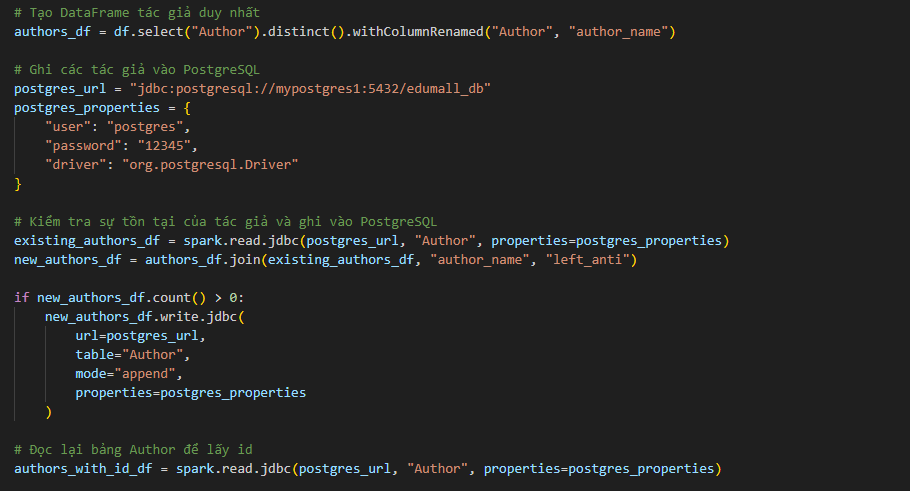
A computer screen shot of a program

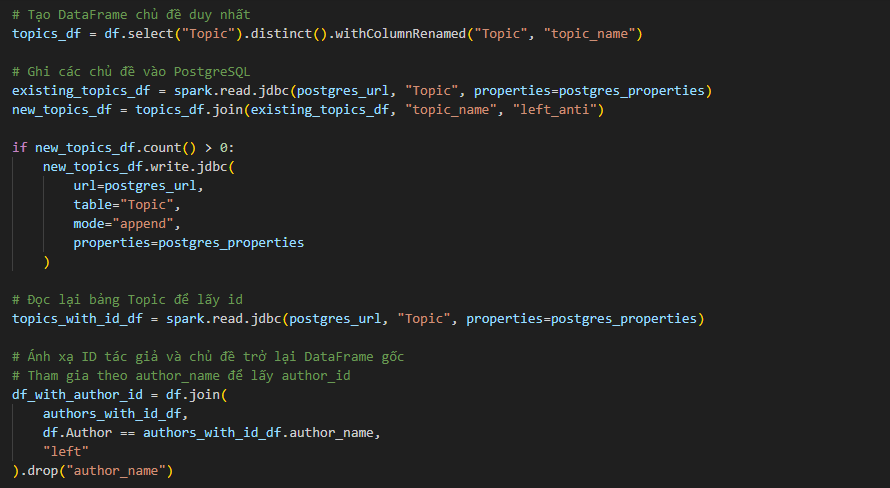
Description automatically generated

1. **MÔ HÌNH HÓA CSDL**

**Tạo mô hình quan hệ cho các bảng sau đó lưu vào Postgres:**

* Khóa học: Tên khóa học, mô tả, giá cũ, giá mới, đánh giá, thời gian học, ngày cập nhật cuối cùng, tháng cập nhật cuối cùng, năm cập nhật cuối cùng, số chương, số bài học, những điều học được, mã giảng viên, mã chủ đề.
* Chủ đề: Mã chủ đề, tên chủ đề.
* Giảng viên: Mã giảng viên, tên giảng viên.
* Nội dung mã nguồn “edumall\_postgresql.py”:





A screen shot of a computer program

Description automatically generated

* *Nội dung Dockerfile để build Postgres container:*

A screenshot of a computer program

Description automatically generated



1. **THIẾT KẾ BÀI TOÁN PHÂN TÍCH DỮ LIỆU**

Bài toán: Các yếu tố ảnh hưởng đến giá cả khóa học.

A computer code on a dark background

Description automatically generated

A computer code with text

Description automatically generated with medium confidence

A screen shot of a computer code

Description automatically generated

A computer screen shot of a computer code

Description automatically generated with medium confidence

A computer code on a dark background

Description automatically generated

A computer screen shot of a blue and white text

Description automatically generated with medium confidence



1. **TỐI ƯU MÔ HÌNH VỚI DATA WAREHOUSE**

* *Mã nguồn file Create\_SQLServer.py, dùng để khởi tạo bảng trong   
  Data Warehouse:*

**A screen shot of a computer

Description automatically generated**

A screen shot of a computer program

Description automatically generated

* *Mã nguồn Postgre\_to\_SQLServer.py, dùng để kết nói dữ liệu từ PogressSQL sang SQL Server.*

A screen shot of a computer

Description automatically generated

A screen shot of a computer program

Description automatically generated

A screen shot of a computer program

Description automatically generated

A screen shot of a computer program

Description automatically generated

1. **SẢN PHẨM CUỐI CÙNG**
2. **Lược đồ kho dữ liệu**

* Kho dữ liệu có thể cung cấp thêm thông tin về mối quan hệ giữa khóa học, chủ đề, tác giả và thời gian cập nhật.

A screenshot of a computer

Description automatically generated

* Sử dụng truy vấn “Select ... from ...” để kiểm tra dữ liệu được chèn trong kho dữ liệu thành công hay chưa.

A screenshot of a computer

Description automatically generated

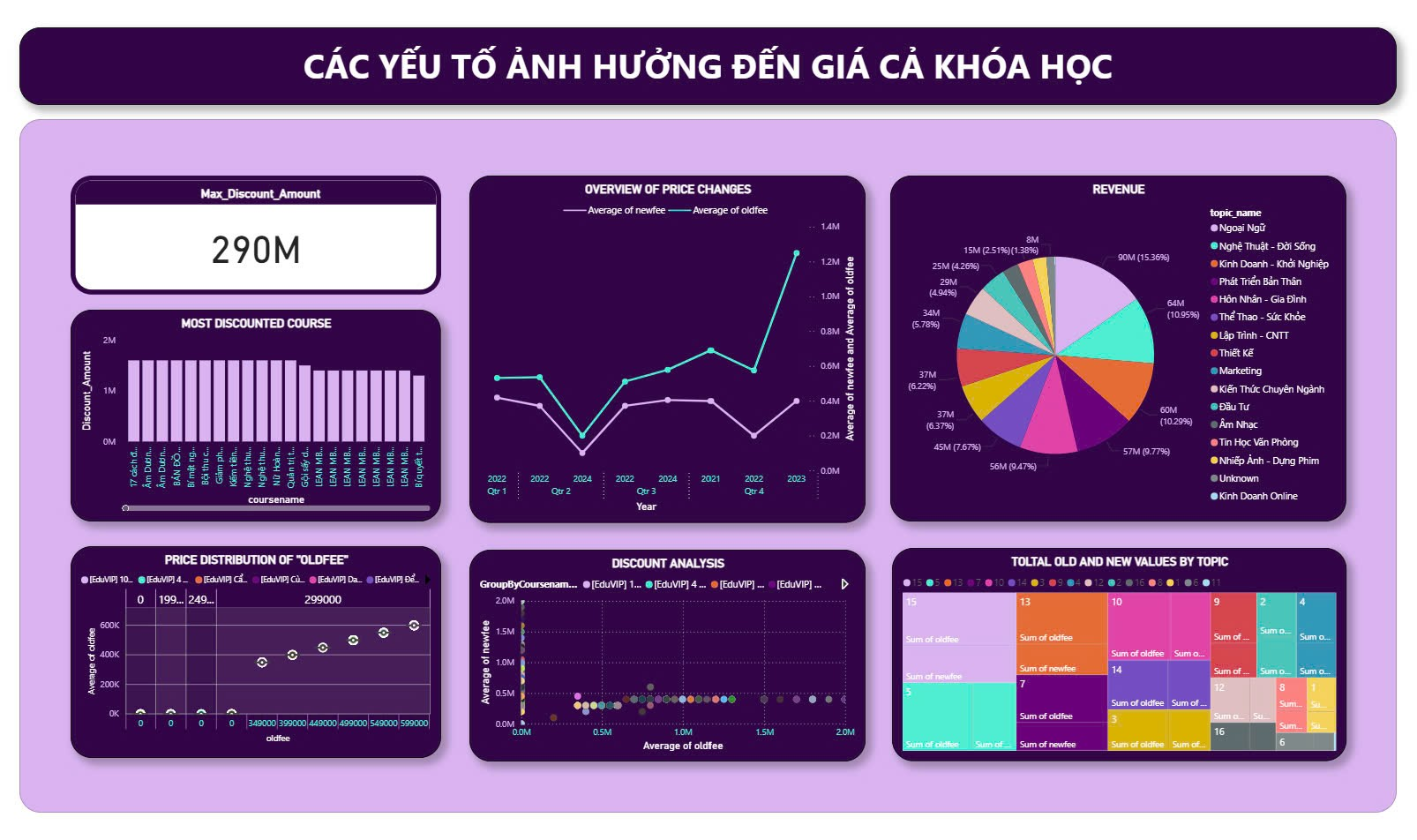
* Sau đó, chuyển sang bước tiếp theo để tạo cơ sở dữ liệu khối.

A screenshot of a computer

Description automatically generated

1. **Bảng điều khiển và phân tích**

* *Tổng quan về sự thay đổi giá cả*



1. **Max\_Discount\_Amount (Card Visual)**

**Ý nghĩa**: Hiển thị **giá trị giảm giá cao nhất** của một khóa học, biểu thị bằng con số cụ thể (290M). Giá trị này được tính bằng Measure Max\_Discount\_Amount.

**Mục tiêu:** Giúp người dùng nhanh chóng nhận diện khóa học có mức giảm giá cao nhất.

1. **Overview of Price Changes (Line Chart)**

**Ý nghĩa:** Theo dõi sự biến đổi giá trung bình (newfee và oldfee) theo từng quý và năm. Đường màu xanh thể hiện giá trung bình newfee, đường màu tím thể hiện oldfee.

**Mục tiêu:** So sánh xu hướng giá cũ và giá mới qua thời gian, từ đó nhận diện thời điểm thay đổi giá đáng kể.

1. **Revenue (Pie Chart)**

**Ý nghĩa:** Biểu đồ tròn hiển thị doanh thu từ giá mới (newfee) phân bổ theo chủ đề (topic\_name). Tooltip có thể hiển thị thêm tổng doanh thu và tỷ lệ phần trăm từng chủ đề.

**Mục tiêu:** Xác định các chủ đề mang lại doanh thu cao nhất. Phân tích tỷ trọng doanh thu giữa các chủ đề.

1. **Most Discounted Course (Column Chart)**

**Ý nghĩa:** Biểu đồ cột sắp xếp các khóa học theo giá trị giảm giá (Discount\_Amount). Cột cao nhất đại diện cho khóa học giảm giá nhiều nhất.

**Mục tiêu:** Tìm khóa học nào có mức giảm giá lớn nhất. So sánh mức giảm giá giữa các khóa học.

1. **Price Distribution of "Oldfee" (Boxplot)**

**Ý nghĩa:** Biểu đồ Boxplot hiển thị phân phối giá cũ (oldfee). Các chỉ số như Q1, Median, Q3, Min, Max được tính toán và thể hiện để minh họa sự khác biệt trong dữ liệu.

**Mục tiêu:** Phân tích giá cũ của các khóa học để nhận diện các ngoại lệ (outliers) và mức độ phân tán.

1. **Discount Analysis (Scatter Plot)**

**Ý nghĩa:** Biểu đồ Scatter so sánh giá trị giảm giá (Discount\_Amount) và giá trung bình của oldfee theo từng khóa học. Tooltip có thể hiển thị thông tin chi tiết như tên khóa học, giá cũ, giá mới và số tiền giảm.

**Mục tiêu:** Nhận diện khóa học có giá trị giảm cao và mối liên hệ với giá cũ.

1. **Total Old and New Values by Topic (Treemap)**

**Ý nghĩa:** Treemap hiển thị tổng giá trị oldfee và newfee phân bổ theo từng chủ đề (topic\_name). Các ô được phân cấp theo tỷ lệ giá trị, tooltip có thể cung cấp chi tiết từng chủ đề.

**Mục tiêu:** Phân tích sự chênh lệch giữa giá cũ và giá mới theo chủ đề. Xác định chủ đề có mức tổng giá trị lớn nhất.