ĐẠI HỌC BÁCH KHOA HÀ NỘI
KHOA TOÁN-TIN

—o0o—

XÂY DỰNG HỆ THỐNG PHÂN TÍCH DỮ LIỆU VỀ BẤT ĐỘNG SẢN TẠI HOA KỲ

ĐỒ ÁN II

Chuyên ngành: Hệ thống thông tin quản lý

Giảng viên hướng dẫn: **TS. NGÔ THỊ HIỀN** _____
Chữ kí của GVHD

Sinh viên thực hiện: **BÙI HỒNG GIANG**
Mã số sinh viên: 20206280
Bộ môn: **Toán tin**

HÀ NỘI, 2024

Nhận xét của giảng viên hướng dẫn

1. Mục tiêu và nội dung của đồ án

(a) Mục tiêu:

(b) Nội dung:

2. Kết quả đạt được

3. Ý thức làm việc của sinh viên:

Hà Nội, ngày ... tháng 1 năm 2024

Giảng viên hướng dẫn

TS. NGÔ THỊ HIỀN

Lời cảm ơn

Lời đầu tiên, em xin được gửi lời cảm ơn đến thầy cô, gia đình và anh chị em bạn bè xung quanh đã giúp đỡ, ủng hộ em trong suốt quá trình làm Đồ án.

Cùng với đó, em xin gửi lòng biết ơn sâu sắc đến quý thầy cô khoa Toán - Tin, Đại học Bách khoa Hà Nội đã dùng tri thức và tâm huyết của mình để có thể truyền đạt cho em những kiến thức quan trọng, quý báu, giúp em ngày một tốt hơn.

Đặc biệt, em xin gửi lời cảm ơn chân thành đến TS. Ngô Thị Hiền người đã tận tâm chỉ bảo, hướng dẫn em trong suốt thời gian làm Đồ án vừa qua. Không chỉ hướng dẫn Đồ án, cô còn là người định hướng con đường học thuật, tạo động lực cho em cố gắng theo đuổi lĩnh vực này. Một lần nữa, em xin cảm ơn cô.

Mục lục

Lời mở đầu

Phân tích dữ liệu bất động sản giúp hiểu rõ các yếu tố ảnh hưởng đến giá cả và xu hướng thị trường, là cơ sở cho các quyết định đầu tư. Nghiên cứu của Zillow vào năm 2021 chỉ ra rằng, dữ liệu lớn và phân tích hành vi người mua có tiềm năng dự báo chính xác các biến động giá nhà. Trên cơ sở đó, xây dựng một hệ thống phân tích dữ liệu bất động sản giúp cho khách hàng có cái nhìn cụ thể và chắc chắn hơn khi quyết định đầu tư. Xuất phát từ nhu cầu thực tế trên, em đã chọn đề tài " Xây dựng hệ thống phân tích dữ liệu về bất động sản tại Hoa Kỳ"

Đồ án ngoài phần mở đầu và kết luận gồm những nội dung chính sau:

- **Chương 1:** Trình bày kiến thức nền tảng, cơ sở lý thuyết, công cụ trực quan hóa dữ liệu Power BI, nhằm cung cấp thông tin hữu ích cho quá trình ra quyết định và phân tích kinh doanh.
- **Chương 2:** Thực hiện phân tích và thiết kế các mô hình của hệ thống Data Warehouse đối với bài toán, đảm bảo các dữ liệu được lưu trữ, quản lý một cách hiệu quả và chính xác nhất
- **Chương 3:** Thực nghiệm: xây dựng hệ quản trị cơ sở dữ liệu trên MySQL và xây dựng các Dashboard dựa trên các yêu cầu của hệ thống đặt ra. Đưa ra nhận định về tình hình.

Hà Nội, tháng 1 năm 2024

Tác giả đồ án

Bùi Hồng Giang

Danh sách hình vẽ

Chương 1

Cơ sở lý thuyết

1.1 Phân tích dữ liệu

Khái niệm

Phân tích dữ liệu (DA - Data Analytics) là quá trình xử lý dữ liệu thô thành dữ liệu sạch và có ích theo yêu cầu. Qua đó, đưa ra cách nhìn tổng thể về doanh số trong một giai đoạn nhất định nào đó và dựa trên trên cơ sở dữ liệu đã phân tích đưa ra định hướng phát triển trong tương lai.

Quy trình

- Xác định mục tiêu và câu hỏi liên quan
- Thu thập dữ liệu
- Xử lý dữ liệu
- Phân tích dữ liệu
- Trực quan hóa dữ liệu
- Dự báo và đưa ra kết luận

Tầm quan trọng của DA

Đối với nhiều doanh nghiệp, quá trình này có thể nắm chủ chốt thành hay bại của doanh nghiệp. Bởi lẽ nhờ có nó, giúp doanh nghiệp tìm ra sai lầm trong quá khứ và tìm ra cách giải quyết. Đồng thời qua đó còn khám phá ra được những cơ hội mới giúp

thúc đẩy và phát triển doanh nghiệp. Cho nên, có thể đưa ra một số lợi ích của DA như sau:

- Có cách nhìn tổng quát về đối tượng khách hàng, qua đó giúp tiếp cận hơn đến với những khách hàng tiềm năng.
- Nắm rõ xu hướng (trend) của thị trường kinh doanh.
- Thay đổi, cải thiện, nâng cao chất lượng sản phẩm và dịch vụ cao cấp hơn.
- Giảm chi phí vận hành.
- Hỗ trợ và tăng năng suất làm việc của nhân viên.
- Ngoài ra, còn có thể theo dõi đối thủ cạnh tranh, đây cũng là yếu tố quan trọng trong việc lập chiến lược kinh doanh phù hợp thông qua dự đoán về đối thủ.

1.2 Kinh doanh thông minh

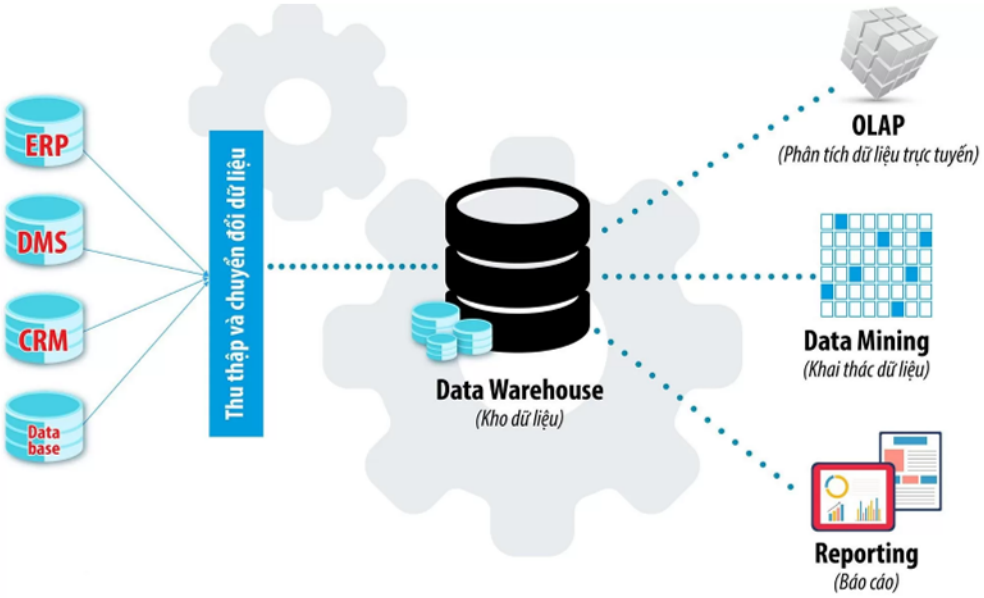
Khái niệm

Kinh doanh thông minh (BI - Business Intelligence) là một khái niệm tổng quát dùng để chỉ các công nghệ, phương pháp và công cụ được sử dụng để phân tích dữ liệu và thông tin kinh doanh để hỗ trợ quyết định và kế hoạch trong doanh nghiệp.

Mô hình hoạt động

BI về cơ bản sẽ chuyển hóa lượng dữ liệu khổng lồ từ một hay nhiều nguồn khác nhau thành những thông tin cần thiết cho hoạt động của doanh nghiệp. Một hệ thống BI bao gồm 3 yếu tố cơ bản sau:

- Kho dữ liệu (Data Warehouse)
- Bộ phận thu thập và chuyển đổi dữ liệu
- Dự báo và phân tích giả lập



Hình 1.1: Mô hình hoạt động của Business Intelligence

Một vài phần mềm công cụ BI

Tableau, Google Data Studio, Power BI, QlikView, Alteryx, Splunk,...

Tầm quan trọng của BI

Công cụ BI giúp tăng tốc độ phân tích thông tin và đánh giá hiệu suất, giúp các công ty giảm thiểu sự kém hiệu quả, khắc phục các vấn đề tiềm ẩn, tìm ra các dòng doanh thu mới và xác định các lĩnh vực phát triển trong tương lai.

Sử dụng BI mang lại nhiều lợi ích cụ thể cho doanh nghiệp, bao gồm:

- Tăng hiệu suất và hiệu quả cho các quy trình hoạt động
- Cung cấp hiểu biết sâu hơn về hành vi của khách hàng
- Giúp theo dõi chính xác hoạt động bán hàng, tiếp thị và tài chính
- Cảnh báo tức thì về sự bất thường của dữ liệu và các vấn đề của khách hàng
- Chia sẻ dữ liệu giữa các phòng ban

1.3 Phân tích nghiệp vụ

Khái niệm

Phân tích nghiệp vụ (BA - Business Analytics) là quá trình tìm hiểu, phân tích và đánh giá các nghiệp vụ của doanh nghiệp nhằm tìm ra cách giải quyết, cải thiện hiệu quả và hiệu suất hoạt động của nó.

Quy trình: Quy trình BA là một quá trình hệ thống giúp các nhà phân tích nghiệp vụ thu thập thông tin và phân tích các yêu cầu của khách hàng để tạo ra một giải pháp kinh doanh hiệu quả. Dưới đây là quy trình BA cơ bản:

- Thu thập thông tin: Trước khi bắt đầu phân tích, các nhà phân tích nghiệp vụ cần thu thập thông tin về các yêu cầu của khách hàng. Các yêu cầu này có thể được thu thập bằng cách phỏng vấn khách hàng, đọc tài liệu hoặc tìm hiểu về ngành công nghiệp của khách hàng.
- Phân tích yêu cầu: Sau khi thu thập thông tin, các nhà phân tích nghiệp vụ sẽ phân tích các yêu cầu của khách hàng và đưa ra các giải pháp kinh doanh phù hợp. Việc này có thể được thực hiện bằng cách sử dụng các công cụ phân tích, bao gồm các phương pháp như SWOT, PESTEL, hoặc đánh giá 5 lực cạnh tranh.
- Xây dựng kế hoạch: Sau khi phân tích yêu cầu, các nhà phân tích nghiệp vụ sẽ xây dựng kế hoạch để triển khai giải pháp kinh doanh. Kế hoạch này bao gồm các bước cụ thể để triển khai giải pháp, các nguồn lực cần thiết và thời gian triển khai.
- Đề xuất giải pháp: Cuối cùng, các nhà phân tích nghiệp vụ sẽ đề xuất giải pháp kinh doanh phù hợp cho khách hàng. Giải pháp này bao gồm các công nghệ và quy trình cần thiết để thực hiện yêu cầu của khách hàng.

Tuy nhiên, quy trình phân tích nghiệp vụ có thể thay đổi tùy thuộc vào từng dự án cụ thể. Điều quan trọng là các nhà phân tích nghiệp vụ cần luôn tập trung vào việc hiểu rõ yêu cầu của khách hàng và tìm kiếm giải pháp kinh doanh phù hợp để giúp khách hàng đạt được mục tiêu kinh doanh của họ.

1.4 Mối liên hệ giữa BI, DA và BA

BI và BA là hai lĩnh vực liên quan đến phân tích dữ liệu để hỗ trợ quyết định kinh doanh. Trong khi đó, DA là một khái niệm rộng hơn, bao gồm cả BI và BA, cùng với các phương pháp phân tích dữ liệu khác như Data Mining và Machine Learning.

Mối liên hệ giữa BI và BA là khá mật thiết, vì BI thường là một phần của quy trình BA. Trong quy trình BA, các thông tin được thu thập từ các nguồn dữ liệu khác nhau được sử dụng để phân tích và đưa ra những quyết định kinh doanh. Các thông tin này được lưu trữ trong các hệ thống BI để có thể truy xuất và sử dụng lại trong tương lai. Điều này cho phép các nhân viên có nhu cầu truy xuất dữ liệu có thể truy cập vào các báo cáo và biểu đồ thống kê để tìm kiếm các thông tin cần thiết.

Trong khi đó, DA bao gồm cả quy trình thu thập, phân tích và đưa ra quyết định từ dữ liệu. Tuy nhiên, BI và BA là những khía cạnh quan trọng trong DA, giúp cho các nhân viên kinh doanh có thể đưa ra những quyết định chính xác hơn dựa trên các thông tin dữ liệu có sẵn. Tóm lại, BI và BA là hai phương pháp quan trọng để hỗ trợ quyết định kinh doanh trong quy trình DA.

Chương 2

Phân tích và thiết kế

2.1 Khảo sát đề tài

2.1.1 Giới thiệu

◇ *Bất động sản và giao dịch mua bán bất động sản là gì*: Bất động sản là một lĩnh vực quan trọng trong nền kinh tế và cuộc sống hàng ngày của chúng ta. Nó bao gồm tài sản như đất đai, các công trình xây dựng và các tài sản liên quan đến địa ốc. Bất động sản có vai trò đa dạng, từ nơi ở cho cá nhân và gia đình đến không gian kinh doanh cho các doanh nghiệp. Giao dịch bất động sản là quá trình mua bán, cho thuê, chuyển nhượng quyền sử dụng các tài sản ở trên thỏa mãn các điều kiện pháp lý. Trong bài báo cáo này, em thực hiện phân tích dữ liệu về việc mua bán bất động sản tại Hoa Kỳ.

◇ *Vai trò của bất động sản trong xã hội hiện nay* :

- Trong đầu tư và tài chính, bất động sản có vai trò quan trọng, thu hút các nhà đầu tư và các tổ chức. Bất động sản hầu hết có xu hướng tăng theo thời gian, thậm chí có thể tăng vọt chỉ trong một thời gian ngắn, điều đó làm cho nó trở thành một hình thức đầu tư lâu dài, hấp dẫn.
- Tạo ra nơi ở và làm việc, và sự phát triển của bất động sản như nhà ở, chung cư, đô thị, trung tâm thương mại, . . . góp phần tạo nên cơ sở hạ tầng của xã hội.
- Bất động sản tác động đến kinh tế: Ngành bất động sản là một trong những ngành kinh tế quan trọng, tạo ra việc làm và thu nhập cho nhiều người. Nó cũng là một chỉ số quan trọng của tình hình kinh tế, với việc tăng giá hoặc

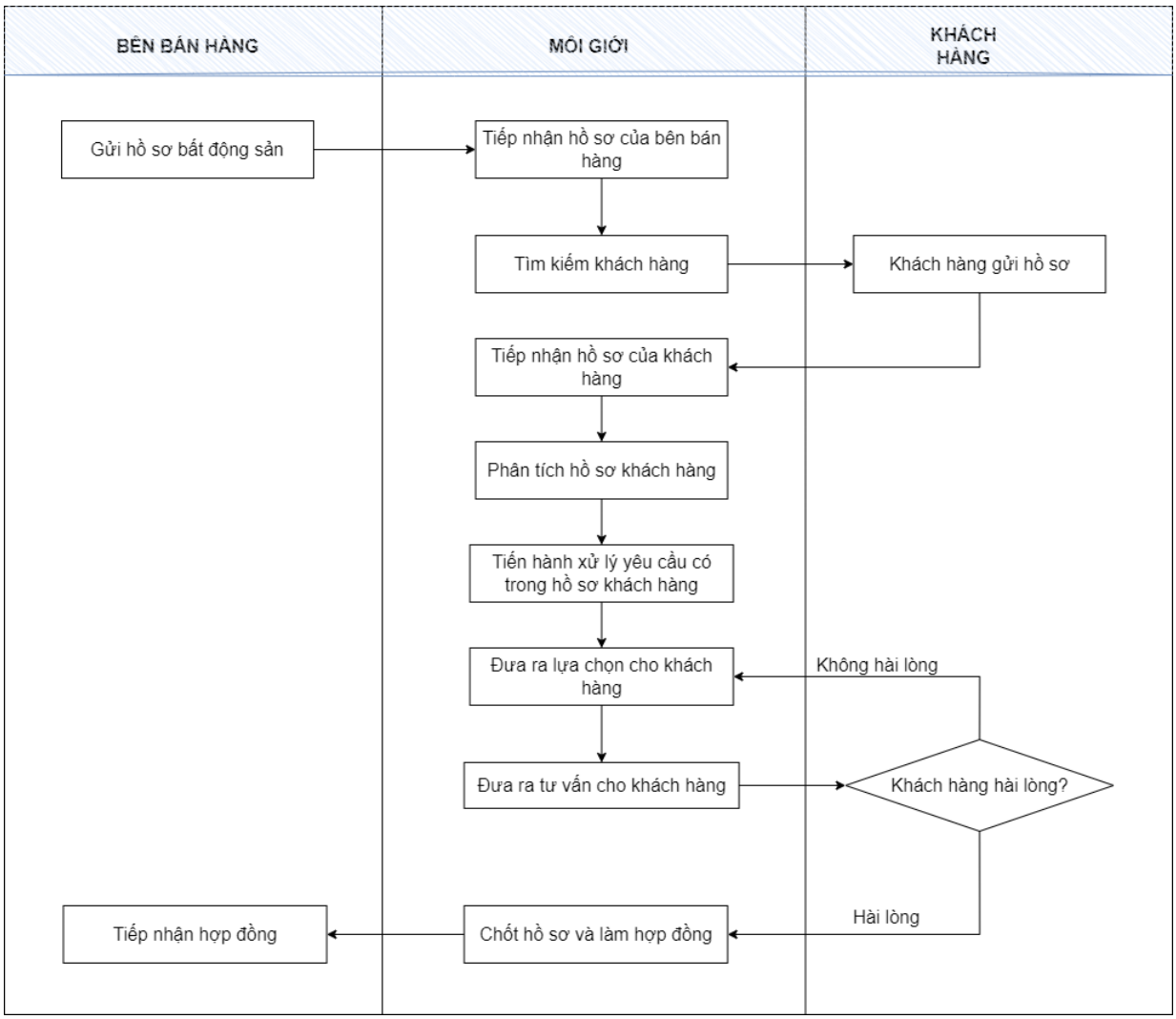
giảm giá của bất động sản thường phản ánh tình hình kinh tế tổng thể.

- Quy hoạch và phát triển bất động sản đóng vai trò quan trọng trong việc kiểm soát sự phát triển của các khu vực mới và cải thiện cơ sở hạ tầng.
- Bất động sản cũng mang giá trị văn hóa và lịch sử, như các công trình kiến trúc cổ, di tích lịch sử, đóng góp vào bản sắc và diện mạo của một đô thị hay quốc gia.

2.1.2 Tổng quan về tình hình bất động sản ở Mỹ những năm 2006 - 2020

- Trong giai đoạn 2006 - 2020, thị trường bất động sản Hoa Kỳ đã phải đối mặt với sự suy giảm đáng kể do cuộc khủng hoảng tài chính năm 2008, khi giá nhà giảm sút và nhiều người mất khả năng thanh toán thế chấp.
- Sau cuộc khủng hoảng, thị trường bắt đầu hồi phục dần và ghi nhận sự tăng trưởng từ giữa những năm 2010.
- Từ đó trở đi, thị trường bất động sản dần ổn định, có xu hướng tăng nhẹ cho đến năm 2020.

2.1.3 Quy trình nghiệp vụ



Hình 2.1: Quy trình nghiệp vụ

1. *Gửi hồ sơ*: Bên bán hàng sẽ gửi thông tin về bất động sản như vị trí, loại bất động sản, định giá, . . . cho bên môi giới. Bên môi giới sẽ tiếp nhận hồ sơ về tài sản và bắt đầu tìm kiếm khách hàng.

2. *Tiếp nhận hồ sơ* : Khi đã tìm được khách hàng có nhu cầu mua bất động sản như trong hồ sơ, khách hàng sẽ gửi hồ sơ bao gồm thông tin về thu nhập, kiểu gia đình của khách hàng và yêu cầu, mong muốn về giá cả, khả năng chi trả, . . . về tài sản và bên môi giới sẽ tiếp nhận hồ sơ đó.

3. *Phân tích hồ sơ*:Bên môi giới xem xét xem thu nhập và khả năng chi trả của khách hàng có phù hợp với các loại tài sản bên mình không. Lọc theo tất cả yêu cầu của khách hàng có những tài sản nào phù hợp, nếu không có hãy trao đổi lại với khách hàng và loại dần một số yêu cầu theo ý kiến của khách hàng sau khi đã thay đổi yêu cầu.

4. *Tiến hành xử lý yêu cầu*: Sau khi đã thống nhất được những yêu cầu cuối cùng của khách hàng, lọc lại một lần nữa những tài sản thỏa mãn yêu cầu, sắp xếp chúng theo thứ tự tốt nhất để đưa ra từng lựa chọn cho khách hàng.

5. *Đưa ra từng lựa chọn cho khách hàng*: Tư vấn và đưa ra thêm những điều kiện thuận lợi của tài sản phù hợp với yêu cầu khách hàng, nếu khách hàng đồng ý thì tiến hành làm hợp đồng và lưu hồ sơ khách hàng, chuyển hồ sơ và hợp đồng của khách hàng cho bên bán. Nếu khách hàng không đồng ý thì chuyển xuống những lựa chọn tiếp theo đến khi khách hàng quyết định.

6. Ngoài ra khi có cơ sở dữ liệu hồ sơ bán hàng trong một thời gian dài, doanh nghiệp sẽ dễ dàng hơn trong việc dự đoán xu thế mua bất động sản của khách hàng và sẽ có những phương án thay đổi, cải thiện về giá bất động sản sao cho phù hợp với nhiều đối tượng khách hàng nhất mà vẫn thu lại lợi nhuận cao.

2.1.4 Vấn đề hiện tại và yêu cầu hệ thống

Vấn đề hiện tại

Các hệ thống doanh nghiệp sử dụng cho việc phân tích dữ liệu thường là Google Sheet. Hệ thống hiện tại có chức năng chính là ghi nhận dữ liệu về sản phẩm, dịch vụ, khách hàng, vận hành nghiệp vụ xử lý khiếu nại của công ty nhưng chưa đáp ứng được các nhu cầu về phân tích dữ liệu, đặc biệt là phân tích dữ liệu nhiều chiều. Một số hạn chế của hệ thống cũ:

- Không thể tự động xuất ra các báo cáo mà cần thống kê một cách thủ công
- Thiếu tính nhất quán trong lưu trữ và đồng bộ dữ liệu, sai sót trong việc lưu trữ dữ liệu, dẫn đến báo cáo phân tích thiếu tính đúng đắn
- Các báo cáo chỉ ở dạng bảng, chưa được trực quan hóa bằng biểu đồ
- Chỉ thống kê dữ liệu, các báo cáo phân tích chưa xem được phân tích đa chiều

Yêu cầu hệ thống

Hệ thống phân tích dữ liệu mới cần:

- Độc lập với hệ thống xử lý giao dịch hiện tại của công ty mà chỉ có chức năng tổng hợp, phân tích dữ liệu từ cơ sở dữ liệu của hệ thống giao dịch
- Phân tích được dữ liệu theo nhiều chiều, trực quan, dữ liệu chính xác và đã được chuẩn hóa.
- Tự động làm mới báo cáo khi dữ liệu thay đổi

Và các báo cáo được xây dựng phục vụ cho một số yêu cầu dưới đây

2.1.5 Yêu cầu phân tích

◇ Tổng quan đánh giá bất động sản

◇ Tình hình bất động sản theo năm

◇ Xu hướng bất động sản theo khu vực

◇ Xu hướng mua bất động sản của khách hàng

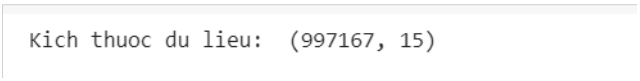
2.2 Thông tin bộ dữ liệu

2.2.1 Giới thiệu bộ dữ liệu

- Tên bộ dữ liệu: Dữ liệu về hồ sơ mua bán bất động sản tại Hoa Kỳ từ 2006 - 2020.
- Nguồn bộ dữ liệu: Bộ dữ liệu được lấy trên web Data.gov - Trang chủ dữ liệu mở của chính phủ Hoa Kỳ.
- Mô tả bộ dữ liệu: Văn phòng Chính sách và Quản lý duy trì danh sách tất cả các giao dịch mua bán bất động sản có giá bán từ 2000 USD trở lên diễn ra từ ngày 1 tháng 10 đến ngày 20 tháng 9 hàng năm. Đối với mỗi hồ sơ bán hàng, hồ sơ bao gồm: thị trấn, địa chỉ tài sản, ngày bán, loại tài sản (nhà ở, căn hộ, thương mại, đất công nghiệp hoặc đất trống), giá bán và đánh giá tài sản.

2.2.2 Kích thước dữ liệu

- Trước khi ETL



- Sau khi ETL

2.2.3 Các công cụ dùng lưu trữ và xử lý

- Microsoft Excel
- Công cụ lập trình Python
- Microsoft Power BI

STT	Tên trường	Mô tả	Kiểu dữ liệu
1	Serial_Number	Số seri của hồ sơ bán hàng	int
2	List_years	Năm bán trong hồ sơ bán hàng	int
3	Date_Recorded	Ngày bán trong hồ sơ bán hàng	date
4	Town	Tên thị trấn của tài sản trong hồ sơ	text
5	Address	Địa chỉ cụ thể của tài sản	text
6	Assessed_Value	Định giá tài sản	int
7	Sale_Amount	Giá bán của tài sản	int
8	Sales_Ratio	Tỷ lệ bán hàng của tài sản	decimal
9	Property_Type	Loại tài sản	text
10	Residential_Type	Loại nhà ở	text
11	Asset_classification	Phân loại tài sản	text
12	Installment period	Thời hạn trả góp tài sản	int
13	Income	Phân loại thu nhập của khách hàng	text
14	Rate	Đánh giá của khách hàng về mức độ hài	int
15	Income_Value	Thu nhập cụ thể của khách hàng	int

Hình 2.2: Thông tin các trường dữ liệu gốc

2.3 Kiến trúc hệ thống phân tích

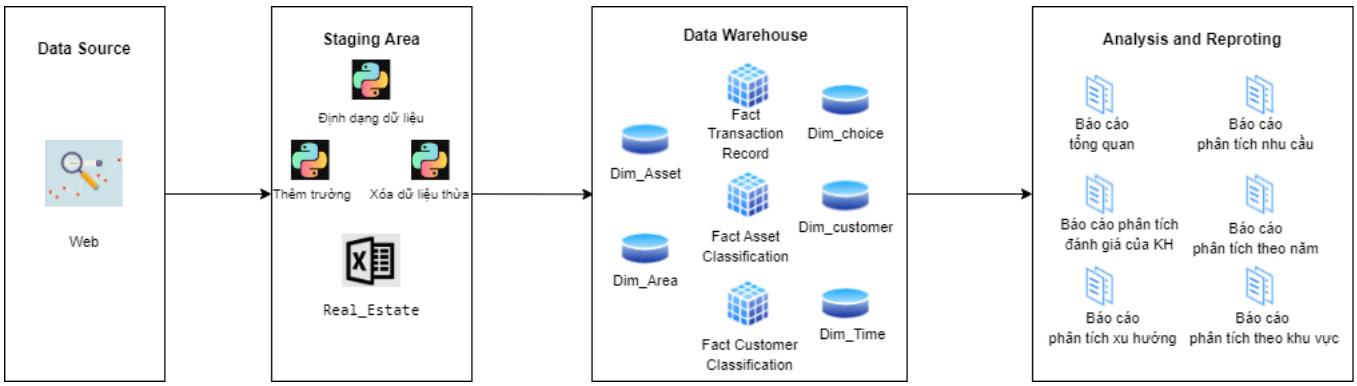
Xây dựng kiến trúc hệ thống phân tích là một công việc quan trọng để đảm bảo rằng dữ liệu được sử dụng một cách hiệu quả trong quá trình phân tích và đưa ra quyết định.

Các lợi ích của việc xây dựng kiến trúc hệ thống phân tích trong data bao gồm:

- Tăng hiệu quả phân tích dữ liệu: Một hệ thống phân tích dữ liệu được thiết kế tốt sẽ giúp cho các nhân viên phân tích dữ liệu có thể tiếp cận các nguồn dữ liệu một cách nhanh chóng và dễ dàng hơn. Họ có thể nhanh chóng tìm kiếm, truy vấn và chuyển đổi các loại dữ liệu khác nhau để phân tích.
- Cải thiện chất lượng dữ liệu: Khi xây dựng kiến trúc hệ thống, ta cần xác định và giải quyết các vấn đề về chất lượng dữ liệu. Có thể áp dụng các quy trình chuẩn hóa, rà soát và phân tích dữ liệu để cải thiện chất lượng dữ liệu.
- Tăng tính linh hoạt: Khi hệ thống phân tích dữ liệu được xây dựng tốt, nó sẽ cho phép người dùng tiếp cận và sử dụng các nguồn dữ liệu từ nhiều nguồn khác nhau. Điều này giúp cho ta có thể kết hợp và phân tích các nguồn dữ liệu khác nhau để đưa ra quyết định.

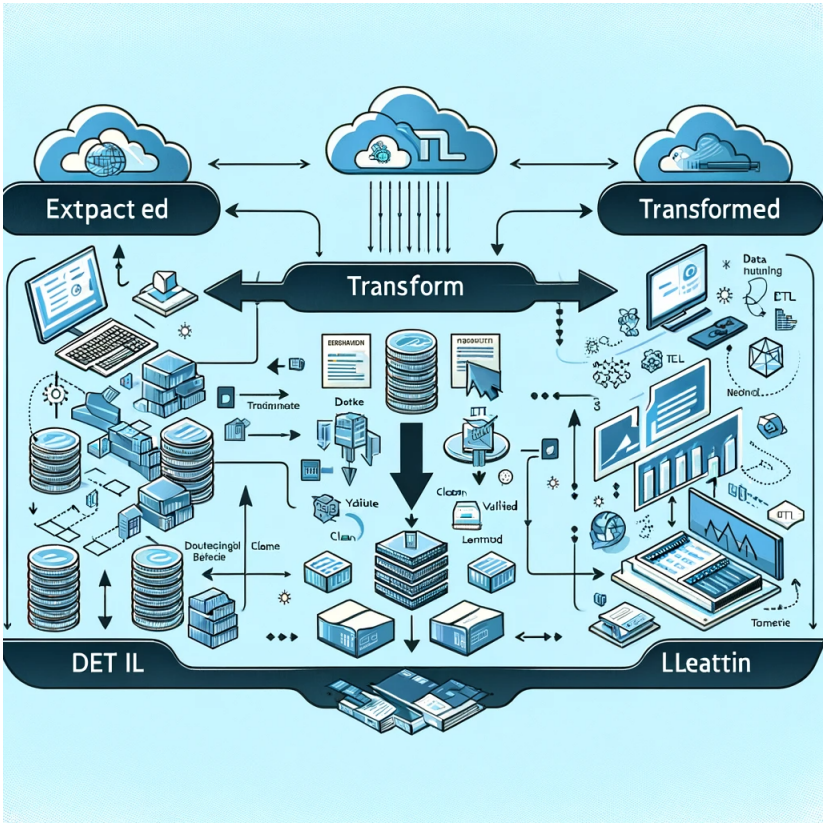
- Tiết kiệm thời gian và chi phí: Ta có thể tối ưu hóa quy trình phân tích để tiết kiệm thời gian và chi phí bằng cách dụng các công cụ tự động hóa để giảm thiểu thời gian và công sức cần thiết cho các tác vụ phân tích dữ liệu.

Từ đặc điểm của bộ dữ liệu và phục vụ cho quá trình báo cáo, phân tích dữ liệu, em xây dựng một kiến trúc hệ thống như sau:



Hình 2.3: Kiến trúc hệ thống phân tích

- Data Source (Nguồn dữ liệu): Dữ liệu được lấy từ web Consumer Financial Protection Bureau (CFPB) - Cục bảo vệ tài chính người tiêu dùng.
- Staging Area (Khu vực tập kết dữ liệu): Ở đây, em sẽ tiến hành xử lý dữ liệu thô bằng cách tiến hành xử lý dữ liệu (ETL dữ liệu). Công cụ giúp xử lý dữ liệu chính là Python.
- Data Warehouse (Kho dữ liệu): chịu trách nhiệm lưu trữ dữ liệu đã được xử lý sạch để phục vụ cho quá trình phân tích bằng PowerBI. Bằng cách phân tích các dim và các fact, dữ liệu sẽ được lưu trữ dưới dạng OLAP phục vụ cho quá trình truy vấn và báo cáo trở nên dễ dàng hơn.
- Analysis and Reporting (Phân tích và báo cáo): Cho phép người dùng trực quan hóa dữ liệu (data visualization), phân tích dữ liệu (data analysis), cũng như xây dựng các dashboard khác nhau bằng PowerBI.



Hình 2.4: Quy trình xử lý dữ liệu

2.4 Xử lý dữ liệu (ETL)

2.4.1 Quy trình xử lý dữ liệu

Quy trình xử lý dữ liệu (ETL) là việc kết hợp ba bước chính: Extract (Trích xuất), Transform (Chuyển đổi), và Load (Tải).Trong đó:

- Extract (Trích xuất): Bước đầu tiên trong quy trình ETL là trích xuất dữ liệu từ nguồn cơ sở dữ liệu.
- Transform (Chuyển đổi): Sau khi dữ liệu được trích xuất, nó được chuyển đổi thành định dạng cần thiết cho phân tích tiếp theo hoặc lưu trữ. Bước này có thể bao gồm việc làm sạch dữ liệu (loại bỏ dữ liệu bị lỗi hoặc không đầy đủ), định dạng lại (thay đổi định dạng dữ liệu), enriching (bổ sung thêm thông tin), và aggregating (tổng hợp dữ liệu).
- Load (Tải): Cuối cùng, dữ liệu được tải vào kho dữ liệu đích, nơi dữ liệu sẽ được lưu trữ hoặc dùng cho việc phân tích.

Quy trình ETL cần được tự động hóa và tối ưu để đảm bảo dữ liệu luôn được cập nhật

và chính xác. Nó là một phần thiết yếu của quản lý dữ liệu trong doanh nghiệp và cần thiết cho việc ra quyết định dựa trên dữ liệu.

2.4.2 Các thao tác ETL trên bộ dữ liệu

◇ **Trích xuất dữ liệu**

```
from google.colab import drive
drive.mount('/content/drive')
import pandas as pd
df=pd.read_csv('/content/drive/My Drive/Colab
   Notebooks/Project2/data.csv')
```

◇ **Hiển thị 5 bản ghi đầu tiên của dữ liệu**

```
df.head()
```

	Serial Number	List Year	Date Recorded	Town	Address	Assessed Value	Sale Amount	Sales Ratio	Property Type	Residential Type	Asset classification	Installment period	income	Rate	income value
0	RE062815	2002	10/21/2002	Andover	16 CHESTER BROOKS LN	203,300	340,912	60%	NaN	NaN	Medium	21	Medium	4	54858
1	RE063785	2002	10/28/2002	Andover	68 STANLEY DR	161,500	262,000	62%	NaN	NaN	Medium	12	High	5	62781
2	RE064144	2002	10/30/2002	Andover	327 HEBRON RD	121,400	230,000	53%	NaN	NaN	Low	5	Medium	4	33713
3	RE064570	2002	10/31/2002	Andover	65 JUROVATY RD	77,200	138,780	56%	NaN	NaN	Low	9	Medium	5	31510
4	RE065795	2002	11/5/2002	Andover	81 STANLEY DR	165,500	277,500	60%	NaN	NaN	Medium	15	High	4	68543

Hình 2.5: Hiển thị 5 bản ghi đầu tiên

◇ **Xóa các dòng có ít nhất 1 giá trị Null**

```
data = df.dropna()
```

◇ **Đổi tên các trường thành kí tự viết liền**

```
data.rename(columns={'Serial Number':
   'Serial_Number'},inplace=True)
data.rename(columns={'List Year':
   'List_Year'},inplace=True)
data.rename(columns={'Date Recorded':
   'Date_Recorded'},inplace=True)
data.rename(columns={'Assessed
   Value':'Assessed_Value'},inplace=True)
data.rename(columns={'Sale Amount': 'Sale_Amount'},
   inplace=True)
data.rename(columns={'Sales Ratio': 'Sales_Ratio'},
   inplace=True)
data.rename(columns={'Property Type': 'Property_Type'},
   inplace=True)
data.rename(columns={'Residential Type':
   'Residential_Type'}, inplace=True)
data.rename(columns={'Asset classification':
   'Asset_classification'}, inplace=True)
data.rename(columns={'Installment period':
   'Installment_period'}, inplace=True)
data.rename(columns={'income value': 'income_value'},
   inplace=True)
```

◇ **Định dạng ngày tháng năm về dạng chuẩn**

```
data['Date_Recorded'] =
   pd.to_datetime(data['Date_Recorded'])
```

Hình 2.6: Định dạng thời gian theo dạng chuẩn

◇ **Định dạng lại kiểu số**

```
if data['Sale_Amount'].dtypes == 'object':
   data['Sale_Amount'] =
   data['Sale_Amount'].str.replace(',', '').astype(float)
if data['Assessed_Value'].dtypes == 'object':
   data['Assessed_Value'] =
   data['Assessed_Value'].str.replace(',',
   '').astype(float)
```

◇ **Xóa bản ghi có giá trị cột "Sale_Amount" = 0**

```
data = data[data['Sale_Amount'] != 0]
```

◇ **Thêm trường dữ liệu "Area_ID "**

Lấy ra danh sách các địa điểm trong bộ dữ liệu, tạo ID của mỗi địa điểm bằng Excel và thêm vào bộ dữ liệu ban đầu.

```
data_khu_vuc=data.groupby('Town'
   ['Sale_Amount'].mean().reset_index()
data_khu_vuc.to_csv('data_khu_vuc.csv', index=False)
files.download('data_khu_vuc.csv')
```

Số địa điểm sau khi tách ra Tạo Area_ID và thêm vào bộ dữ liệu

	Town
0	***Unknown***
1	Andover
2	Ansonia
3	Ashford
4	Avon
..	...
165	Windsor Locks
166	Wolcott
167	Woodbridge
168	Woodbury
169	Woodstock

```
data_town=pd.read_csv('/content/drive/My Drive/Colab
   Notebooks/Project2/data_khu_vuc.csv')
data_town.drop('Sale Amount', axis=1, inplace=True)
data = pd.merge(data, data_town, on='Town')
```

◇ **Xuất file csv sau khi đã ETL xong**

```
import pandas as pd
from google.colab import files
data.to_csv('data_sauetl.csv', index=False)
files.download('data_sauetl.csv')
```

Serial_Number	List_Year	Date_Recorded	Town	Address	Assessed_Value	Sale_Amount	Sales_Ratio	Property_Type	Residential_Type	Asset_classification	Installment_period	income	Rate	income_value	Area_ID
RE378447	2007	2007-10-02	Andover	149 HEBRON AVE	138400	217500	64	Single Family	Single Family	Low	9	Medium	5	32528	TOWN_002
RE379969	2007	2007-10-17	Andover	65 LAKE RD	213800	345000	62	Single Family	Single Family	Medium	13	High	3	63996	TOWN_002
RE380431	2007	2007-10-22	Andover	28 OLD COVENTRY RD	218500	299900	73	Single Family	Single Family	Medium	13	High	5	79846	TOWN_002
RE381156	2007	2007-10-29	Andover	10 CHESTER BROOKS LN	313400	425000	74	Single Family	Single Family	Medium	19	Medium	5	57958	TOWN_002
RE381409	2007	2007-10-30	Andover	191 LONG HILL RD	181700	267000	68	Single Family	Single Family	Medium	27	Low	4	26964	TOWN_002
RE383576	2007	2007-11-19	Andover	22 WOOD FERN WAY	68700	410930	17	Single Family	Single Family	Medium	23	Medium	5	42305	TOWN_002
RE384255	2007	2007-11-28	Andover	345 HEBRON RD	198900	282000	71	Single Family	Single Family	Medium	11	High	1	60219	TOWN_002
RE384845	2007	2007-12-03	Andover	147 WHEELING RD	167200	177100	94	Single Family	Single Family	Low	10	Low	2	28210	TOWN_002
RE385793	2007	2007-12-11	Andover	163 LAKE RD	126500	148000	85	Single Family	Single Family	Low	11	Low	5	28400	TOWN_002
RE386404	2007	2007-12-17	Andover	62 HENDEE RD	238400	371000	64	Single Family	Single Family	Medium	25	Low	3	26023	TOWN_002
RE387639	2007	2007-12-31	Andover	51 WINDRUSH LN	214600	231000	93	Single Family	Single Family	Medium	12	High	3	73132	TOWN_002
RE389610	2007	2008-01-28	Andover	587 ROUTE 6	147200	170000	87	Single Family	Single Family	Low	10	Low	2	27059	TOWN_002
RE391718	2007	2008-02-26	Andover	76 ROUTE 6	216400	300000	72	Single Family	Single Family	Medium	22	Medium	3	48461	TOWN_002
RE392407	2007	2008-03-04	Andover	30 ROUTE 6	114900	179770	64	Single Family	Single Family	Low	5	Medium	4	52094	TOWN_002
RE393207	2007	2008-03-13	Andover	4 SHADBLOW LN	79100	391500	20	Single Family	Single Family	Medium	14	High	5	78021	TOWN_002

Hình 2.7: Dữ liệu sau khi ETL

2.5 Khai phá dữ liệu

2.5.1 Phân tích đặc trưng của dữ liệu bất động sản

Phân tích đặc trưng của bộ dữ liệu bất động sản là một quá trình quan trọng để hiểu rõ thông tin liên quan và đưa ra những quyết định có ý nghĩa. Bộ dữ liệu bất động sản thường chứa đựng nhiều thông tin quan trọng, trong đó có các đặc trưng như:

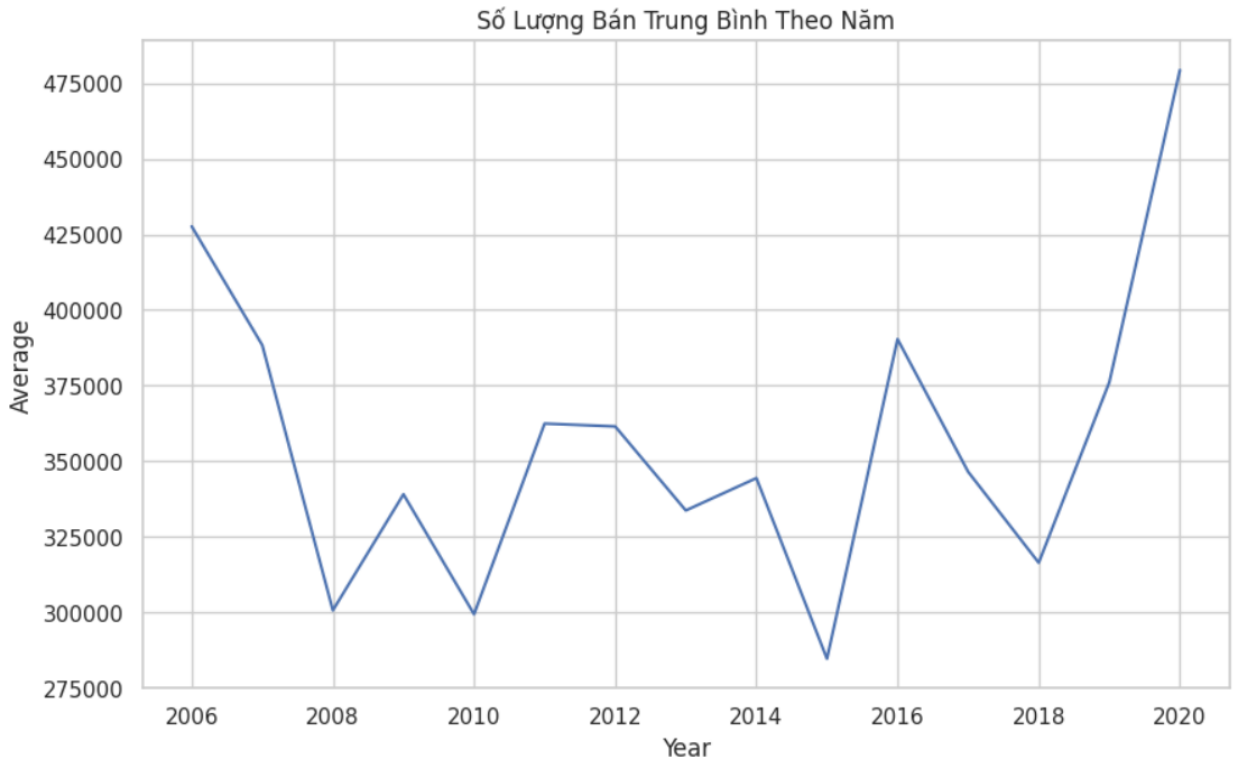
- Loại bất động sản: Mô tả về loại hình tài sản, bao gồm nhà ở, đất đai, căn hộ chung cư, ...
- Giá trị bất động sản: Thông tin về định giá hoặc giá bán của bất động sản.
- Vị trí địa lý: Thông tin về địa điểm cụ thể của bất động sản.

Quá trình phân tích đặc trưng của bộ dữ liệu bất động sản giúp định rõ các yếu tố ảnh hưởng đến giá trị và tính khả dụng của các tài sản. Điều này giúp cung cấp thông tin hữu ích cho những người quan tâm đến thị trường bất động sản, từ nhà đầu tư đến những người muốn mua nhà hoặc đang tìm kiếm chỗ ở mới.

2.5.2 Xây dựng biểu đồ

Từ bộ dữ liệu đã làm sạch, em đi vào khai thác và phân tích tình hình bất động sản Hoa Kỳ trong khoảng năm 2006 - 2020 như sau theo một số chiều như sau:

1. Số lượng bán trung bình theo từng năm
2. Số lượng bán trung bình theo tháng

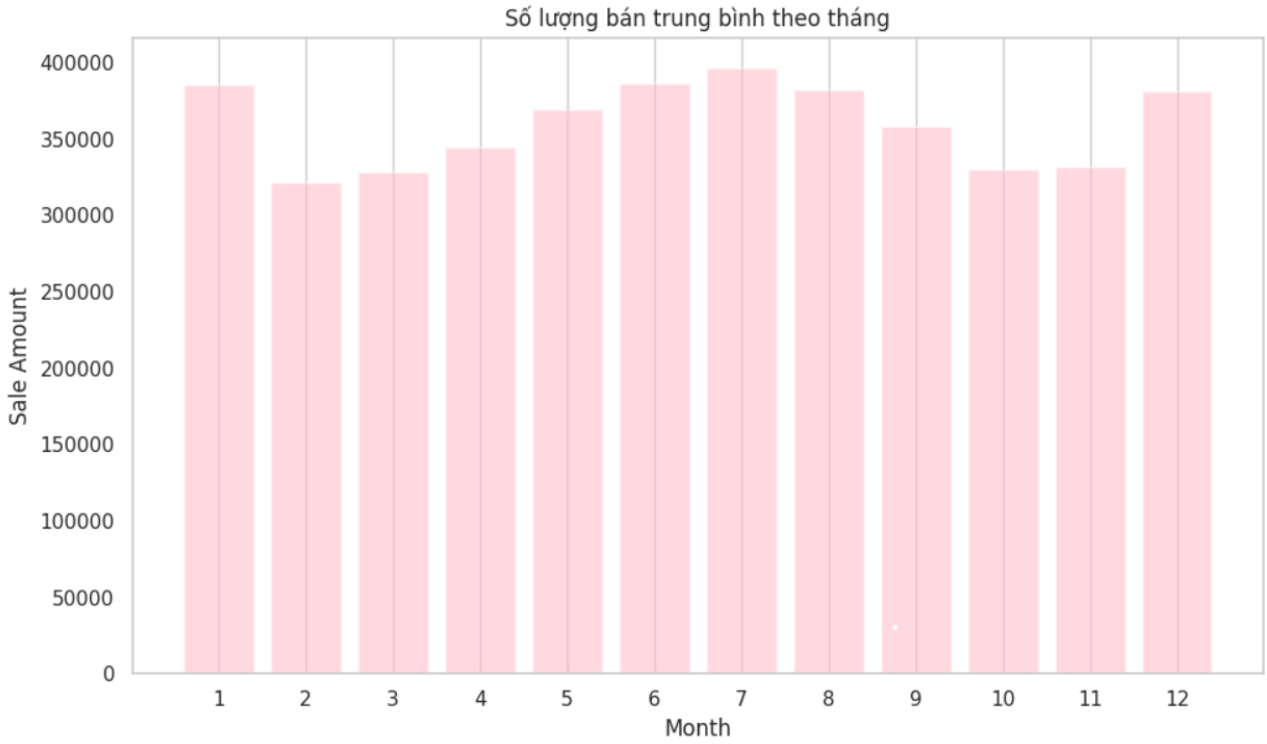


Hình 2.8: Biểu đồ lượng bán trung bình theo năm

Doanh thu bán trung bình theo từng năm

- Nhìn vào biểu đồ ta có thể dễ dàng thấy được sự thay đổi của tình hình bất động sản tại Hoa Kỳ trong chuỗi thời gian 2006 - 2020.
- Khoảng năm 2006 - 2009, trung bình giá bất động sản có xu hướng giảm hơn. Điều này phản ánh sự suy thoái của thị trường bất động sản, dẫn đến cuộc khủng hoảng tài chính toàn cầu bắt đầu vào năm 2007 và đạt đến đỉnh điểm vào năm 2008 - 2009.
- Từ khoảng năm 2010 - 2020 có sự biến động nhất định nhưng nhìn chung tổng thể cho thấy xu hướng tăng. Từ năm 2010 trở đi, có vẻ như thị trường bắt đầu phục hồi và giá bán trung bình bắt đầu tăng lên. Điều này có thể phản ánh sự phục hồi của nền kinh tế Hoa Kỳ, cũng như các biện pháp kích thích của chính phủ và Cục Dự trữ Liên bang.

Doanh thu bán trung bình theo tháng



Hình 2.9: Biểu đồ lượng bán trung bình theo tháng

Từ biểu đồ ta có thể nhìn thấy rõ rằng :

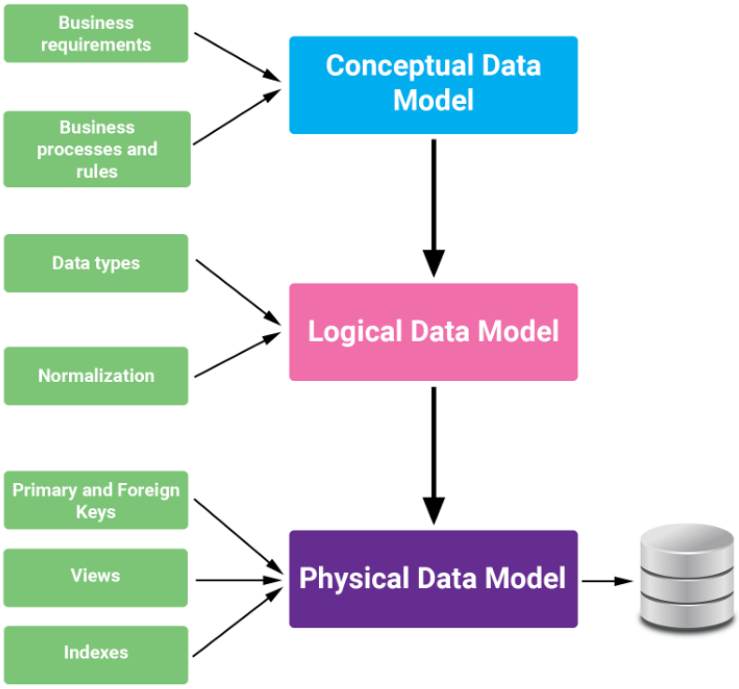
- Tháng bán được nhiều nhất là tháng 7 và tháng 12, tháng bán được ít hơn cả là tháng 2 và tháng 9.
- Có thể thấy thị trường bất động sản cũng theo mùa vụ hoặc cũng có bị ảnh hưởng bởi thị trường kinh tế vì sự chênh lệch giữa các tháng là đáng kể.

2.6 Xây dựng mô hình dữ liệu

Mô hình dữ liệu cung cấp những thông số về dữ liệu, thuộc tính, mối quan hệ hoặc liên kết với dữ liệu khác. Hiểu một cách đơn giản, mô hình dữ liệu cung cấp cho người dùng cái nhìn tổng quan nhất về dữ liệu đại diện cho kịch bản và dữ liệu nghiệp vụ.

Như vậy, việc xây dựng mô hình dữ liệu là quá trình phân tích, chọn lọc và biến đổi dữ liệu thành các đặc trưng (features) có tính chất hữu ích và có thể được sử dụng để

dự đoán hoặc phân loại một dữ liệu mới.



Hình 2.10: Mô hình dữ liệu

Mô hình hóa dữ liệu là quá trình tạo mô hình dữ liệu. Trước tiên, chúng ta phải xác định dữ liệu, các thuộc tính và mối quan hệ của nó với dữ liệu khác và xác định các ràng buộc hoặc hạn chế đối với dữ liệu.

Để giúp cho việc xây dựng mô hình dữ liệu trở nên dễ dàng hơn, em sẽ đi vào xác định và phân tích các dimension và fact table.

2.6.1 Phân tích các dim và fact

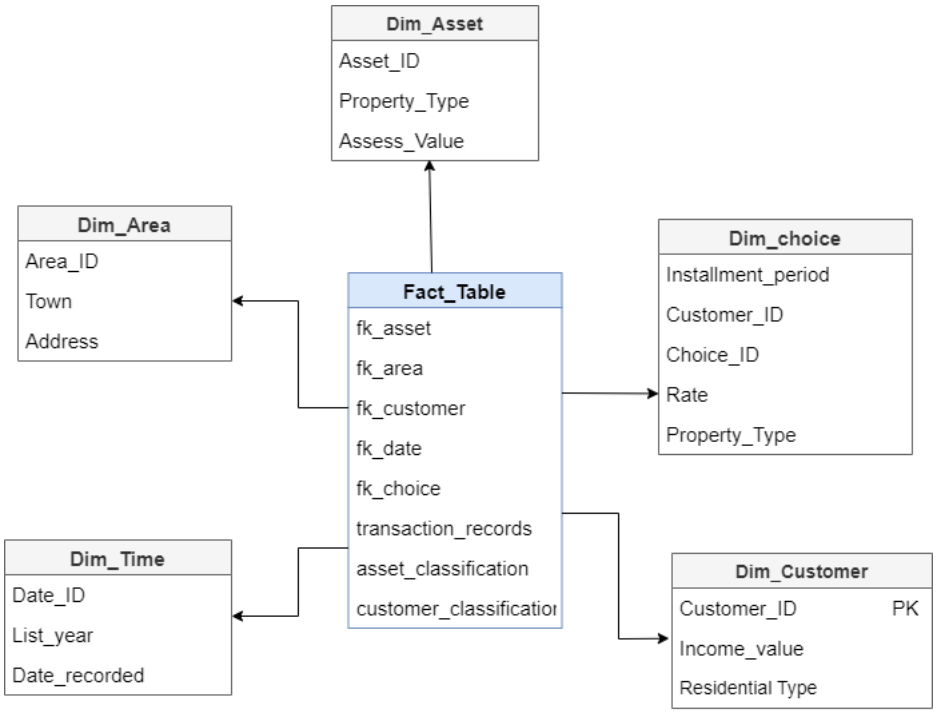
Trong xây dựng mô hình dữ liệu, hai khái niệm quan trọng đó là Dimension Table và Fact Table.

- Dimension table (dim) là bảng dữ liệu được thiết kế dựa trên các chiều và các chỉ số để đại diện cho các thông tin. Mỗi chiều có thể là các thuộc tính của một thực thể.
- Fact table (fact) là bảng dữ liệu được thiết kế để lưu trữ thông tin chi tiết về các sự kiện hoặc các giao dịch trong hệ thống. Mỗi hàng trong bảng fact đại diện cho

một giao dịch hoặc một sự kiện và chứa thông tin về các chỉ số và các khóa ngoại đến các bảng dimension liên quan.

Như vậy, sự phân chia rõ ràng giữa Dimension table và Fact table giúp cho việc tìm kiếm và truy xuất dữ liệu trở nên nhanh chóng và hiệu quả hơn, đồng thời cũng giúp cho việc xử lý và phân tích dữ liệu trở nên dễ dàng và linh hoạt hơn. Do đó, việc xây dựng mô hình dữ liệu có sử dụng cả dim và fact table là rất quan trọng trong phân tích dữ liệu và xây dựng các dashboard để hỗ trợ quản lý và ra quyết định.

Từ bộ dữ liệu Real_Estate sau khi đã chuẩn hóa, em xây dựng lược đồ dữ liệu mua bán bất động sản gồm 5 dimension được kết nối với bảng fact như sau:



Hình 2.11: Lược đồ Dim và Fact Table

Khóa đo lường là một hàm được tính toán trên tất cả các bảng thứ nguyên trả về số vấn đề khiếu nại sau khi áp dụng truy vấn OLAP phức tạp. Điểm mấu chốt đằng sau phản hồi nhanh của truy vấn OLAP là sử dụng bảng thứ nguyên chứa khái niệm phân cấp. Khái niệm này gần như xây dựng kích thước như địa chỉ cho phép thao tác OLAP dễ dàng. Có nhiều lợi thế khi sử dụng lược đồ hình sao chẳng hạn như phản hồi nhanh của truy vấn OLAP, xử lý các thay đổi của kích thước theo thời gian, cho phép nhiều thứ bậc cho các kích thước và lược đồ dễ dàng và đơn giản để xây dựng và hiểu.

1. Mô tả các chiều dữ liệu:

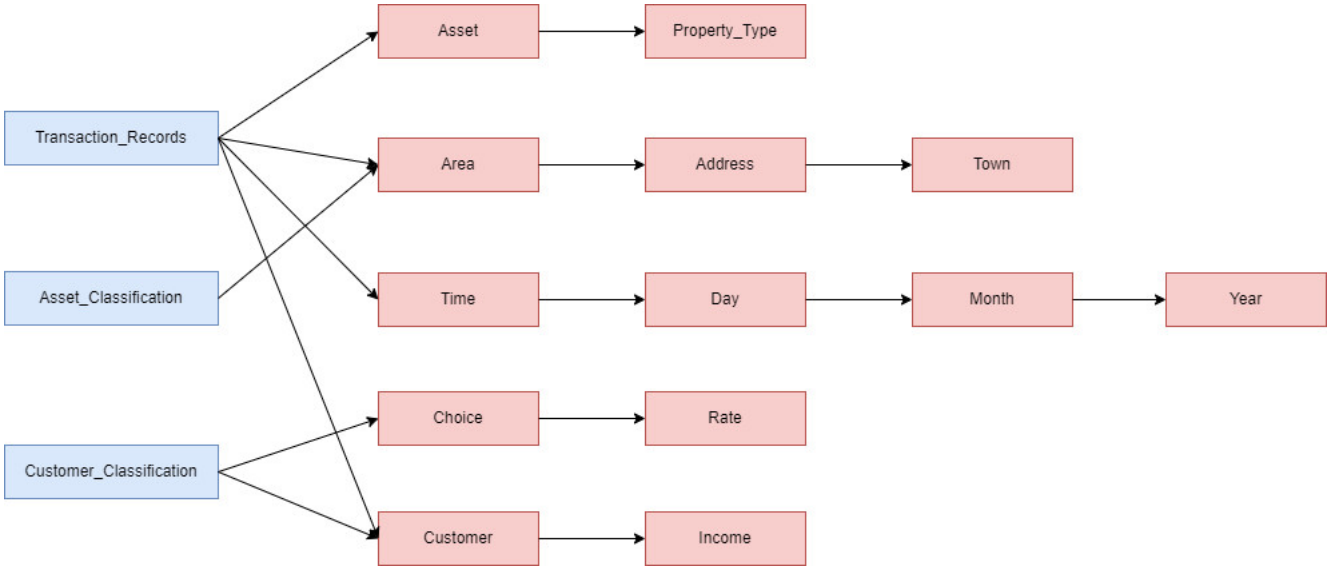
    - Dim_Time: chứa thông tin về thời gian thực hiện giao dịch của mỗi bộ hồ sơ, bao gồm ngày, tháng, năm.
    - Dim_Asset: chứa thông tin về bất động sản.
    - Dim_Area : chứa thông tin vị trí các bất động sản.
    - Dim_Choice: chứa thông tin về lựa chọn của khách hàng về kiểu bất động sản, thời gian trả góp bất động sản và đánh giá về bất động sản đó.
    - Dim_customer: chứa thông tin về khách hàng.

2. Mô tả các chủ điểm phân tích:

    - Fact_transacion_records : chứa thông tin của mỗi bộ hồ sơ theo các chiều khác nhau.
    - Fact_Asset_Classification: Tổng quan về tài sản bất động sản theo các chiều khác nhau.
    - Fact_Customer_Classification: Tổng quan về khách hàng theo các chiều khác nhau

2.6.2 Mô hình dữ liệu logic

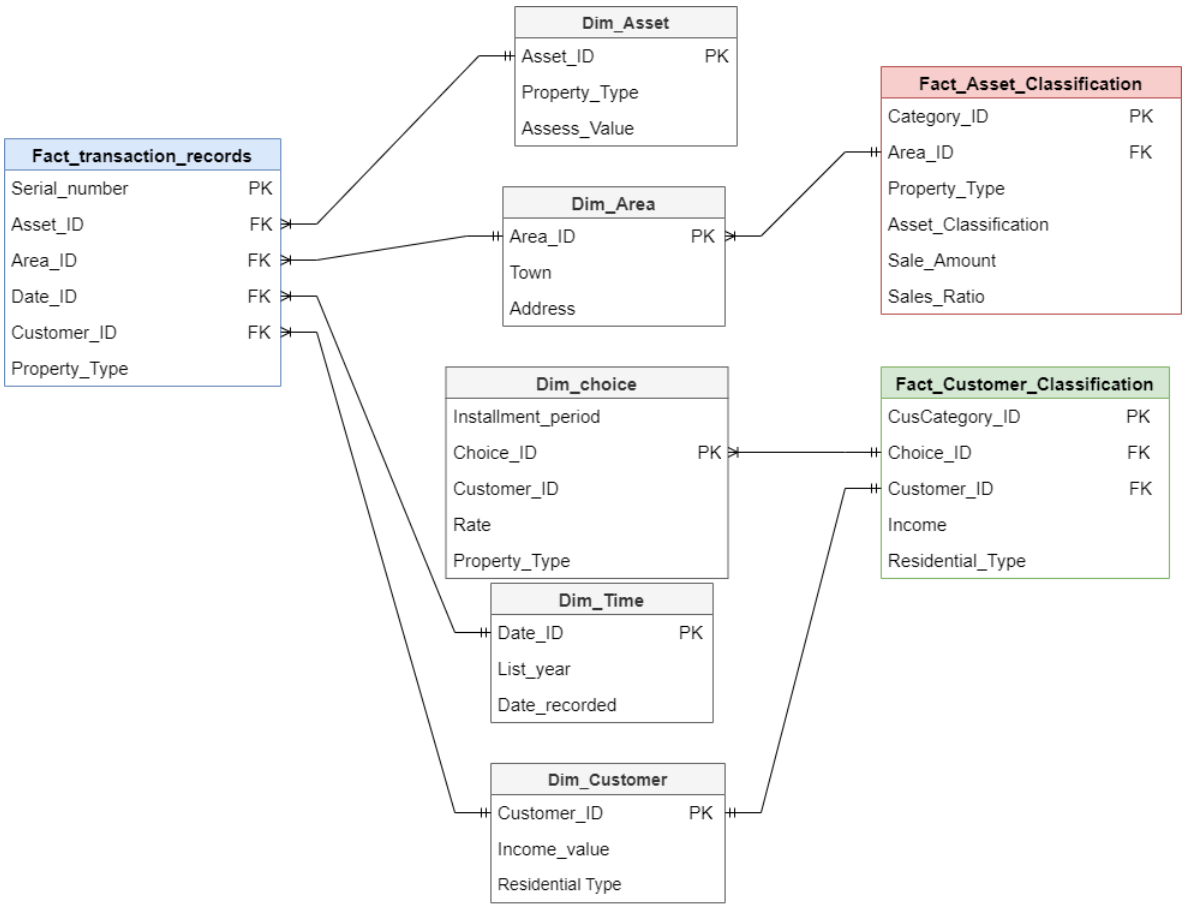
Mô hình dữ liệu logic được sử dụng để xác định cấu trúc các yếu tố dữ liệu và thiết lập mối quan hệ giữa chúng. Mô hình dữ liệu logic cung cấp thông tin bổ sung cho các yếu tố mô hình dữ liệu khái niệm. Lợi ích của việc sử dụng mô hình dữ liệu logic là cung cấp nền tảng để tạo ra mô hình dữ liệu vật lý. Tuy nhiên, cấu trúc mô hình vẫn giữ nguyên tính chất chung chung. Kết hợp các chiều, các chủ điểm phân tích, em thiết kế được mô hình logic như sau:



Hình 2.12: Mô hình dữ liệu logic

2.6.3 Mô hình dữ liệu vật lý

Em xây dựng lược đồ hình sao dựa trên mô hình dim table và fact table. Khi sử



Hình 2.13: Lược đồ OLAP

dụng lược đồ hình sao, hệ thống kho dữ liệu có thể đáp ứng được nhu cầu phân tích của người dùng với phản hồi nhanh của truy vấn OLAP. Hơn nữa, lược đồ hình sao cũng hỗ trợ xử lý các thay đổi của kích thước theo thời gian, cho phép nhiều thứ bậc cho các kích thước và dễ dàng và đơn giản để xây dựng và hiểu, qua đó cải thiện tính khả dụng của hệ thống và tối ưu hóa hiệu suất phân tích dữ liệu.

2.7 Thiết kế hệ thống kho dữ liệu

Các bảng trong Cơ sở dữ liệu kho dữ liệu (Data Warehouse) được thiết kế như sau:

Tên bảng	**Tên trường**	**Kiểu dữ liệu**	**Mô tả**
Fact_ transaction_ records	Serial_ number (PK)	text	Mã của mỗi bộ hồ sơ bất động sản
	Asset_ ID (FK)	int	Mã bất động sản
	Area_ ID (FK)	int	Mã khu vực
	Date_ ID (FK)	int	Mã thời gian của hồ sơ
	Customer _ ID (FK)	int	Mã khách hàng
	Property _ Type	text	Loại bất động sản
Fact_ Asset_ Classification	Category_ ID (PK)	int	Mã phân loại tài sản
	Area_ ID (FK)	int	Mã khu vực
	Property _ Type	text	Loại bất động sản
	Asset_ Classification	text	Phân loại bất động sản
	Sale_ Amount	int	Giá bán của bất động sản
	Sale_ Ratio	int	Tỷ lệ bán hàng của bất động sản
Fact_ Customer_ Classification	CusCategory_ ID (PK)	int	Mã phân loại khách hàng
	Choice_ ID (FK)	int	Mã theo lựa chọn của khách hàng
	Customer _ ID (FK)	int	Mã khách hàng
	income	text	Phân loại thu nhập của khách hàng
	Residential_ Type	text	Phân loại sử dụng bất động sản
Dim_ Asset	Asset_ ID (FK)	int	Mã bất động sản
	Property _ Type	text	Loại bất động sản
	Asset_ Value	int	Định giá tài sản
Dim_ Area	Area_ ID (FK)	int	Mã khu vực
	Town	text	Tên của thị trấn trong hồ sơ
	Address	text	Địa chỉ cụ thể của bất động sản

Dim_ Choice	Choice_ ID (FK)	int	Mã theo lựa chọn của khách hàng
	Customer _ ID	int	Mã khách hàng
	Installment_ Period	int	Thời gian trả góp tài sản
	Rate	int	Đánh giá của khách hàng về mức độ hài lòng
	Property _ Type	text	Loại bất động sản
Dim_ Time	Date_ ID (FK)	int	Mã thời gian của hồ sơ
	List_ year	int	thời gian về năm trong hồ sơ
	Date_ recorded	varchar(10)	Thời gian về ngày, tháng, năm trong hồ sơ
Dim_ Customer	Customer _ ID (FK)	int	Mã khách hàng
	Income_ Value	int	Thu nhập cụ thể của khách hàng
	Residential_ Type	text	Phân loại sử dụng bất động sản

Chương 3

Thực nghiệm

3.1 Xây dựng tầng khu vực tập kết dữ liệu

◇ **Tạo database**

```
CREATE Database StagingArea ;
```

◇ **Tạo các trường dữ liệu, các table trong dữ liệu**

```
CREATE TABLE Real_Estate(
 Serial_Number VARCHAR(255),
 List_years INT,
 Date_Recorded DATETIME,
 Town VARCHAR(255),
 Address VARCHAR(255),
 Assessed_Value INT,
 Sale_Amount INT,
 Sales_Ratio INT,
 Property_Type VARCHAR(255),
 Residential_Type VARCHAR(255),
 Asset_classification VARCHAR(255),
 Installment_period INT,
 Income VARCHAR(255),
 Rate INT,
```

```
Income_Value INT,
Area_ID VARCHAR(255),
PRIMARY KEY(Serial_Number)
);
```

◇ **Đổ dữ liệu vào StagingArea**

```
LOAD DATA INFILE "C:\\ProgramData\\MySQL\\MySQL Server
   8.0\\Uploads\\Real_Estate.csv"
INTO TABLE Real_Estate
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;
```

3.2 Xây dựng tầng kho dữ liệu

◇ **Tạo database OLAPDATA**

```
CREATE DATABASE OlapData;
```

◇ **Tạo các bảng Dimension**

```
CREATE TABLE Dim_Asset (
    Asset_ID INT NOT NULL PRIMARY KEY,
    Property_Type VARCHAR(255),
    Asessed_Value INT
);
CREATE TABLE Dim_Area (
    Area_ID INT NOT NULL PRIMARY KEY,
    Town VARCHAR(255),
    Address VARCHAR(255)
);
CREATE TABLE Dim_Choice (
    Choice_ID INT NOT NULL PRIMARY KEY,
    Installment_period INT,
    Customer_ID INT,
    Rate INT,
    Property_Type VARCHAR(255)
);
CREATE TABLE Dim_Time (
    Date_ID INT NOT NULL PRIMARY KEY,
    List_years INT,
    Date_recorded DATETIME
);
CREATE TABLE Dim_Customer (
    Customer_ID INT NOT NULL PRIMARY KEY,
    Income_Value INT,
    Residential_Type VARCHAR(255)
);
```

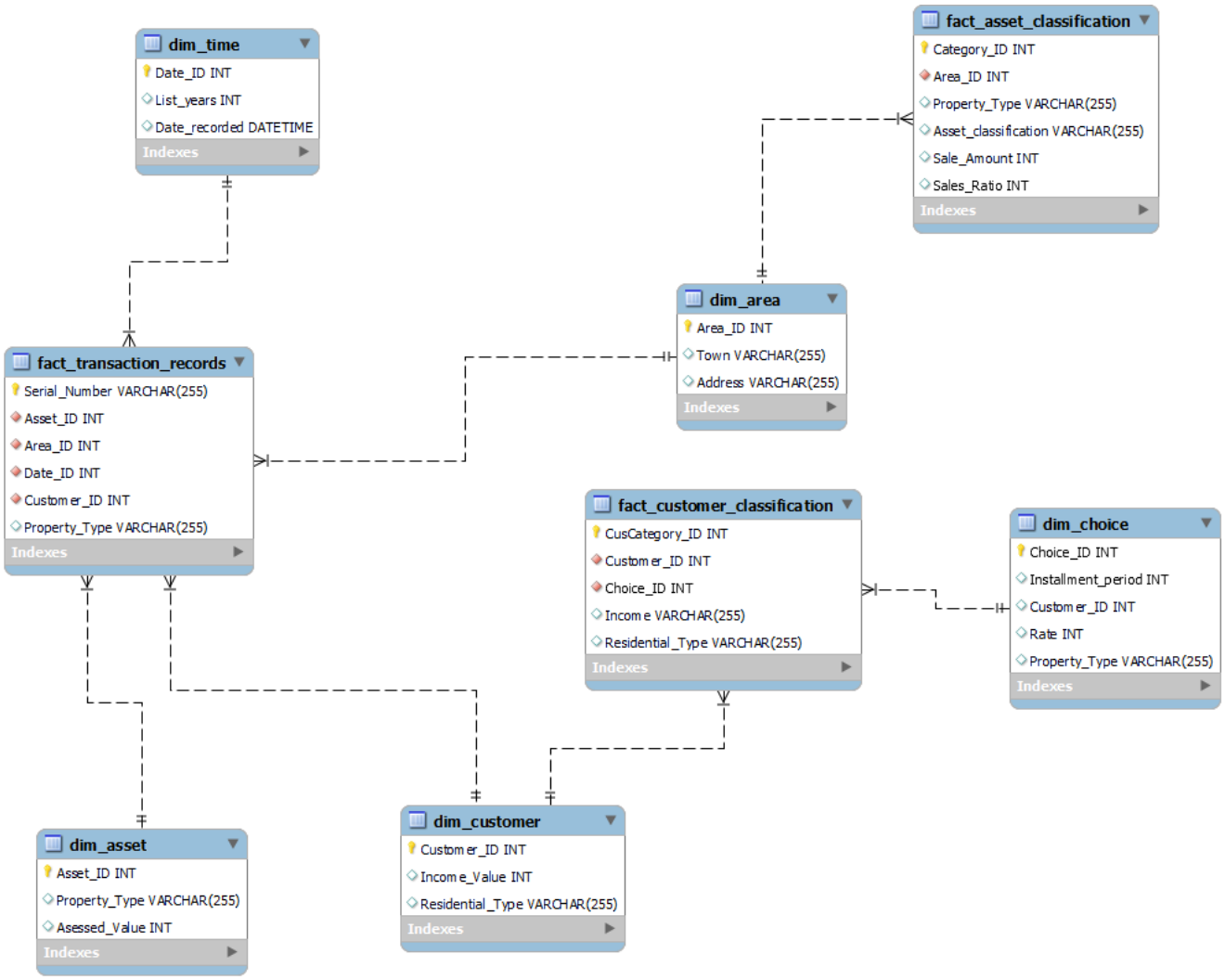
◇ **Tạo các bảng Fact**

```
CREATE TABLE Fact_transaction_records (
    Serial_Number VARCHAR(255) PRIMARY KEY,
    Asset_ID INT NOT NULL REFERENCES Dim_Asset (Asset_ID),
    Area_ID INT NOT NULL REFERENCES Dim_Area (Area_ID),
    Date_ID INT NOT NULL REFERENCES Dim_Time (Date_ID),
    Customer_ID INT NOT NULL REFERENCES Dim_Customer
   (Customer_ID),
    Property_Type VARCHAR(255)
    );
CREATE TABLE Fact_Asset_Classification (
    Category_ID INT NOT NULL PRIMARY KEY,
    Area_ID INT NOT NULL REFERENCES Dim_Area (Area_ID),
    Property_Type VARCHAR(255),
    Asset_classification VARCHAR(255),
  Sale_Amount INT,
  Sales_Ratio INT
);
CREATE TABLE Fact_customer_classification (
    CusCategory_ID INT NOT NULL PRIMARY KEY,
    Customer_ID INT NOT NULL REFERENCES Dim_Customer
   (Customer_ID),
    Choice_ID INT NOT NULL REFERENCES Dim_Choice (
   Choice_ID),
    Income VARCHAR(255),
    Residential_Type VARCHAR(255)
);
```

◇ **Tạo kết nối giữa các bảng Fact và Dim**

```
ALTER TABLE fact_transaction_records
ADD CONSTRAINT FK_Dim_Asset_transaction
FOREIGN KEY (Asset_ID)
REFERENCES Dim_Asset (Asset_ID);

ALTER TABLE fact_transaction_records
ADD CONSTRAINT FK_Dim_Area_transaction
FOREIGN KEY (Area_ID)
REFERENCES Dim_Area (Area_ID);

ALTER TABLE fact_transaction_records
ADD CONSTRAINT FK_Dim_Time_transaction
FOREIGN KEY (Date_ID)
REFERENCES Dim_Time (Date_ID);

ALTER TABLE fact_transaction_records
ADD CONSTRAINT FK_Dim_Customer_transaction
FOREIGN KEY (Customer_ID)
REFERENCES Dim_Customer (Customer_ID);

ALTER TABLE fact_asset_classification
ADD CONSTRAINT FK_Dim_Area_asset
FOREIGN KEY (Area_ID)
REFERENCES Dim_area (Area_ID);

ALTER TABLE fact_customer_classification
ADD CONSTRAINT FK_Dim_Customer_customer
FOREIGN KEY (Customer_ID)
REFERENCES Dim_customer (Customer_ID);

ALTER TABLE fact_customer_classification
ADD CONSTRAINT FK_Dim_Choice_customer
```

```
FOREIGN KEY (Choice_ID)
REFERENCES Dim_Choice (Choice_ID);
```



Hình 3.1: Hệ thống OLAP

◇ **Đổ dữ liệu từ Staging vào OLAP**

```
DELIMITER //
CREATE PROCEDURE TransferDataToOlap()
BEGIN
INSERT IGNORE INTO olapdata.dim_asset (Asset_ID,
  Property_Type, Assessed_Value)
 SELECT DISTINCT ROUND(RAND() * 1000000), Property_Type,
  Assessed_Value
 FROM stagingarea.complaints ;

 INSERT IGNORE INTO olapdata.dim_area (Area_ID, Town,
 Address)
 SELECT DISTINCT ROUND(RAND() * 1000000), Town, Address
 FROM stagingarea.complaints ;

  INSERT IGNORE INTO olapdata.dim_choice
  (Installment_period, choice_ID, Customer_ID,
  Rate,Property_Type)
 SELECT DISTINCT Installment_period, ROUND(RAND() *
  1000000),ROUND(RAND() * 1000000), Rate,Property_Type
 FROM stagingarea.complaints ;

 INSERT IGNORE INTO olapdata.dim_time (Date_ID,
  List_years,
Date_Recorded)
 SELECT DISTINCT ROUND(RAND() * 1000000), List_years,
  Date_Recorded
 FROM stagingarea.complaints ;

 INSERT IGNORE INTO olapdata.dim_customer (Customer_ID,
  Income_Value,
Residential_Type)
 SELECT DISTINCT ROUND(RAND() * 1000000), Income_Value,
```

```
  Residential_Type
 FROM stagingarea.complaints ;

 INSERT IGNORE INTO olapdata.Fact_transaction_records (
    Serial_Number,
    Asset_ID,
    Area_ID,
    Date_ID,
    Customer_ID,
    Property_Type
)
SELECT
    s.Serial_Number,
    a.Asset_ID,
    ar.Area_ID,
    t.Date_ID,
    c.Customer_ID,
    s.Property_Type
FROM
    StagingArea.Real_Estate s
JOIN
    olapdata.Dim_Asset a ON s.Asset_ID = a.Asset_ID
JOIN
    olapdata.Dim_Area ar ON s.Area_ID = ar.Area_ID
JOIN
    olapdata.Dim_Time t ON s.List_years = t.List_years
JOIN
    olapdata.Dim_Customer c ON s.Income_Value =
   c.Income_Value;

INSERT IGNORE INTO olapdata.Fact_Asset_Classification (
    Category_ID,
    Area_ID,
```

```
    Property_Type,
    Asset_classification,
    Sale_Amount,
    Sales_Ratio
)
SELECT
    ROUND(RAND() * 1000000),
    ar.Area_ID,
    s.Property_Type,
    s.Asset_classification,
    s.Sale_Amount,
    s.Sales_Ratio
FROM
    StagingArea.Real_Estate s
JOIN
    olapdata.Dim_Area ar ON s.Area_ID = ar.Area_ID;

INSERT IGNORE INTO olapdata.Fact_customer_classification (
    CusCategory_ID,
    Customer_ID,
    Choice_ID,
    Income,
    Residential_Type
)
SELECT
    ROUND(RAND() * 1000000),
    c.Customer_ID,
    ch.Choice_ID,
    s.Income,
    s.Residential_Type
FROM
    StagingArea.Real_Estate s
JOIN
```

```
    olapdata.Dim_Customer c ON s.Income_Value =
  c.Income_Value
JOIN
    olapdata.Dim_Choice ch ON s.Installment_period =
  ch.Installment_period
        AND s.Property_Type = ch.Property_Type;
END //
DELIMITER ;
```

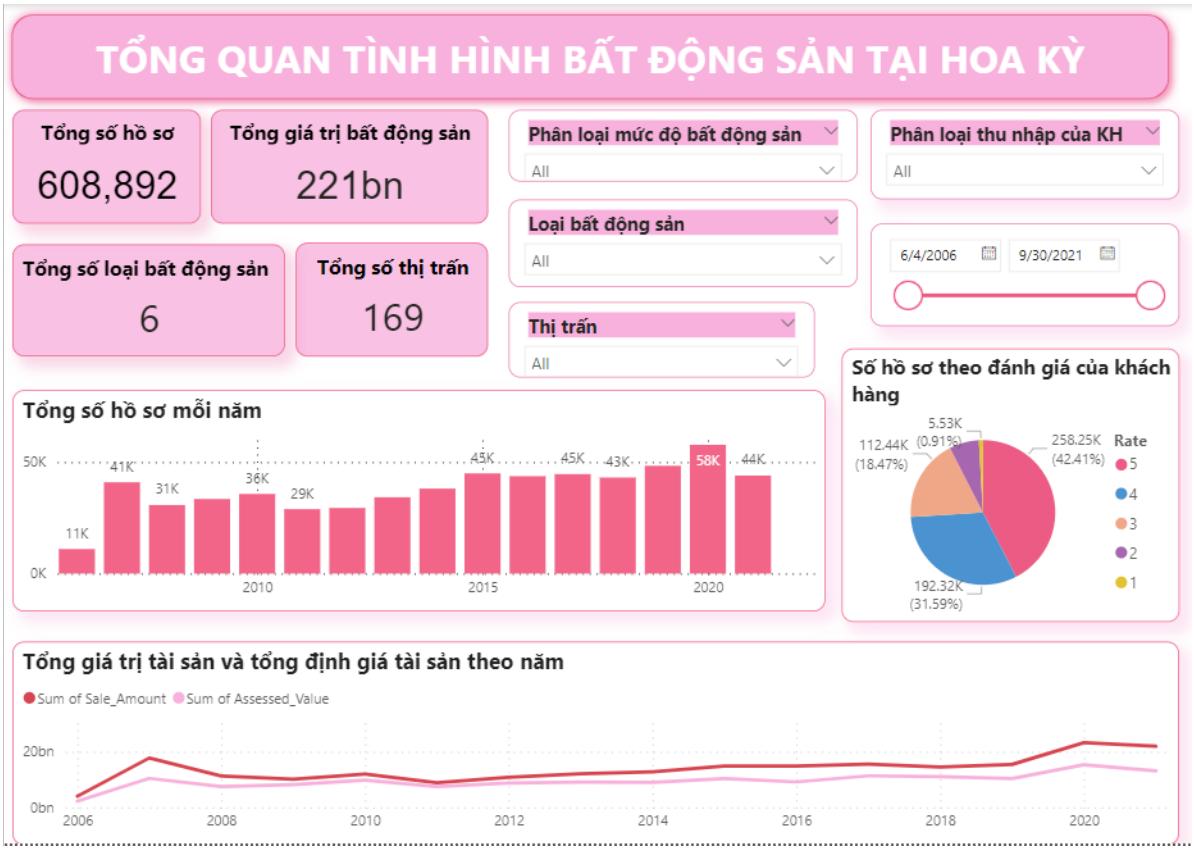
3.3 Xây dựng Dashboard

Phần này xây dựng Dashboard sử dụng Power BI giúp tối ưu cho việc trực quan hóa dữ liệu, từ nguồn dữ liệu được ETL bằng Python và xuất ra file excel Real_Estate.csv. Power BI là một công cụ tạo báo cáo và trực quan hóa dữ liệu của Microsoft, được phát triển để giúp người dùng tạo các báo cáo, biểu đồ, và bảng điều khiển trực quan từ các nguồn dữ liệu khác nhau. Power BI cho phép:

- Kết nối, lọc và chuyển đổi dữ liệu từ nhiều nguồn khác nhau để tạo ra báo cáo ý nghĩa
- Cung cấp nhiều công cụ và tính năng trực quan hóa dữ liệu thông qua các biểu đồ, đồ thị, ... giúp ta hiểu rõ hơn về các thông tin quan trọng của doanh nghiệp và từ đó ra quyết định dựa trên chúng.

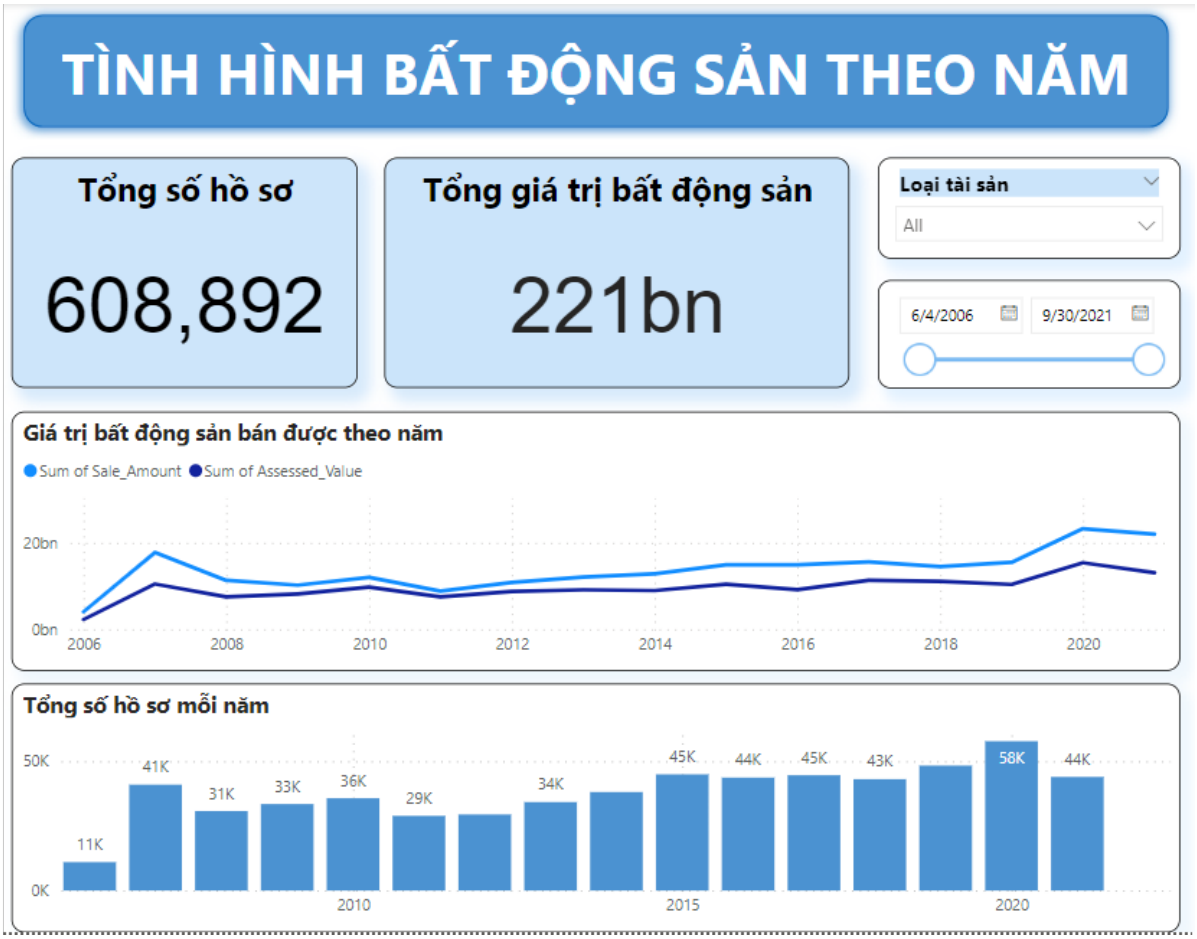
Các Dashboard mà xây dựng tập trung chính vào phân tích các yêu cầu đặt ra ở chương I.

3.3.1 Tổng quan tình hình bất động sản



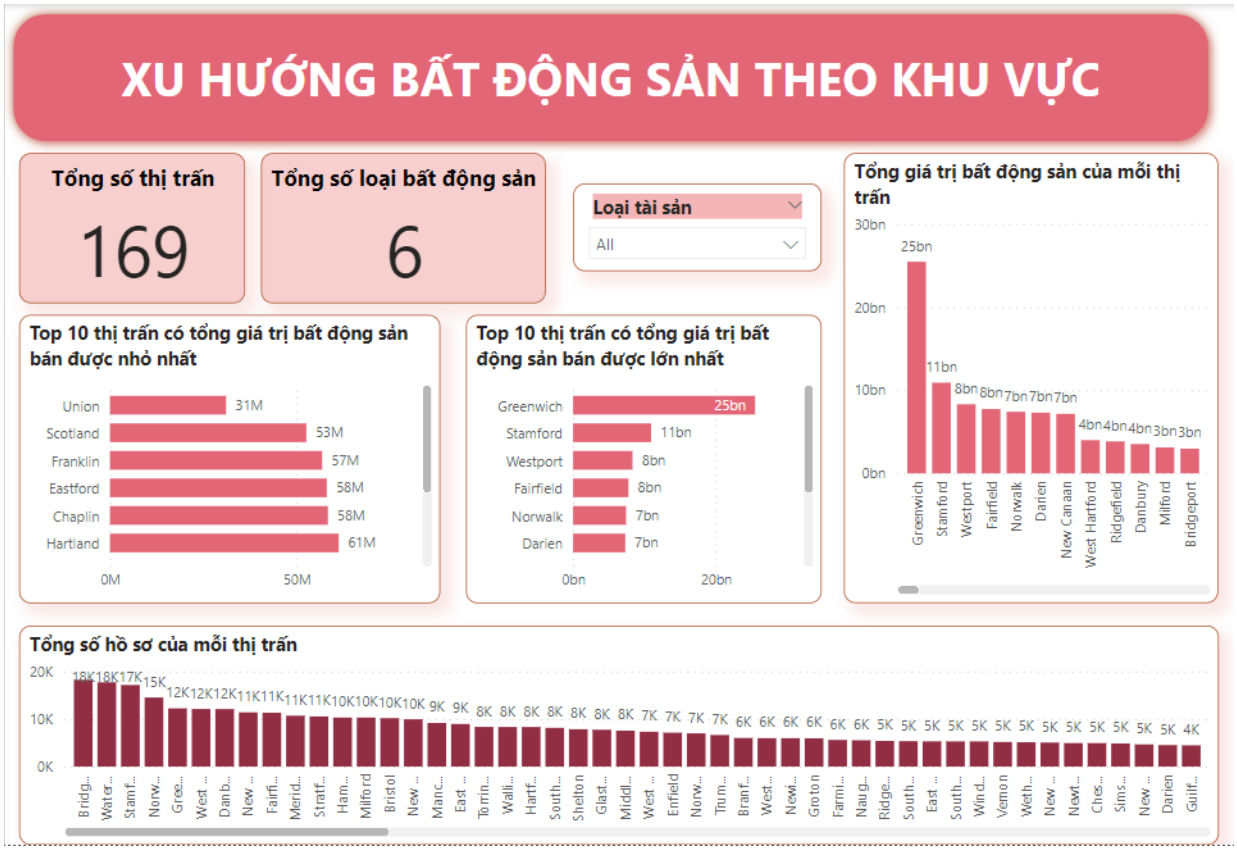
Hình 3.2: Dashboard_ Tổng quan tình hình bất động sản tại Hoa Kỳ

3.3.2 Tình hình bất động sản theo năm



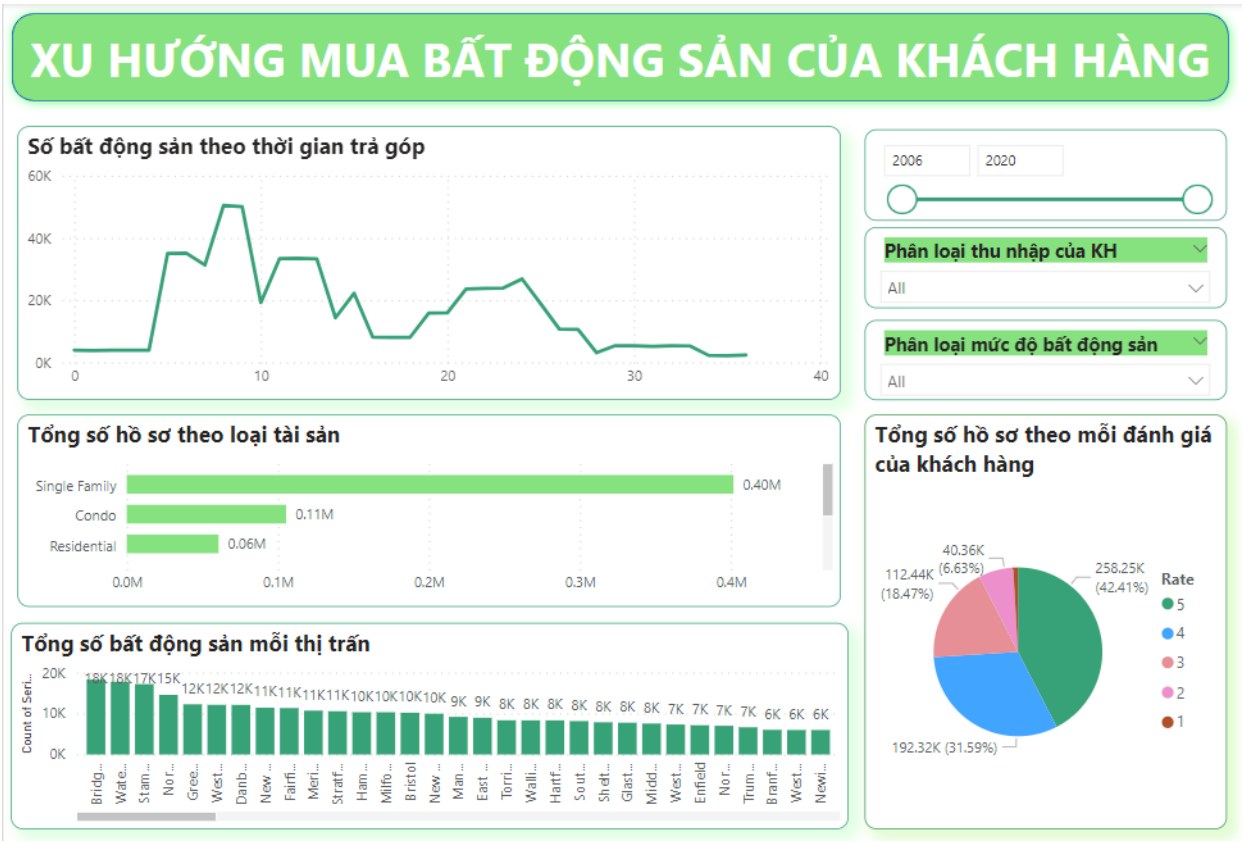
Hình 3.3: Dashboard_ Tình hình bất động sản tại Hoa Kỳ theo năm

3.3.3 Xu hướng bất động sản theo khu vực



Hình 3.4: Dashboard_ Xu hướng bất động sản tại Hoa Kỳ theo khu vực

3.3.4 Xu hướng mua bất động sản của khách hàng



Hình 3.5: Dashboard_ Xu hướng mua bất động sản của khách hàng tại Hoa Kỳ

3.4 Nhân định tình hình

Từ quá trình phân tích các Dashboard, chúng ta có thể đưa ra một số nhận định về thị trường bất động sản tại Hoa Kỳ như sau:

- Dựa vào tình hình thị trường như số lượng bất động sản được bán, thời gian bán trung bình và tỷ lệ giữa giá niêm yết và giá bán có thể cung cấp cái nhìn toàn diện về thị trường. Có thể mở rộng nhiều sự lựa chọn bất động sản có giá dao động quanh kinh tế mà không phải chỉ chọn những bất động sản có giá bằng khả năng kinh tế của mình.

- Tìm hiểu giá bất động sản theo khu vực: Khách hàng hiểu rõ mức giá trung bình của bất động sản ở từng khu vực cụ thể trong Hoa Kỳ. Từ đây, khách hàng quyết định về việc mua nhà hoặc đầu tư vào khu vực nào phù hợp với ngân sách của họ mà không sợ bị mua quá đắt với giá thực tế.

- Theo dõi, tìm hiểu giá của bất động sản trong một thời gian dài. Thị trường bất động sản là thị trường lâu dài, việc chỉ xem xét trong 1-2 năm là khó để nắm bắt được xu hướng, vì vậy việc tìm hiểu sự biến động của bất động sản theo thời gian có thể dự đoán được xu hướng của loại bất động sản đang quan tâm, quyết định xem thời điểm nào thích hợp để mua nó.

- Phân tích sự ưa chuộng các loại bất động sản cụ thể là các loại như nhà ở, căn hộ, đất đai hay bất động sản thương mại để khách hàng hiểu rõ hơn về sự đa dạng của thị trường và điều chỉnh lựa chọn theo nhu cầu cụ thể.

- Xác định những vùng đất tiềm năng: kết hợp nhiều điều kiện như giá cả, vị trí địa lý, sự ưa chuộng và khả năng phát triển trong tương lai để đưa ra những bất động sản được cho là tiềm năng và định hình lại chiến lược đầu tư sao cho chắc chắn nhất. Ví dụ như với bộ dữ liệu trên, thông qua Dashboard phân tích tình hình bất động sản theo khu vực, thị trấn Bridgeport là thị trấn có số lượng hồ sơ bất động sản bán được nhiều nhất nhưng về tổng giá trị bất động sản bán được thì chỉ đứng thứ 10. Vì vậy chúng ta cần phân tích dựa trên nhiều yếu tố và dựa trên mục đích mua bất động sản để lựa chọn bất động sản hợp lý với mục tiêu của mình nhất.

- Và điều cuối cùng là quan trọng nhất nhưng không thể hiện trong bộ dữ liệu, đó là quy định về Quy tắc và Luật pháp, khách hàng phải tìm hiểu kĩ về thông tin

về quy tắc và luật pháp liên quan đến mua và sở hữu bất động sản tại Hoa Kỳ.

- Nói chung, để hiểu rõ và phân tích chính xác thị trường bất động sản không phải là điều dễ và không phải ai cũng có cái nhìn chiến lược như những người có chuyên môn. Nhưng việc tìm hiểu và phân tích dựa trên tầm hiểu biết của bản thân và tham khảo ý kiến của người có chuyên môn để tránh những trường hợp rủi ro từ nhỏ nhất khi bắt đầu tham gia vào thị trường bất động sản.

Kết luận

Trong quá trình thực hiện đề tài này, em đã tìm hiểu và vận dụng các kiến thức được cô trang bị và có những cố gắng nhất định để hoàn thành báo cáo. Đề tài " Xây dựng hệ thống phân tích dữ liệu về bất động sản tại Hoa Kỳ" đã thực hiện được những công việc sau:

- Hiểu được về những kiến thức nền tảng cơ bản của kinh doanh thông minh, phân tích dữ liệu và phân tích kinh doanh
- Biết cách xử lý dữ liệu thô, ETL dữ liệu cơ bản bằng công cụ lập trình Python
- Phân tích và thiết kế mô hình hệ thống mới
- Xây dựng được hệ quản trị cơ sở dữ liệu trên hệ quản trị hệ thống MySQL
- Xây dựng báo cáo theo những chiều khác nhau bằng Microsoft Power BI đáp ứng yêu cầu của đề bài.

Trong tương lai, đề tài tiếp tục phát triển theo các hướng sau:

- Cần xử lý thêm cả dữ liệu chữ
- Tăng hiệu năng cho hệ thống phân tích
- Mở rộng thêm dữ liệu từ nhiều hệ thống OLTP và xử lý vấn đề cập nhật dữ liệu khi hệ thống OLTP cập nhật liên tục.

Do điều kiện thời gian còn có hạn chế nên báo cáo không tránh khỏi những thiếu sót. Rất mong nhận được những ý kiến đóng góp quý báu của thầy cô và các bạn. Em xin chân thành cảm ơn!

Tài liệu tham khảo

Tài liệu tiếng Việt

[1] Nguyễn Danh Tú, *Slide Bài giảng Kho dữ liệu và Kinh doanh thông minh*, 2023.

Tài liệu tiếng Anh

[2] Alaa Khalaf Hamoud, Hisham Noori Hussien, Arwa Akram Fadhil, Zahraa Raad Ekal, *"Improving Service Quality Using Consumers' Complaints Data Mart which Effect on Financial Customer Satisfaction"*, 2020.

[3] Vageesh M Shivaprasad, *"Analysis of Customer Complaint Data of Consumer Financial Protection Bureau Using Different Text Classification Methods"*, 2019.