**Chương 1. Tổng quát**

* 1. **Giới thiệu đề tài**

Trong bối cảnh giá nhà đang tăng cao, việc đầu tư, mua bán và phát triển bất động sản trong thị trường hiện nay rất quan trọng, Các nhà đầu tư cần biết rõ thị hiếu của khách hàng, bên cạch đó xác định các khu vực tiềm năng để tối ưu hóa lợi nhuận đạt được. Đề tài ‘Phân tích giá nhà ‘ nhằm mục đích cung cấp một cách tiếp cận khoa học cho các nhà đầu tư, vừa tạo điều kiện dễ dàng cho người dùng tiếp cận, tra cứu nhưng thông tin chung cư phù hợp với điều kiện của mình.

* 1. **Nhiệm vụ đề tài**

Áp dụng những kĩ thuật thu thập dữ liệu để thu thập thông tin về các đặc điểm và giá hiện tại của các chung cư. Từ đó xây dụng cở sơ dữ liệu phù hợp để tối ưu quá trình xử lí, sử dụng và quản lí dữ liệu một cách mạnh mẽ. Từ đó phân tích dữ liệu qua các biểu đồ trực quan để giúp người dùng tiếp cận một cách hiệu quả. Điều này giúp các nhà đầu tư và người dùng có cái nhìn sâu sắc và tổng quát về thị trường bất động sản, qua đó định hướng được chiến lược phát triển và đầu tư và mua bán.

* + 1. *Tính cấp thiết đề tài*
    2. *Ý nghĩa khóa học và thực tiễn của đề tài*
  1. **Mục tiêu**

Đề tài cung cấp một bộ cở sở dữ liệu về nhưng đặc điểm và giá cả của nhưng chung cư trong thị trường bất động sản HCM, xây đựng một hệ thống truy vấn mạnh mẽ và phân tích các điểm nóng bất động sản tại HCM.

* + 1. *Mục tiêu tổng quát*
    2. *Mục tiêu cụ thể*

Mục tiêu của chúng tôi là xây dựng một bộ dữ liệu tổng quát về thông tin các đặc điểm, tiện ích và giá cả tại các chung cư trên địa bàn HCM. Quy trình thực hiện bao gồm: Lựa chọn trang web, thiết kế cơ sở dữ liệu, thu thập dữ liệu, lưu dữ liệu và cuối cùng là phân tích. Với từng bước một, chúng tôi lựa chọn web batdongsan để thu thap dữ iệu, sử dụng Môngdb để tổ chức và lưu trữ dữ liệu cuối cùng Selenium và sklieam để thu thập và phân tích dữ liệu

* 1. **Đối tượng và phạm vi**
     1. **Đối tượng**
     2. **Phạm vi**
  2. **Phương pháp nghiên cứu**
  3. **Đóng góp của đề tài**

**Chương 2. Cở sở lí thuyết**

* 1. **Selenium Webdriver** 
     1. *Giới thiệu về selenium webdriver*

WebDriver không chỉ là công cụ tự động hóa phổ biến và mạnh mẽ cho kiểm thử các ứng dụng web mà còn rất hữu ích trong việc thu thập dữ liệu từ các trang web. WebDriver cho phép bạn thực hiện một số hành động như: tương tác trực tiếp, click vào đối tượng, điền văn bản trên web một cách tự động.

WebDriver hỗ trợ nhiều trình duyệt phổ biến như Chrome, Firefox, Safari và Edge, điều này giúp bạn thu thập từ các trình duyệt khác nhau mà không cần thay đổi mã nguồn. Bên cạnh đó, WebDriver hỗ trợ nhiều đa dạng ngôn ngữ lập trình như Python, Js, Java ... Là một open souch được phát triển bởi selenium, Web Driver hỗ trợ nhiều hệ điều hành khác nhau như Win , Ubuntu, Linux, Mac ... Nhờ vào tính linh hoạt này, WebDriver được phổ biến trong công đồng phát triển.

* + 1. *Ưu điểm và hạn chế*

Ưu điểm:

* Selenium webdriver hỗ trợ tương tác trực tiếp các ứng dụng web
* Hỗ trợ đa ngôn ngữ lập trình và hệ điều hành cho nhà phát triển
* Sử dụng mã cho nhiều trình duyệt web khác nhau

Hạn chế:

* Selenium webdriver chỉ hỗ trợ các ứng dụng web
* Không hỗ trợ hệ điều hành android
* Không tự động vượt qua capa trên các ứng dụng web
  1. **Cở sở dữ liệu phân tán**
     1. *Giới thiệu về NoSQL*

Not only SQL **(**NoSQL) là một loại cở sở dữ liệu không sử dụng các mô hình quan hệ truyền thống để lưu trữ và quản lí dữ liệu. NoSQL cho phép thay đổi linh hoạt trong cách tổ chức dữ liệu, phù hợp với dữ liệu phi cấu trúc và bán cầu trúc.

Cở sở dữ liệu NoSQL sử dụng các mô hình lưu trữ khác nhau như lưu trữ theo tài liệu, cặp khóa – giá trị, cột hay đồ thị, điều này cho phép NoSQL lữu trữ dữ liệu linh hoạt hơn so với các cơ sở dữ liệu quan hệ, nơi mà dữ liệu được lưu trữ dưới dạng bảng có cấu trúc cố định.

* + 1. *Các loại cơ sở dữ liệu phân tác*
* Cơ sở dữ liệu tài liệu: dữ liệu được lưu trữ dưới dạng tài liệu thường là BSON hay JSON.
* Cơ sở dữ liệu khóa – giá trị: Lưu trữ dưới dạng khóa – giá trị. Mỗi khóa trỏ đến một giá trị duy nhất.
* Cơ sở dữ liệu cột: Lưu trữ dữ liệu dưới dạng cột thay vì hàng như SQL.
* Cơ sở dữ liệu đồ thị: Tối ưu cho việc lưu trữ và truy vấn dữ liệu quan hệ phức tạp thông qua đồ thị.
  + 1. *Lợi ích của NoSQL*

NoSQL có nhiều lợi ích cho việc lưu trữ và quản lí dữ liệu. Đầu tiên, với khả năng phân tán dữ liệu ở nhiều máy chủ khác nhau, điều này giúp dễ dàng đáp ứng nhu cầu lưu trữ dữ liệu lớn mà không cần năng cấp phần cứng máy chủ hiện tại. Thứ hai, khả năng lưu trữ dữ liệu phi cấu trúc và bán cấu trúc phù hợp với nhiều loại cơ sở dữ liệu khác nhau giúp dễ dàng phát triển các ứng dụng một cách hiệu quả. Thứ ba, tính linh hoạt của NoSQL còn thể hiện rõ hơn khi thay đổi cấu trúc của dữ liệu một cách dễ dàng mà không cần thay đổi hay phát triển lại như trong cơ sở dữ liệu quan hệ, giảm thiếu thời gian và công sức của nhà phát triển, doanh nghiệp. Cuối cùng, NoSQL thường cung cấp hiệu suất tốt hơn cho các ứng dụng yêu cầu xử lý dữ liệu lớn hoặc có yêu cầu cao về tốc độ truy cập cao. NoSQL có thể xử lý hàng triệu yêu cầu mỗi giây mà không làm giảm tốc độ hoặc hiệu suất.

* 1. **MongoDb**
     1. *Giới thiệu về MongoDB*

MongoDB là một hệ cơ sở quản trị dữ liệu NoSQL nổi bất, được thiết kế để lưu trữ và quản trị dữ liệu phi cấu trúc hoặc bán cấu trúc. MongoDB được thiết kế nhầm đáp ứng sử phát triển của dữ liệu trong bối cảnh yêu cầu về tốc độ xử lí các dữ liệu lớn.

MongoDB lưu trữ dữ liệu dưới dạng document với dịnh dạng BSON, điều này giúp người dùng lưu trữ dữ liệu một cách linh hoạt, khi thay đổi cấu trúc của một dữ liệu, chúng ta không cần sửa đổi toàn bộ dữ liệu trong cở sở dữ liệu.

MongoDB sử dụng chỉ mục để cải thiện tốc độ truy vấn và tìm kiếm, bên cạnh đó MongoDB hỗ trợ nhiều ngôn ngữ lập trình khác nhau, điều này làm chúng đễ đang tích hợp vào trong các ứng dụng một cách dễ dàng và hiệu quả.

* + 1. *Ưu điểm và hạn chế*

Ưu điểm:

* Hỗ trợ phân tán dữ liệu trên nhiều máy chúng
* Sử dụng chỉ mục để tăng tốc độ truy vấn, giúp tìm kiếm dữ liệu nhanh
* Cho phép thực hiện các phép tính phức tạp mà không cần chuyển đổi

Hạn chế:

* Quản lí dữ liệu phức tạp do không thể liên kết các dữ liệu
* Lưu trữ dưới dạng BSON tiêu tốn nhiều bộ nhớ hơn
* Chi phí cao do tiêu tốn nhiều bộ nhớ