TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC
School of Applied Mathematics and Informatics

Báo cáo cuối kỳ

KHO DỮ LIỆU VÀ KINH DOANH THÔNG MINH

Chủ đề:
Healthycare - Chăm sóc sức khỏe

Giảng viên hướng dẫn: ThS. Nguyễn Danh Tú

Nhóm thực hiện:	Nhóm 9
Nguyễn Thị Duyên	20195866
Phạm Thị Hoa	20195874
Trần Thị Hồng	20195880
Phạm Thu Trang	20195931
Trần Thị Hồng Vân	20195941
Lớp:	133598 - MI4214
Học kỳ:	20212

Ngày 6 tháng 5 năm 2023

Mục lục

Lời mở đầu

Với sự phát triển khoa học kỹ thuật trên toàn cầu, nguồn dữ liệu thông tin khảo sát của mỗi một doanh nghiệp, tập đoàn ngày càng nhiều và tăng trưởng theo cấp số nhân. Với nhu cầu tiếp nhận, phân tích và xử lý dữ liệu dưới góc nhìn đa chiều và tổng hợp hiện nay, việc thống kê dòng dữ liệu là vô cùng cần thiết, từ đó khái niệm kho dữ liệu ra đời nhằm đảm lưu trữ đầy đủ dữ liệu cho bước phân tích tiếp theo và nâng cao tốc độ của các kết quả trả về của hệ thống. Cùng với Data Warehouse thì Dashboard cũng là một công cụ không thể thiếu trong các hoạt động kinh doanh, quản lý của tổ chức. Nhờ có Dashboard mà nhà quản trị có cái nhìn tổng quan, chi tiết và cụ thể cho hướng đi của doanh nghiệp.

Thế giới có nhịp độ nhanh ngày nay đã ảnh hưởng không nhỏ đến sức khỏe của chúng ta. Điều rất quan trọng là phải tiêu thụ thực phẩm lành mạnh và có một chế độ ăn uống cân bằng, tập luyện để giữ cho chúng ta khỏe mạnh về thể chất và tinh thần.

Dựa vào nhu cầu đó nhóm 9 quyết định chọn chủ đề HEALTH CARE để tìm hiểu và báo cáo về các khảo sát sức khỏe của mọi người dựa vào các nhân tố ảnh hưởng đến sức khỏe như tuổi tác, các bệnh nền từ bộ dữ liệu của hệ thống giám sát yếu tố rủi ro hành vi BRFSS.

Chúng em xin gửi lời cảm ơn đến thầy ThS. Nguyễn Danh Tú đã hướng dẫn và đưa ra những góp ý cho nhóm hoàn thành đề tài.

Hà Nội,ngày 26 tháng 7 năm 2022
Thay mặt nhóm báo cáo

Nhóm 9

Đánh giá các thành viên

Các thành viên trong nhóm

1. Nguyễn Thị Duyên - 20195866
2. Phạm Thị Hoa - 20195874
3. Trần Thị Hồng -20195880
4. Phạm Thu Trang -20195931
5. Trần Thị Hồng Vân - 20195941

Đánh giá:

Bảng đánh giá các thành viên trong nhóm

Họ và tên	Nguyễn Thị Duyên
Lớp	Toán tin 01 - K64
Nhóm	9

STT	Tên thành viên	Làm tốt nhiệm vụ được giao	Liên hệ được khi cần	Khả năng đóng góp sáng kiến, ý kiến cho hoạt động nhóm	Sẵn sàng giúp đỡ	Đóng góp chung vào kết quả của nhóm
1	Nguyễn Thị Duyên	4	3	4	4	5
2	Phạm Thị Hoa	5	4	3	5	3
3	Trần Thị Hồng	5	3	3	5	4
4	Phạm Thu Trang	4	4	5	3	4
5	Trần Thị Hồng Vân	3	5	4	4	3

Hình 1: Bảng đánh giá thành viên

Trong quá trình hoạt động, các thành viên trong nhóm đều có tinh thần làm việc tốt, hoàn thành tốt nhiệm vụ được giao và đều đóng góp vào trong kết quả của nhóm, từ khâu chọn dữ liệu đến việc phân tích, xử lý dữ liệu và vẽ Dashboard. Với tinh thần làm việc nhóm tốt, mọi người đã hoàn thành công việc đúng hạn và báo cáo Dưới đây là công sức của cả 5 thành viên đã cố gắng trong 16 tuần vừa qua.

Tự đánh giá báo cáo nhóm

1. Mục tiêu và nội dung của báo cáo

(a) Mục tiêu:

Sau khi hoàn thành báo cáo, chúng em sẽ nắm rõ được những khái niệm cơ bản về phân tích dữ liệu; hiểu được quy trình phân tích dữ liệu và triển khai vào công việc thực tiễn; Làm chủ được các công cụ trực quan hóa dữ liệu như Excel, Power BI và Hệ quản trị cơ sở dữ liệu để lựa chọn giải pháp tốt nhất trong thực tế.

(b) Nội dung của bài Báo cáo:

- Tổng quan về Data Warehouse.
- Tổng quan về Heathycare
- Phân tích nghiệp vụ.
- Đưa ra requirement khi xây dựng kho dữ liệu.
- Kiến trúc Data Warehouse
- Sơ đồ quá trình ETL và các nội dung ETL bằng Power Query và SQL
- Xử lý dữ liệu và vẽ sơ đồ dữ liệu OLTP
- Sơ đồ các chiều dữ liệu
- Đưa dữ liệu từ cơ sở dữ liệu vào công cụ phân tích Power BI và OLAP hoá.
- Vẽ sơ đồ dữ liệu OLAP
- Vẽ các dashboard theo chủ đề
- Phân tích các dashboard.

2. Kết quả đạt được:

- Hiểu được những kiến thức nền tảng cơ bản về Kho dữ liệu, Kinh doanh thông minh, phân tích dữ liệu và phân tích kinh doanh. Bên cạnh đó xây dựng được kiến trúc của kho dữ liệu.
- Nắm rõ được quy trình phân tích dữ liệu.
- Phân tích và thiết kế được các mô hình của hệ thống mới.
- Làm chủ được một số công cụ trực quan hóa dữ liệu như Excel, Power BI.
- Khảo sát và làm việc trên bộ dữ liệu thực tế
- Xây dựng được kiến trúc kho dữ liệu ứng với bài toán thực tế của nhóm.
- ETL dữ liệu trên các công cụ Excel, Power Query
- Vẽ được sơ đồ OLTP
- Xác định được các chiều phân tích của bài toán

- Xây dựng được mô hình dữ liệu OLAP và đưa vào công cụ trực quan hóa Power BI.
- Xây dựng được các dashboard phân tích và tổng quan.
- Phân tích được các dashboard theo các chiều dữ liệu đã xác định.
- Tiếp cận đến các công cụ mới để phân tích dữ liệu và xây dựng DW như SQL Server Analysis Service...

3. Nội dung chưa làm được:

- Chưa xử lý thành thạo bộ dữ liệu kích thước lớn, và còn thiếu dữ liệu ban đầu.
- Sử dụng công cụ trực quan hóa Power BI chưa được chuyên nghiệp, các Dashboard chưa phân tích sâu các chiều dữ liệu.

4. Bài học thu được:

- Việc xây dựng kho dữ liệu là vô cùng cần thiết cho mục đích phân tích dữ liệu.
- Dữ liệu trong thực tế không phải lúc nào cũng được chuẩn hóa theo một quy tắc và không dữ liệu nào là hoàn hảo. Vì vậy luôn cần phải xử lý trước khi đưa vào hệ thống.
- Quá trình ETL dữ liệu rất quan trọng và tốn nhiều thời gian nên đòi hỏi cần phải tập trung quan tâm thực hiện. Cần nghiên cứu bộ dữ liệu một cách tỉ mỉ để xử lý dữ liệu một cách đúng đắn để tránh xảy ra sai sót và mất dữ liệu
- Dữ liệu có kích thước lớn gây ảnh hưởng đến thời gian chạy của máy tính, cần chọn máy có cấu hình phù hợp cho việc xử lý và phân tích. Và có thể chia nhỏ các file để dễ dàng thực hiện.
- Kiến thức nghiệp vụ của mỗi quy trình là vô cùng quan trọng, vì khi hiểu được nghiệp vụ chúng ta mới có thể xây dựng được một hệ thống BI đúng và hiệu quả. Vì vậy, đầu tiên cần phải khảo sát thật kỹ nghiệp vụ của từng quy trình hoạt động.
- Mô hình dữ liệu đa chiều giúp phân tích dữ liệu trên nhiều góc nhìn khác nhau, và có thể phân cấp.
- Các dashboard nên được xây dựng theo hướng chủ đề, phải đảm bảo tính logic và thẩm mỹ.

Danh sách hình vẽ

1 Tổng quan về Data Warehouse

1.1 Khái niệm:

Kho dữ liệu (Data Warehouse) được hiểu là một tập hợp các dữ liệu tương đối ổn định (không hay thay đổi), cập nhật theo thời gian, được tích hợp theo hướng chủ đề nhằm hỗ trợ quá trình tạo quyết định về mặt quản lý (W.H. Inmon).

Kho dữ liệu về bản chất là một cơ sở dữ liệu bình thường, các hệ quản trị cơ sở dữ liệu quản lý và lưu trữ nó như các cơ sở dữ liệu thông thường. Tuy nhiên, nó có thể quản lý dữ liệu lớn và hỗ trợ truy vấn. Nên điểm khác biệt giữa kho dữ liệu và cơ sở dữ liệu là ở quan niệm, cách nhìn nhận vấn đề.

1.2 Đặc tính:

1. Hướng chủ đề: Cung cấp một khung nhìn đơn giản và súc tích xung quanh các sự kiện của các chủ đề ứng với mỗi loại tổ chức. Kho dữ liệu được thiết kế để hỗ trợ việc phân tích dữ liệu và hỗ trợ ra quyết định sau khi loại bỏ những dữ liệu không hữu ích.

2. Tính tích hợp: Là đặc tính quan trọng nhất. Khi dữ liệu từ nhiều nguồn khác nhau đưa vào kho dữ liệu, chúng sẽ được chuyển đổi, định dạng lạ,.. để đảm bảo sự đồng nhất trong các quy ước tên, cấu trúc mã hóa, các đơn vị đo, thuộc tính,... giữa các nguồn khác nhau.

3. Gắn thời gian và có tính lịch sử: Dữ liệu trong kho dữ liệu bao gồm cả quá khứ và hiện tại. Mỗi dữ liệu trong kho dữ liệu đều được gắn với thời gian và có tính lịch sử.

4. Chỉ đọc, không biến động: Là một lưu trữ vật lý của dữ liệu được chuyển đổi từ môi trường tác nghiệp. Kho dữ liệu tách rời với môi trường tác nghiệp, nên dữ liệu trong đó là dữ liệu chỉ đọc, không chỉnh sửa hoặc thêm mới.

1.3 Lợi ích

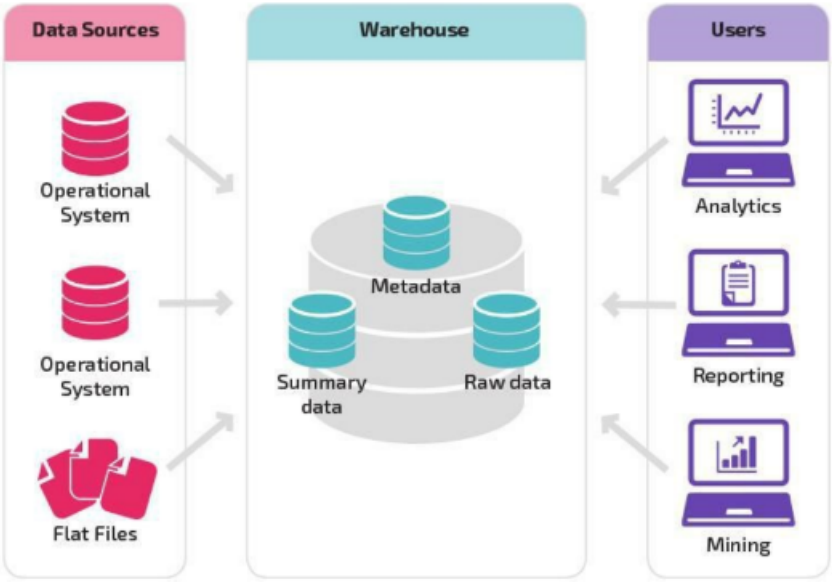
- Dữ liệu sau khi đưa vào kho dữ liệu đều tuân theo những quy tắc thống nhất.
- Dữ liệu được tổ chức tạo thuận lợi cho việc truy vấn phân tích và tạo tiền đề để đưa ra những quyết định có ảnh hưởng lớn.
- Cải thiện tính bảo mật và hiệu suất mà không cần tách động tới hệ thống dữ liệu gốc.
- Công việc kinh doanh trở nên thông minh hơn, nâng cao dịch vụ khách hàng.

1.4 Kiến trúc:

1.4.1 Phân loại kiến trúc:

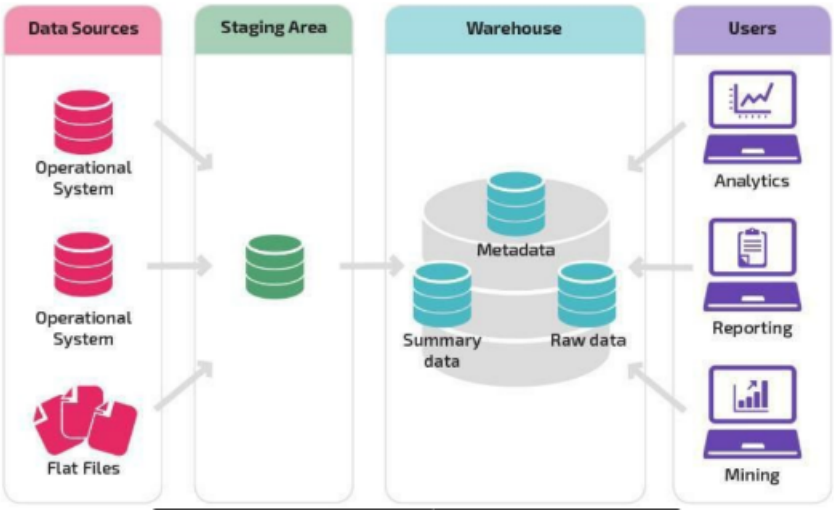
Phụ thuộc rất nhiều vào vị trí của từng bộ phận trong tổ chức. Các kiến trúc phổ biến của kho dữ liệu bao gồm:

1. **Kiến trúc cơ bản:** Kiến trúc cơ bản rất ít được sử dụng trong thực tế. Mặc dù, kiến trúc này loại bỏ dư thừa dữ liệu giúp giảm thiểu lượng dữ liệu được lưu trữ nhưng không phù hợp với các doanh nghiệp có yêu cầu dữ liệu phức tạp và nhiều nguồn dữ liệu. Trong kiến trúc cơ bản chỉ có tầng nguồn là tầng có sẵn về mặt vật lý. Kho dữ liệu của kiến trúc này là ảo tức nó được xây dựng dưới dạng một cái nhìn đa chiều về dữ liệu hoạt động và được tạo bởi phần mềm trung gian cụ thể hoặc một lớp xử lý trung gian.



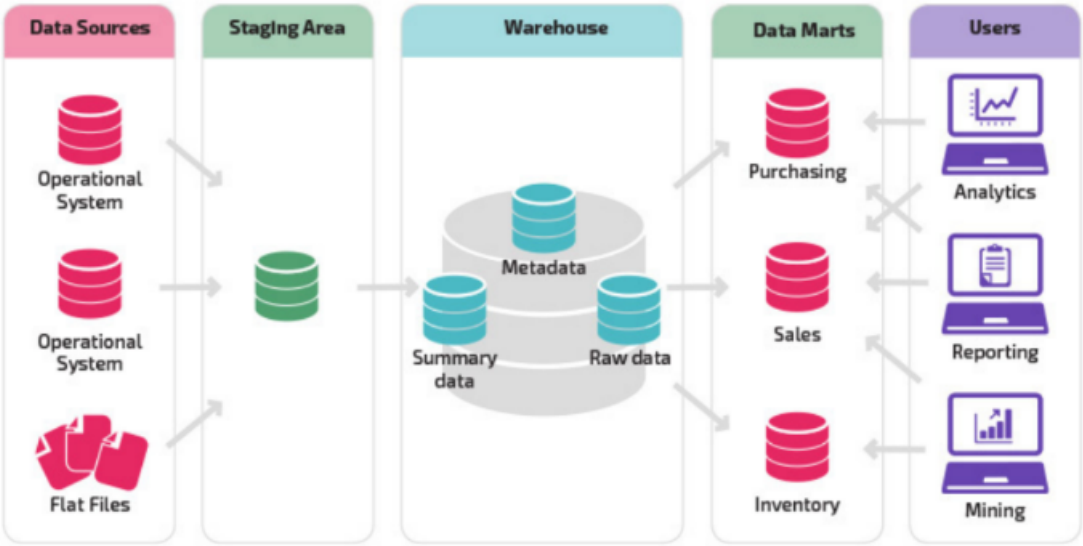
Hình 2: Kiến trúc kho dl cơ bản

2. **Kiến trúc với vùng tập kết dữ liệu (Staging Area):** Kiến trúc này cung cấp thông tin luôn ở chất lượng tốt, ngay cả khi quyền truy cập vào các nguồn bị từ chối tạm thời vì lý do kỹ thuật hoặc tổ chức; truy vấn phân tích kho dữ liệu không ảnh hưởng đến việc quản lý các giao dịch, độ tin cậy cao, kho dữ liệu được cấu trúc logic theo mô hình đa chiều, kho dữ liệu có thể sử dụng các giải pháp thiết kế cụ thể để tối ưu hóa hiệu suất của các ứng dụng phân tích và báo cáo.



Hình 3: Kiến trúc kho dữ liệu vs Staging Area

3. **Kiến trúc với vùng tập kết dữ liệu (Staging Area) và kho dữ liệu chủ đề (Data Marts):** Đây là kiến trúc phổ biến nhất trong ba loại. Ưu điểm chính của tầng tập kết dữ liệu (Staging Area) là tạo ra một mô hình dữ liệu tham chiếu chung cho cả doanh nghiệp. Đồng thời, tách biệt rõ ràng các vấn đề khai thác và tích hợp dữ liệu nguồn với các vấn đề của tổng thể kho dữ liệu. Đáng chú ý, tầng tập kết dữ liệu cũng được sử dụng trực tiếp để hoàn thành tốt một số nhiệm vụ hoạt động. Bên cạnh đó, kho dữ liệu chủ đề (Data Mart) giúp dữ liệu được truy cập dễ dàng hơn và cải thiện hiệu suất hệ thống.



Hình 4: Kiến trúc kho dl với Staging và Data Marts

1.4.2 Nguồn dữ liệu

Nguồn dữ liệu của kho dữ liệu rất đa dạng:

- Dữ liệu từ các hệ thống tác nghiệp.
- Hệ thống kế thừa.
- Các nguồn dữ liệu bên ngoài.

1.4.3 Tập kết dữ liệu

Tập kết dữ liệu (Data Staging) là nơi lưu trữ trung gian được sử dụng để xử lý dữ liệu trong thời gian chiết xuất, chuyển đổi và tải (ETL) quá trình. Dữ liệu tại đây có thể bị xóa trước khi chạy quy trình ETL hoặc ngay sau khi hoàn thành thành công quy trình ETL. Tuy nhiên, trong một số trường hợp Data Staging được thiết kế để lưu giữ dữ liệu trong thời gian dài cho các mục đích lưu trữ hoặc xử lý sự cố.

1.4.4 Công cụ trích xuất, chuyển đổi và tải dữ liệu

Quy trình ETL được thiết kế phù hợp trích xuất dữ liệu từ hệ thống nguồn, thực thi các tiêu chuẩn về chất lượng và tính nhất quán của dữ liệu, tuân thủ dữ liệu để các nguồn riêng biệt có thể được sử dụng cùng nhau và cuối cùng cung cấp dữ liệu ở định dạng sẵn sàng trình bày để các nhà phát triển ứng dụng có thể xây dựng ứng dụng và người dùng cuối có thể đưa ra quyết định.

1. **Trích xuất (Extract):** Trích xuất dữ liệu một cách tạo tiền đề cho sự thành công của các quy trình tiếp theo. Hầu hết các dự án kho dữ liệu kết hợp dữ liệu từ các hệ thống nguồn khác nhau. Mỗi hệ thống riêng biệt cũng có thể sử dụng một tổ chức và/hoặc định dạng dữ liệu khác nhau. Việc truyền trực tuyến nguồn dữ liệu được trích xuất và tải nhanh chóng đến cơ sở dữ liệu đích là một cách khác để thực hiện ETL khi không cần lưu trữ dữ liệu trung gian. Quá trình trích xuất nhằm xác thực dữ liệu xem được lấy từ các nguồn có các giá trị chính xác hay không.

2. **Biến đổi (Transform):** Trong giai đoạn chuyển đổi dữ liệu , một loạt các quy tắc hoặc chức năng được áp dụng cho dữ liệu được trích xuất để chuẩn bị cho việc tải vào mục tiêu cuối cùng. Một chức năng quan trọng của chuyển đổi là làm sạch dữ liệu, nhằm chuyển dữ liệu "thích hợp"đến mục tiêu. Việc tương tác giữa các hệ thống khác nhau gặp cản trở do các bộ ký tự có thể có sẵn trong hệ thống này nhưng có thể không có ở các hệ thống khác.

3. **Load:** Tùy thuộc vào yêu cầu của tổ chức, quá trình tải dữ liệu rất khác nhau. Một số kho dữ liệu có thể ghi đè thông tin hiện có bằng thông tin tích lũy; cập nhật dữ liệu trích xuất thường xuyên được thực hiện theo chu kỳ. Các kho dữ liệu khác (hoặc thậm chí các phần khác của cùng một kho dữ liệu) có thể thêm dữ liệu mới ở dạng lịch sử theo các khoảng thời gian đều đặn. Thời gian và phạm vi thay thế hoặc kết nối thêm là các lựa chọn thiết kế chiến lược phụ thuộc vào thời gian có sẵn và nhu cầu của doanh nghiệp. Các hệ thống phức tạp hơn có thể duy trì lịch sử và dấu vết

kiểm tra tất cả các thay đổi đối với dữ liệu được tải vào kho dữ liệu.
Khi giai đoạn tải tương tác với cơ sở dữ liệu, các ràng buộc được xác định trong lược đồ cơ sở dữ liệu - cũng như trong các trình kích hoạt được kích hoạt khi tải dữ liệu - áp dụng. Ví dụ: Tính duy nhất, tính toàn vẹn tham chiếu, các trường bắt buộc), cũng góp phần vào hiệu suất chất lượng dữ liệu tổng thể của quy trình ETL.

1.4.5 Siêu dữ liệu

Siêu dữ liệu (Metadata) lưu các định nghĩa logic các bảng, thuộc tính của kho dữ liệu, tên các nguồn dữ liệu tác nghiệp, định nghĩa vật lý các bảng và các cột.

1.4.6 Kho dữ liệu chủ đề

Kho dữ liệu chủ đề (Data Mart) có những đặc điểm giống với kho dữ liệu (Data Warehouse) nhưng với quy mô nhỏ hơn và lưu trữ dữ liệu về một lĩnh vực, một chuyên ngành. Các kho dữ liệu chủ đề có thể được hình thành từ một tập con dữ liệu của kho dữ liệu hoặc cũng có thể được xây dựng độc lập và sau khi xây dựng xong, các kho dữ liệu chủ đề có thể được kết nối tích hợp lại với nhau tạo thành kho dữ liệu. Vì vậy có thể xây dựng kho dữ liệu bắt đầu bằng việc xây dựng các kho dữ liệu chủ đề hay ngược lại xây dựng kho dữ liệu trước sau đó tạo ra các kho dữ liệu chủ đề.

1.4.7 Các công cụ và nền tảng hỗ trợ

- Hệ quản trị CSDL: MySQL, Oracle, Microsoft SQL Server, MariaDB,..
- Trực quan hoá dữ liệu, tạo báo cáo (Report, Dashboard): Power BI, Tableau,...
- Phân tích dữ liệu: Python, R, SAS, Excel. Orange...

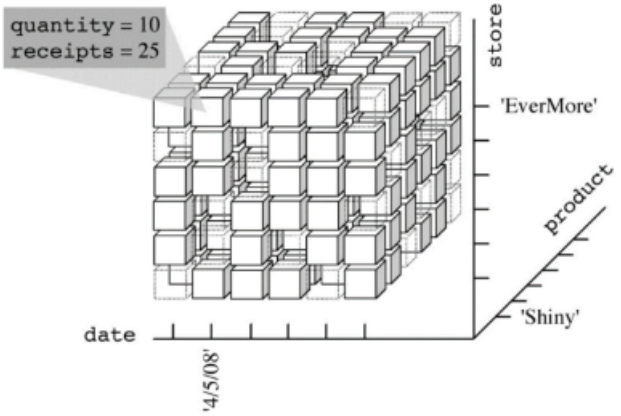
1.5 Kiến trúc khối và Các dạng lược đồ dữ liệu đa chiều

1.5.1 Mô hình dữ liệu đa chiều

Data Warehouse và các hệ thông OLAP được xây dựng theo mô hình dữ liệu đa chiều. Dữ liệu trong kho dữ liệu được thể hiện dưới dạng đa chiều gọi là khối (Cube). Mỗi chiều mô tả một đặc trưng nào đó của dữ liệu. (Nếu số chiều dữ liệu lớn hơn 3, gọi là Hyper Cube).

1. **Chiều (Dimension) & Độ đo (Measure)**
   - Chiều cung cấp các thông tin, ngữ cảnh của bảng Fact.
   - Độ đo là đại lượng có thể tính toán được trên các thuộc tính của bảng Fact.
2. **Cây phân cấp & Số liệu tổng hợp**
   - Mức độ chi tiết của các tiêu chí thể hiện cho người dùng được gọi là mức dữ liệu, được quyết định bằng việc kết hợp các mức dữ liệu của từng phân lớp.
   - Số liệu tổng hợp: Việc tổng hợp số liệu xảy ra khi người dùng thay đổi mức chi tiết của dữ liệu lấy ra từ Cube, bằng cách duyệt qua cây phân cấp.

3. **Các mô hình thiết kế Date Warehouse**
   - OLAP kiểu quan hệ (Relational OLAP ∼ ROLAP)
   - OLAP kiểu đa chiều (Multi-dimensional OLAP ∼ MOLAP)
   - OLAP kết hợp (Hybird OLAP HOLAP = ROLAP + MOLAP)



Hình 5: Minh họa khối dữ liệu

1.5.2 Lược đồ dữ liệu đa chiều

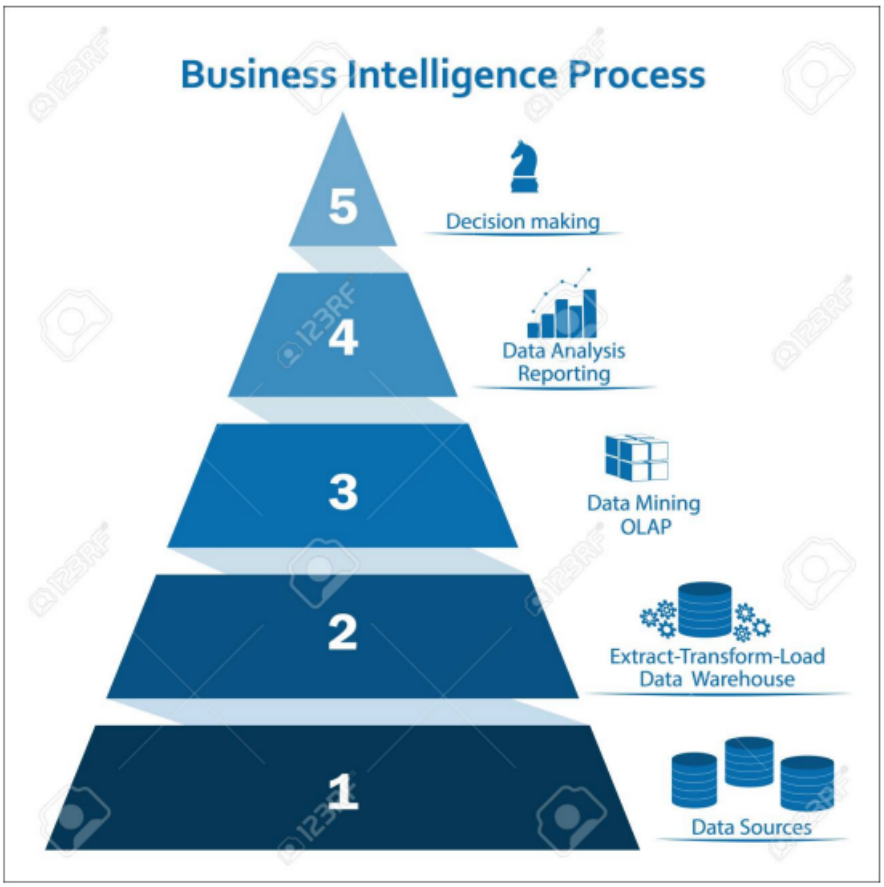
1. **Lược đồ hình sao**
   Gồm 1 bảng Fact nằm ở trung tâm và những bảng Dimension bao quanh. Các câu hỏi nhằm vào bảng Fact và được cấu trúc bởi các bảng Dimension.
   - Ưu điểm: Bảng Fact, Dimension được mô tả rõ ràng, dễ hiểu. Bảng Dim chứa dữ liệu tĩnh, và bảng Fact chứa dữ liệu động được nạp bằng các thao tác. Khoá của Fact được tạo bởi khoá của các bảng Dim. Nghĩa là khoá chính của các bảng Dim chính là khoá ngoại của bảng Fact.
   - Nhược điểm: Dữ liệu không được chuẩn hoá
2. **Lược đồ hình bông tuyết**
   Lược đồ hình bông tuyết là một sự mở rộng của lược đồ hình sao tại đó mỗi "cánh ngôi sao". Các chiều được cấu trúc rõ ràng. Bảng Dimension được chia thành hai chiều: chính và phụ.
   - Ưu điểm: Số chiều được phân cấp thể hiện dạng chuẩn của bảng Dimension.
   - Nhược điểm: Cấu trúc phi dạng chuẩn của lược đồ hình sao phù hợp hơn cho việc duyệt các chiều.
3. **Lược đồ ngân hà**
   Lược đồ ngân hà được hình thành nhờ sự kết hợp giữa lược đồ hình sao và lược đồ hình bông tuyết. Lược đồ này chứa nhiều bảng Fact sử dụng chung một số bảng Dim và là sự kết hợp của nhiều Data Mart.

2 Tổng quan về BI

2.1 Khái niệm

Kinh doanh thông minh (Business Intelligence - BI) là quy trình/hệ thống công nghệ cho phép phân tích và thể hiện thông tin giúp cho các nhà quản lý và người sử dụng của tổ chức đưa ra các quyết định phù hợp.
Kinh doanh thông minh bao gồm một loạt các công cụ, ứng dụng và phương thức cho phép các tổ chức thu thập thông tin từ các hệ thống nội bộ và bên ngoài; chuẩn bị sẵn sàng cho việc phân tích, phát triển và chạy các truy vấn đối với dữ liệu, tạo các báo cáo, bảng điều khiển và hình ảnh hóa dữ liệu để cung cấp kết quả phân tích cho những người sử dụng và những người ra quyết đinh.

2.2 Các bước trong quy trình kinh doanh thông minh



Hình 6: Quy trình kinh doanh thông minh

- **Nguồn dữ liệu (Data Source)** là vị trí bắt nguồn dữ liệu đang được sử dụng. Nguồn dữ liệu có thể là vị trí ban đầu nơi dữ liệu được sinh ra hoặc nơi thông tin vật lý được số hóa lần đầu tiên, tuy nhiên, ngay cả những dữ liệu tinh tế nhất cũng có thể đóng vai trò là nguồn, miễn là một quy trình khác truy cập và sử dụng nó. Cụ thể, nguồn dữ liệu có thể là một cơ sở dữ liệu đến từ các hệ quản trị cơ sở dữ liệu MySQL, SQL, Oracle, MSSQL, một tệp phẳng, các phép đo trực tiếp từ các thiết bị vật lý, dữ liệu web cóp nhặt hoặc bất kỳ dịch vụ dữ liệu trực tuyến và tĩnh nào có rất nhiều trên internet...
- **Kho dữ liệu (Data Warehouse)** là cơ sở dữ liệu được thiết kế theo mô hình Online Analytical Processing (OLAP), dữ liệu trong data warehouse chỉ có thể đọc, không được ghi hay xóa mà chỉ được update bởi gói ETL chuyển đổi dữ liệu từ Data Sources vào Data Warehouse.
- **Khám phá dữ liệu (Data Exploration)** là bước đầu tiên trong phân tích dữ liệu, trong đó người dùng khám phá một tập dữ liệu lớn theo cách không có cấu trúc để khám phá các mẫu, đặc điểm và điểm quan tâm ban đầu.Khám phá dữ liệu tạo ra các truy vấn, báo cáo, biểu đồ phân tích, thống kê từ mô hình dữ liệu OLAP ở Kho dữ liệu.
- **Khác dữ liệu (Data mining)** là quá trình phân tích khối lượng lớn dữ liệu để khám phá thông tin kinh doanh giúp các công ty giải quyết vấn đề, giảm thiểu rủi ro và nắm bắt cơ hội mới.

2.3 Lợi ích

Những lợi ích chính mà doanh nghiệp có thể nhận được từ các ứng dụng BI:

- Tăng tốc và cải thiện việc ra quyết định
- Tối ưu hóa quy trình kinh doanh nội bộ
- Phát hiện các vấn đề kinh doanh cần được giải quyết
- Xác định các xu hướng kinh doanh và thị trường mới nổi
- Phát triển các chiến lược kinh doanh mạnh mẽ hơn
- Thúc đẩy doanh số bán hàng cao hơn và doanh thu mới
- Đạt được lợi thế cạnh tranh so với các công ty đối thủ

Một số công cụ thông dụng hiện nay:

1. Power BI
2. Oracle BI
3. QlikView
4. Spago
5. Pentaho
6. IBM Cognos

2.4 Công cụ trực quan hóa dữ liệu Power BI

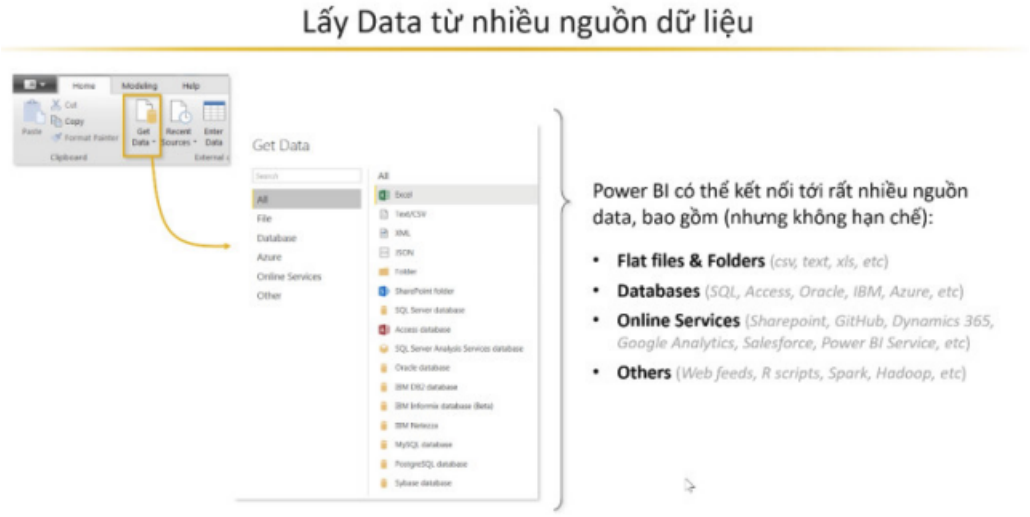
2.4.1 Giới thiệu chung

Power BI được ra đời vào năm 2011, được phát triển bởi Microsoft, sau đó nó được đưa vào sử dụng chính thức vào năm 2015. Power BI tập hợp rất nhiều các dịch vụ về phần mềm, các ứng dụng, các trình kết nối hoạt động song song cùng nhau để biến đổi các nguồn dữ liệu từ nhiều nguồn khác nhau thành các thông tin chi tiết liền mạch và trực quan. Power Bi được phát triển, sử dụng trên nền tảng Desktop, Website Service và Mobile App, nó hoàn toàn thân thiện và dễ dàng thích ứng với mọi người dùng mặc dù mỗi người đều có những nhu cầu khác nhau.

2.4.2 Các chức năng của Power BI

Kết nối dữ liệu từ nhiều nguồn
Chúng ta có thể truy cập dữ liệu từ nhiều nguồn khác nhau dựa trên nền tảng Power BI, bao gồm các nguồn như sau:

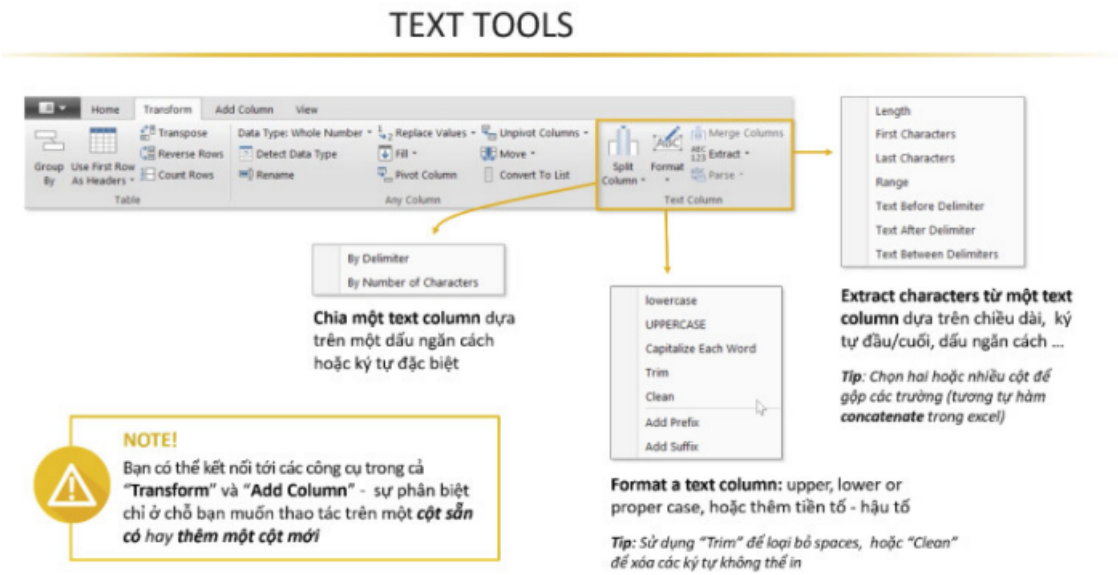
- File: các dạng Excel, Text/CSV, XML, JSON, Folder, PDF, SharePoint folder.
- Database: SQL Server database, Access database, Oracle database, IBM Db2 database, MySQL. . .
- Power Platform: Power BI datasets, Power BI dataflows, Common Data Service. . .
- Azure
- Online Services: SharePoint Online List, Microsoft Exchange Online, Dy-namics 365...



Hình 7: Kết nối dữ liệu trong PBI

Tiền xử lý dữ liệu

Hầu hết trong doanh nghiệp, dữ liệu thu thập được đều đã trải qua quá trình tiền xử lý dữ liệu để có thể sẵn sàng sử dụng tạo ra các báo cáo. Trong quá trình tiền xử lý dữ liệu, các dữ liệu được trích xuất từ một nguồn dữ liệu, sau đó được chuyển đổi, xác thực, chuẩn hóa, sửa chữa, kiểm tra và cuối cùng được tải vào kho dữ liệu.

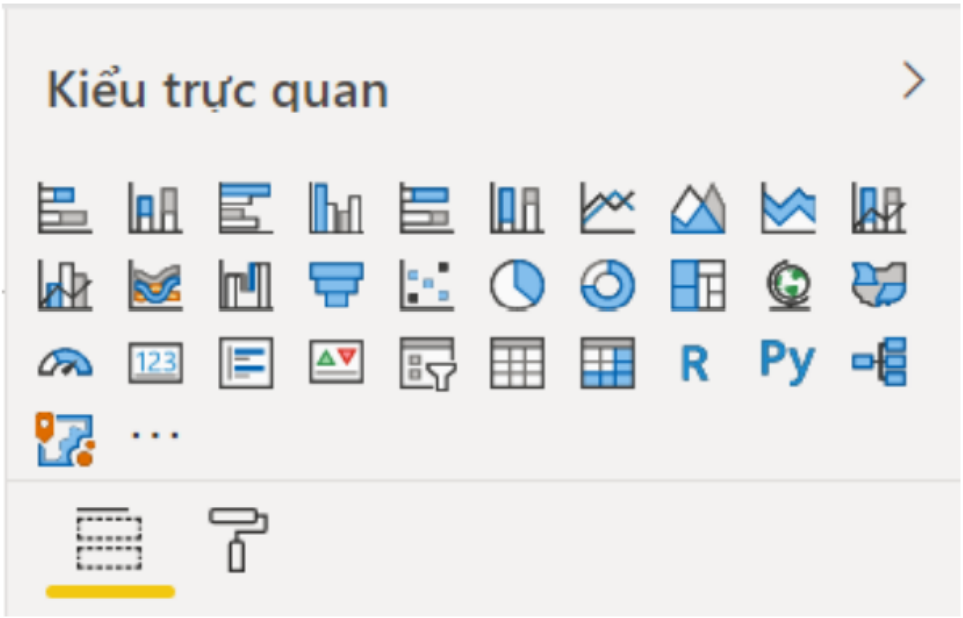


Hình 8: Kết nối dữ liệu trong PBI

Quy trình tiền xử lý dữ liệu được thực hiện bởi các ứng dụng như SQL Server Integration Services (SSIS) hoặc các công cụ của bên thứ ba khác. Tuy nhiên, trong một số doanh nghiệp, công việc tiền xử lý dữ liệu được thực hiện ngay trong Excel, được gọi là chuyển đổi dữ liệu. Tuy nhiên, quy trình ETL trong Excel là một quy trình thủ công, mất nhiều thời gian và khó có thể tự động hóa.
Vì thế Microsoft đã tạo ra công cụ để có thể làm cho quá trình này trở nên nhanh và dễ dàng hơn nhiều đó là Power Query, Power BI Desktop. Hai công cụ này cung cấp cho người dùng khả năng tự động hóa quá trình nhập, chuyển đổi và tải dữ liệu vào các bảng nội bộ trong Power BI, sau đó có thể được sử dụng làm nguồn cho các báo cáo hay dashboard của Power BI. Vì Power Query duy trì bản ghi từng bước của mọi hành động được thực hiện để nhập, chuyển đổi và tải dữ liệu, các bước này sẽ được lặp lại khi có thêm dữ liệu được thêm vào. **Mô hình hóa dữ liệu (Data modeling)** Mô hình hóa dữ liệu là quá trình tạo ra một biểu diễn trực quan của toàn bộ hệ thống thông tin hoặc các bộ phận của nó để giao tiếp các kết nối giữa các điểm và cấu trúc dữ liệu. Mục đích là minh họa các loại dữ liệu được sử dụng và lưu trữ trong hệ thống, mối quan hệ giữa các loại dữ liệu này, cách dữ liệu có thể được nhóm và tổ chức cũng như các định dạng và thuộc tính của nó.
Dữ liệu có thể được mô hình hóa ở nhiều mức độ trừu tượng khác nhau. Quá trình bắt đầu bằng cách thu thập thông tin về các yêu cầu kinh doanh từ các bên liên quan và người dùng cuối. Các quy tắc nghiệp vụ này sau đó được chuyển thành cấu trúc dữ liệu để hình thành một thiết kế cơ sở dữ liệu cụ thể.
Mô hình dữ liệu có thể được so sánh với lộ trình, bản thiết kế của kiến trúc sư hoặc bất kỳ sơ đồ chính thức nào giúp hiểu sâu hơn về những gì đang được thiết kế.
Mô hình hóa dữ liệu sử dụng các lược đồ chuẩn hóa và các kỹ thuật chính thức. Điều này cung cấp một cách chung, nhất quán và có thể dự đoán được để xác định và quản lý tài nguyên dữ liệu trong một tổ chức hoặc thậm chí xa hơn.
Quy trình mô hình hóa dữ liệu:

- Xác định các thực thể
- Xác định các thuộc tính chính của từng thực thể
- Xác định mối quan hệ giữa các thực thể
- Ánh xạ các thuộc tính cho các thực thể hoàn toàn
- Gán các khóa khi cần thiết và quyết định mức độ chuẩn hóa cân bằng giữa nhu cầu giảm dư thừa với các yêu cầu về hiệu suất.
- Hoàn thiện và xác thực mô hình dữ liệu.

Trực quan hóa dữ liệu



Hình 9: Kết nối dữ liệu trong PBI

Trực quan hóa dữ liệu là biểu diễn đồ họa của thông tin và dữ liệu. Bằng cách sử dụng các yếu tố trực quan như biểu đồ, đồ thị và bản đồ, các công cụ trực quan hóa dữ liệu cung cấp một cách dễ tiếp cận để xem và hiểu các xu hướng, ngoại lệ và mẫu trong dữ liệu.
Trong thế giới của Dữ liệu lớn, các công cụ và công nghệ trực quan hóa dữ liệu là rất cần thiết để phân tích một lượng lớn thông tin và đưa ra các quyết định dựa trên dữ liệu.
Trực quan hóa dữ liệu giúp bạn biến tất cả dữ liệu chi tiết đó thành thông tin kinh doanh dễ hiểu, hấp dẫn về mặt hình ảnh — và hữu ích.
Trực quan hóa dữ liệu làm cho dữ liệu trở nên sống động, khiến bạn trở thành người kể chuyện bậc thầy về những thông tin chi tiết ẩn trong các con số của bạn. Thông qua trang tổng quan trực tiếp, báo cáo tương tác, biểu đồ, đồ thị và các biểu thị trực quan khác, trực quan hóa dữ liệu giúp người dùng phát triển thông tin chi tiết mạnh mẽ về doanh nghiệp một cách nhanh chóng và hiệu quả.

3 Ứng dụng Data Warehouse và BI vào bài toán

3.1 Giới thiệu bài toán

3.1.1 Đặt vấn đề

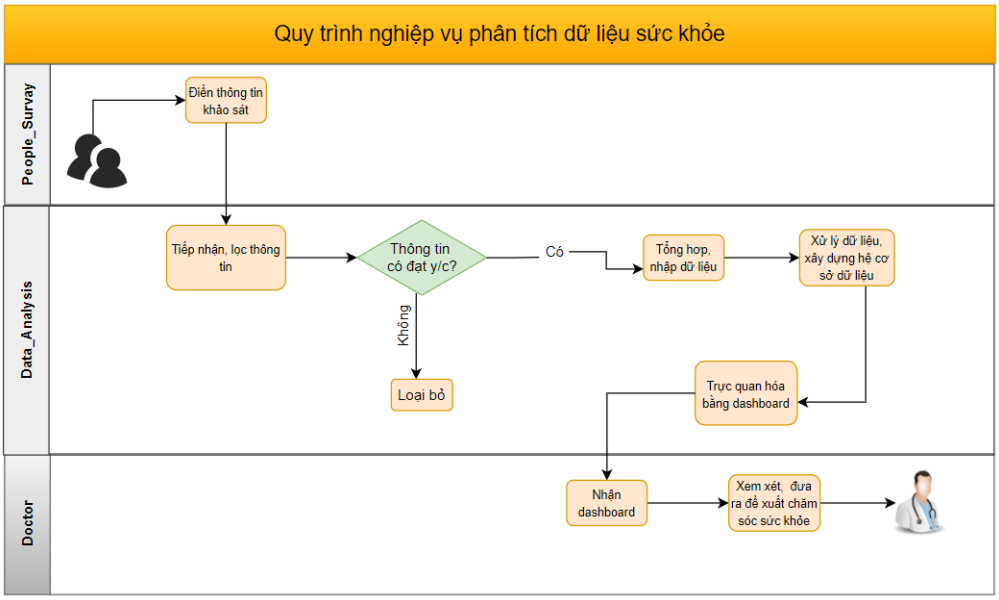
Tổng quan về Health care:
Health care hay chăm sóc sức khỏe là sự cải thiện sức khỏe thông qua phòng ngừa, chuẩn đoán, điều trị, cải thiện hoặc chữa bệnh, chấn thương và suy giảm thể chất và tinh thần khác ở người bệnh.

Tổng quan về ảnh hưởng của thuốc lá đến phổi:
Thuốc lá là nguyên nhân gây ra nhiều bệnh như: viêm họng, viêm phế quản, viêm phế quản phổi, bệnh phổi tắc nghẽn mạn tính, hen, ung thư phổi, ...
Bệnh phổi tắc nghẽn mạn tính là chỉ những tổn thương ở phổi có liên quan đến sự tắc nghẽn đường thở không phục hồi hoàn toàn. Mối liên quan giữa bệnh phổi tắc nghẽn mạn tính và hút thuốc cũng mạnh như với ung thư phổi. Thuốc lá là nguyên nhân quan trọng nhất gây ra bệnh phổi tắc nghẽn mạn tính, có khoảng 15% những người hút thuốc lá sẽ có triệu chứng lâm sàng của bệnh phổi tắc nghẽn mạn tính và 80 - 90% người mắc bệnh phổi tắc nghẽn mạn tính là nghiện thuốc lá.
Ở hầu hết các nước, thuốc lá là nguyên nhân gây ra hơn 90% ca tử vong vì ung thư phổi. Trung bình người hút thuốc làm tăng nguy cơ bị ung thư phổi lên từ 5 đến 10 lần so với người không hút thuốc.

Lợi ích của kho dữ liệu đối với chăm sóc sức khỏe

- Báo cáo hiệu quả.
- Quyết định lâm sàng tốt hơn.
- Yêu cầu và thanh toán bảo hiểm được tối ưu hóa.
- Cải thiện kinh nghiệm và kết quả của bệnh nhân.
- Chăm sóc dựa trên giá trị được cá nhân hóa.
- Lập kế hoạch chiến lược nâng cao.

3.1.2 Quy trình nghiệp vụ



Hình 10: Quy trình nghiệp vụ

Quy trình xử lý

- Đầu tiên, mọi người điền thông tin khảo sát.
- Sau đó, Data analysis sẽ tiến hành tiếp nhận và lọc thông tin, những thông tin không đạt yêu cầu sẽ bị loại bỏ (Ví dụ những bản điền của người nào thiếu thông tin, thông tin điền không đúng ...), còn những thông tin đạt yêu cầu sẽ được tổng hợp, và xây dựng thành 1 bộ dữ liệu.
- Data analysis trực quan hóa bộ dữ liệu bằng dashboard và chuyển dashboard đến bác sĩ.
- Bác sĩ nhận dashboard, dựa vào đó để đưa ra những xem xét và đánh giá tình hình cũng như cách cải thiện sức khỏe của từng bệnh nhân.

3.1.3 Quy mô dữ liệu

Giới thiệu về bộ dữ liệu thì được chúng em lấy từ hệ thống giám sát yếu tố rủi ro hành vi là hệ thống điều tra qua điện thoại liên quan đến sức khỏe hàng đầu của quốc gia BRFSS. Mục tiêu của hệ thống BRFSS: Thu thập dữ liệu tiểu bang về cư dân Hoa Kỳ về các hành vi nguy cơ liên quan đến sức khỏe, tình trạng sức khỏe mãn tính và việc sử dụng các dịch vụ phòng ngừa.
Kích thước bộ dữ liệu:

- Dữ liệu gồm:
  - 3 file dữ liệu tương ứng với từng năm 2013 - 2015
  - Mỗi file chứa 1.5 triệu bản ghi.
- Dữ liệu chủ yếu là dạng có cấu trúc
  - Kích thước: 114MB
- Chọn phân tích:
  - Dữ liệu khảo sát tại 6 tiểu bang: Alaska, California, Massachusetts, New York, Texas, Washington.
  - 26 cột dữ liệu liên quan đến đối tượng khảo sát và vấn đề bệnh phổi mãn tính.

Mô tả 1 số trường dữ liệu quan trọng

DATE	Ngày khảo sát
IDATE	Mã ngày khảo sát
STATE_ID	Mã các bang
SEX	Giới tính người tham gia khảo sát
MARTAL	Tình trạng hôn nhân của người tham gia khảo sát
EDUCA	Trình độ học vấn của người được khảo sát
EMPLOY	Nơi làm việc của người tham gia khảo sát
INCOME	Tổng thu nhập hàng năm của họ
AGE	Độ tuổi
AGE_GROUP	Người tham gia khảo sát thuộc nhóm tuổi nào
SMOKDAY	Tình trạng hút thuốc lá hiện nay ntn?
STOPMSK	Trong 12 tháng qua, có ngừng hút thuốc không
SMOKE100	Đã hút ít nhất 100 điếu thuốc hay chưa
LASTMK	Bao lâu rồi chưa hút thuốc lá
USENOW	Hiện tại có sử dụng thuốc lá nào không
GENHLTH	Cảm nhận của người khảo sát về tình hình chung của sức khỏe
HLTHPLN	Có loại bảo hiểm sức khỏe nào không?
PERSDOC	Có bác sĩ riêng chăm sóc sức khỏe không
MEDCOST	Trong 12 tháng qua, có lần nào cần đi khám nhưng không đi vì vấn đề chi phí hay không?
CHECKUP	Bao lâu rồi chưa kiểm tra sức khỏe định kỳ
CHCCOPD	Từng bị bệnh tắc nghẽn phổi mãn tính hoặc COPD, viêm phế quản mãn tính
WEIGHT	Cân nặng của người được khảo sát
HEIGHT	Chiều cao cảu người được khảo sát

Hình 11: Quy trình nghiệp vụ

3.1.4 Requirements

- **Input:** Dữ liệu khảo sát về ảnh hưởng của thuốc lá đối với phổi trong năm 2013 – 2015.

- **Output:** Dashboard trả lời những câu hỏi:

  1. Phân tích và khảo sát thông tin người tham gia khảo sát:
     - Giới tính
     - Độ tuổi
     - Công việc
     - Tình trạng hôn nhân
     - Học vấn
     - ...
  2. Phân tích và khảo sát tình trạng sức khỏe và khả năng tiếp cận y tế:
     - Tình trạng sức khỏe
     - Có bảo hiểm y tế không?
     - Có từng mắc bệnh phổi mãn tính không?
     - Bao lâu chưa đi khám sức khỏe?
     - ...
  3. Phân tích và khảo sát thực trạng của việc hút thuốc lá:
     - Có hút thuốc lá không?
     - Tần suất sử dụng thuốc lá?
     - Đã từng hút ít nhất 100 điếu thuốc chưa?
     - Đã từng có ý định bỏ thuốc lá chưa?
     - ...

3.2 Phân tích và thiết kế hệ thống

3.2.1 Kiến trúc Dataware house

1. **Kiến trúc cũ:**

Hình 12: Kiến trúc DW cũ

Mô hình cũ thực hiện các hoạt động phân tích dữ liệu ngay trên hệ thống lưu trữ. Điều này đem lại một số hạn chế như:

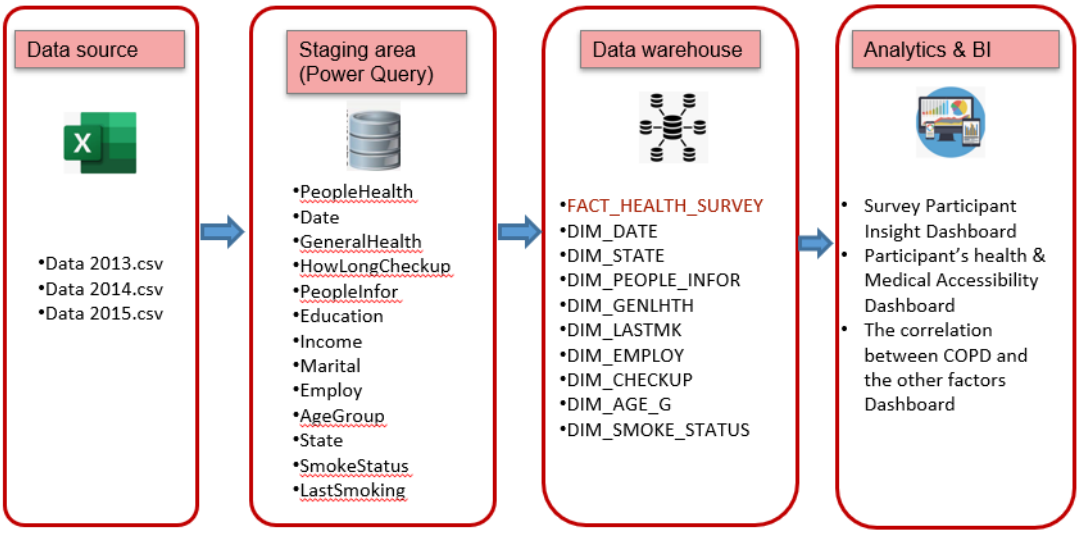
- Làm giảm hiệu năng hệ thống
- Dữ liệu để phân tích không ổn định
- Tốc độ xử lý chậm.

Chính vì thế ta cần xây dựng mô hình mới để khắc phục được những hạn chế này:

2. **Kiến trúc mới:**

Đặc điểm hệ thống mới: Sử dụng kiến trúc có vùng Staging, trong đó:

- Nguồn dữ liệu là file CSV chứa dữ liệu thô.
- Vùng Staging là khu vực lưu trữ trung gian được sử dụng để thực hiện các bước ETL dữ liệu, ví dụ như xử lý dữ liệu Null, dữ liệu đa trị, định dạng kiểu dữ liệu, tách thành các bảng,...
- Data Warehouse là nơi dữ liệu được lưu trữ và quản lí để phục vụ cho việc phân tích thống kê, báo cáo, khai thác và trực quan hóa dữ liệu.
- Lớp phân tích được sử dụng để đưa ra các báo cáo và dashboard theo các chủ đề đã nêu thông qua một số công cụ trực quan hoá dữ liệu.

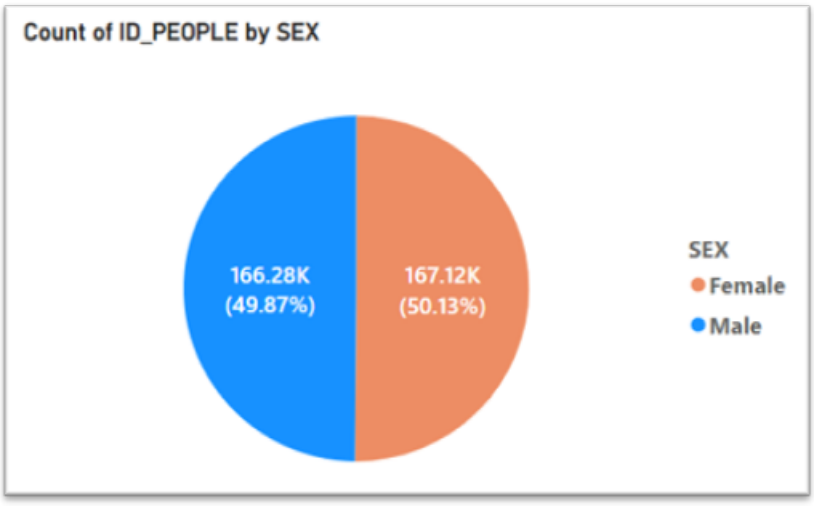


Hình 13: Kiến trúc DW mới

3.2.2 Data Exploration

Thông tin của người thực hiện khảo sát

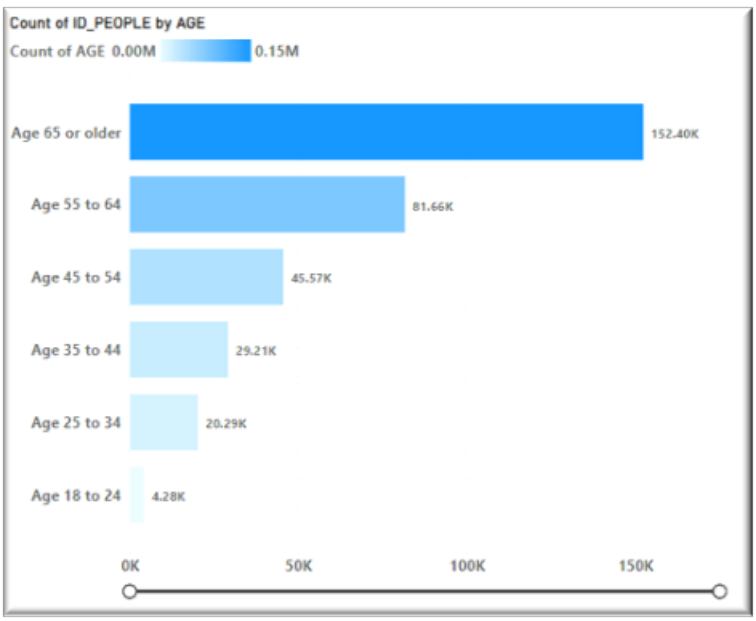
– Tỷ lệ giới tính



Hình 14: Tỷ lệ giới tính

+ Từ biểu đồ ta thấy tỷ lệ giới tính của số người tham gia khảo sát là gần như ngang nhau : với tỷ lệ "Male" chiếm 49.78% và tỷ lệ "Female" chiếm 50.13%.

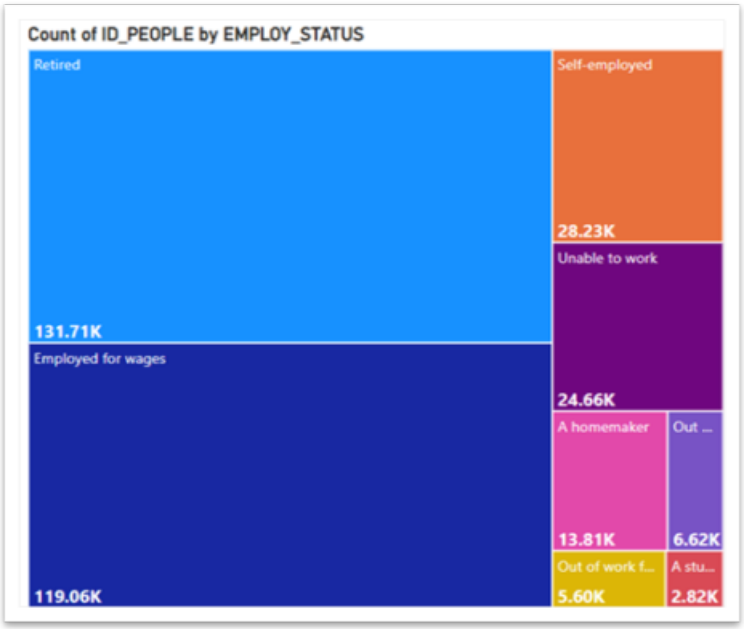
– Độ tuổi tham gia khảo sát



Hình 15: Độ tuổi tham gia khảo sát

+ Ở khảo sát này thì có chia làm 6 nhóm tuổi và bắt đầu từ 18 tuổi trở lên.
+ Nhìn vào biểu đồ ta có thể thấy số lượng người tham gia khảo sát ở độ tuổi 65 tuổi trở lên rất cao lên đến 152.40K người, và số lượng người tham gia giảm dần theo chiều giảm của độ tuổi. Điều này cho thấy càng lớn tuổi con người càng có xu hướng quan tâm đến sức khỏe bản thân nhiều hơn.

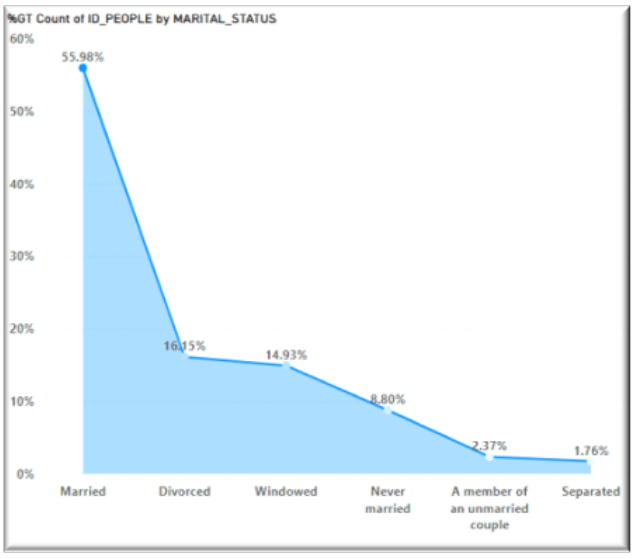
– Tình trạng việc làm



Hình 16: Tình trạng việc làm

+ Cũng tương tự như Độ tuổi tham gia khảo sát, có số lượng người từ 65 tuổi trở lên chiếm đa số thì tình trạng việc làm tương ứng là "Retired" (Đã nghỉ hưu) cũng chiếm đa số, lên đến 131.71K người.
+ Số lượng người làm công ăn lương cũng chiếm một phần không nhỏ lên đến 119.06K người và ít nhất là học sinh, sinh viên.

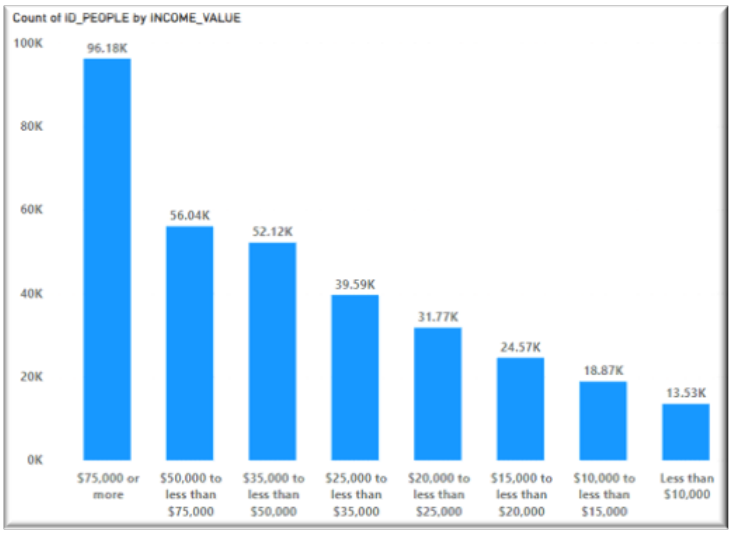
– Tình trạng hôn nhân



Hình 17: Tình trạng hôn nhân

+ Tình trạng hôn nhân của người tham gia khảo sát được chia thành 6 nhóm : Đã kết hôn, Đã ly hôn , Góa vợ hoặc chồng, Không bao giờ kết hôn, Chưa kết hôn, Ly thân.
+ Trong đó tỷ lệ người đã kết hôn chiếm đến 55.98% điều đó cũng do hầu hết người được khảo sát từ 18 tuổi trở lên và phần lớn mọi người đã đủ khả năng để kết hôn.

– Mức thu nhập

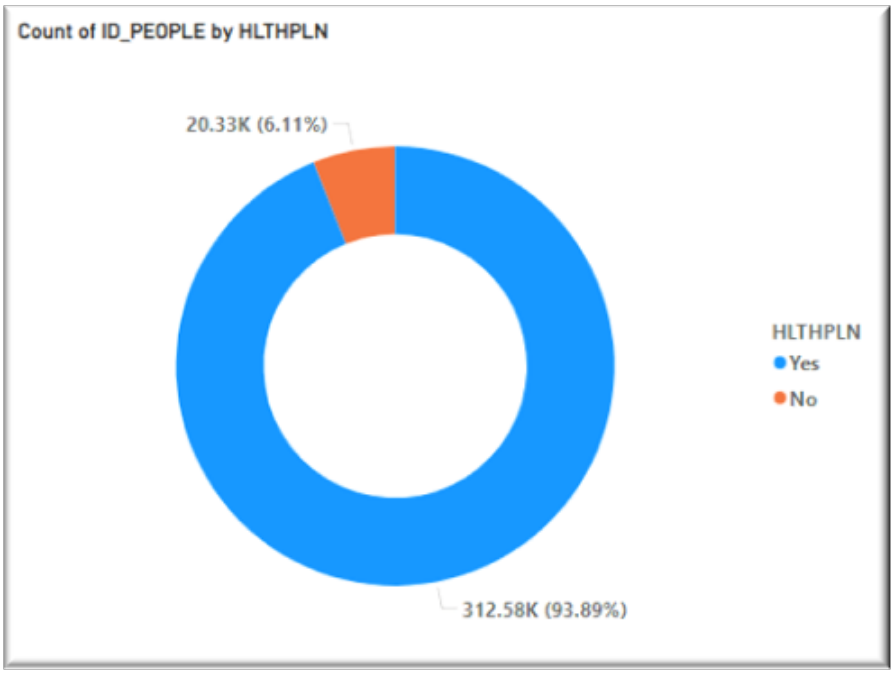


Hình 18: Mức thu nhập

+ Mức thu nhập được chia thành 8 nhóm trải dài từ ít hơn $10.000 đến nhiều hơn $75000.
+ Số người có mức thu nhất $75.000 hoặc lớn hơn chiếm đa số lên đến 96.18K người, và giảm dần theo chiều giảm của mức thu nhập. Điều đó cho ta thấy phần lớn người tham gia khảo sát có thu nhập khá cao

Khả năng tiếp cận y tế

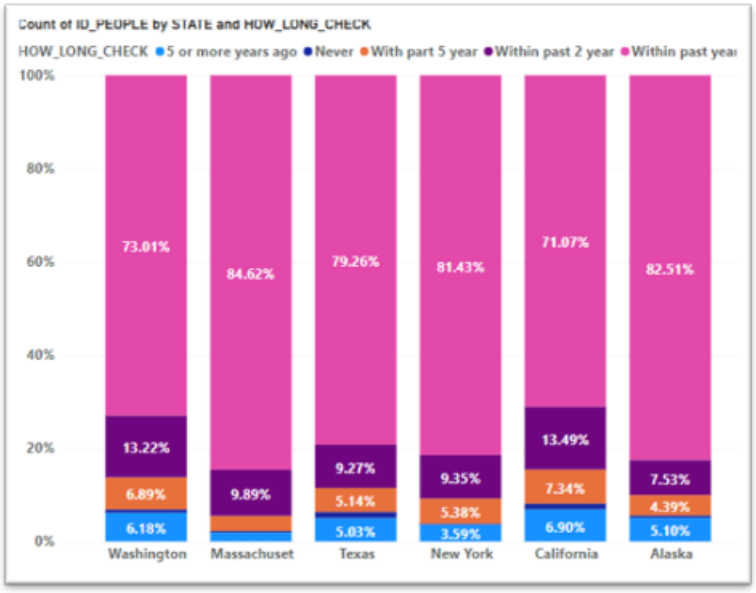
– Tỷ lệ người có bảo hiểm y tế



Hình 19: Bảo hiểm y tế

+ Số lượng người có bảo hiểm y tế lên đến 312.58K chiếm đến 93.89%, điều này hoàn toàn phù hợp với mức thu nhập cao của số người tham gia khảo sát.
+ Tuy nhiên vẫn còn đến 20.33K người chưa có bảo hiểm y tế, cho thấy khả năng tiếp cận y tế của họ vẫn chưa cao.

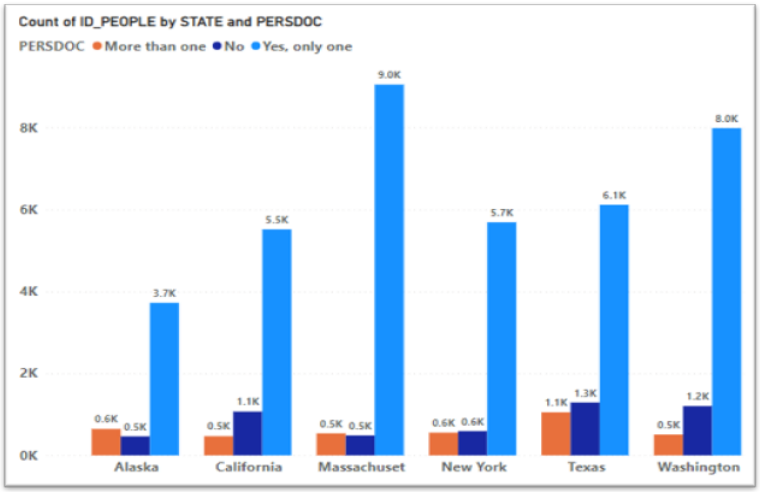
– Tỷ lệ giữa các khoảng thời gian chưa đi kiểm tra sức khỏe theo từng khu vực



Hình 20: Thời gian đi kiểm tra sức khỏe

+ Thời gian đi kiểm tra sức khỏe được chia làm 4 nhóm theo độ dài của thời gian.
+ Ta có thể thấy phần lớn mọi người đi kiểm tra sức khỏe trước đó chưa đến 1 năm
+ Bên cạnh đó vẫn có nhiều người khám sức khỏe trước đó 2 năm, 5 năm và thậm chí là nhiều hơn 5 năm.

– Số người có bác sĩ riêng theo từng khu vực

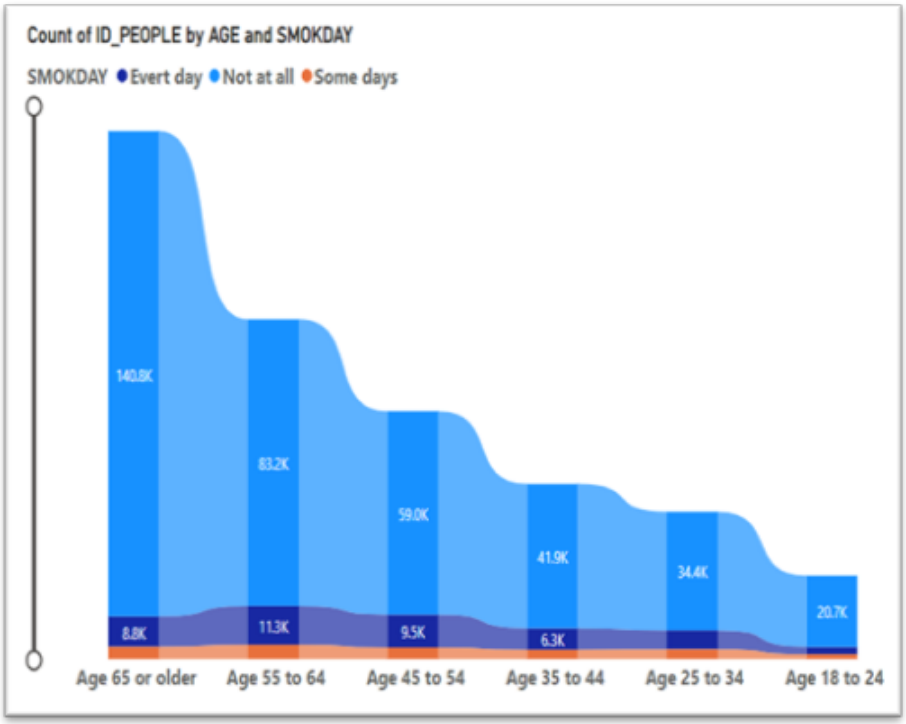


Hình 21: Bác sĩ riêng

+ Nhìn vào báo cáo ta có thể thấy, phần lớn người tham gia khảo sat có 1 bác sĩ riêng, và cũng có một phần người có nhiều hơn 1 bác sĩ riêng.
+ Có một phần người tham gia khảo sát không có bác sĩ riêng, họ là những người có mức thu nhập thấp và có khả năng tiếp cận y tế thấp.

Thực trạng của việc hút thuốc lá

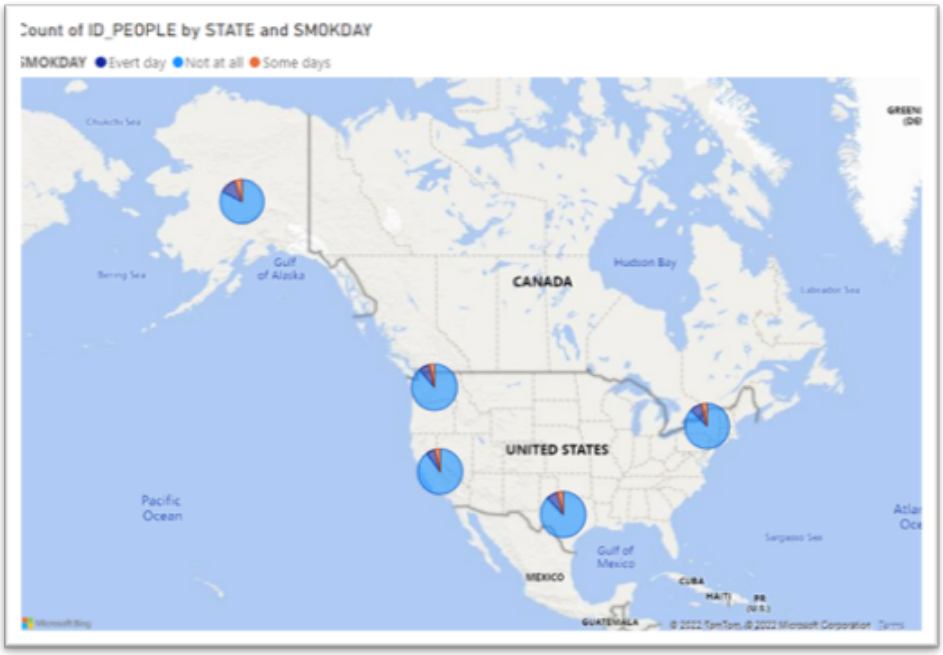
– Số người sử dụng thuốc lá theo từng nhóm tuổi



Hình 22: Theo độ tuổi

+ Biểu đồ cho thấy phần lớn người tham gia khảo sát không hút thuốc lá
+ Nhưng vẫn có một bộ phận người hút thuốc lá hàng ngày và cao nhất ở nhóm tuổi từ 55 đến 64.

– Tình trạng hút thuốc lá ở từng khu vực



Hình 23: Theo khu vực

+ Tỉ lệ hút thuốc lá ở các khu vực là gần như nhau, phần lớn mọi người đều không sử dụng thuốc lá.

3.2.3 ETL

Dữ liệu gốc chứa gần 1,5 triệu bản ghi các dữ liệu đều đã được mã hóa dưới dạng số. Tha thực hiện ETL dữ liệu bằng Excel gồm các thao tác như thay thế giá trị , đổi kiểu dữ liệu , tách cột , xử lý các cột ngày tháng Đầu tiên đưa dữ liệu vào Power query .Nhận thấy có 3 file dữ liệu của năm 2013 , 2014, 2015 đều có các cột tương ứng giống nhau . Để giúp cho việc thống kê và báo cáo được dễ dàng hơn ta thực hiện nối các bảng lại với nhau bằng cách sử dụng *Append Quieries* .

Xử lý dữ liệu mã hóa

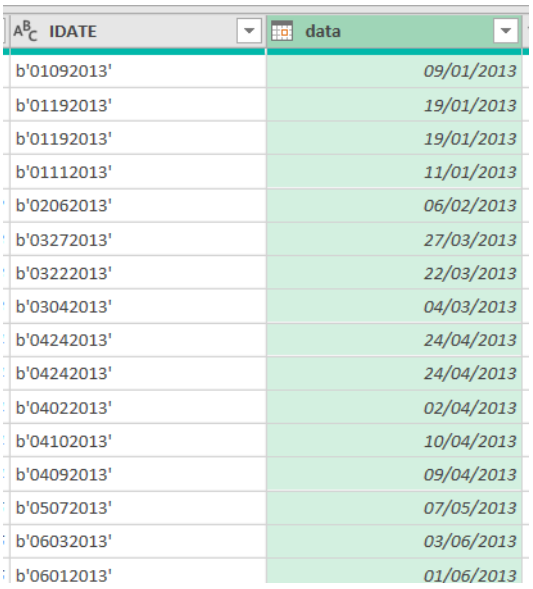
Các dữ liệu trong bảng 2013, 2014,2015 đều đã được mã hóa dưới dạng số. Do đó cần phải giải mã chúng để đưa về các giá trị cụ thể chi tiết hơn. Có rất nhiều cách để giải mã như tách cột, thêm cột, thêm cột có điều kiện, thay thế giá trị, ... Cụ thể:

- Giải mã dữ liệu cột IDAY , IMONTH , IYEAR
  Nhận thấy dữ liệu trong các cột *IDAY, IMONTH ,IYEAR* đều được bao bọc bởi cặp dấu ” do đó cần phải tách dữ liệu ra bằng cách xử dụng *Split Column* với dấu phân cách là ’. Ta thu được kết quả khi tách cột *IMONTH* như hình dưới :

IMONTH.1	IMONTH.2	IMONTH.3	IDAY
b	1		b'09'
b	1		b'19'
b	1		b'19'
b	1		b'11'
b	2		b'06'
b	3		b'27'
b	3		b'22'
b	3		b'04'
b	4		b'24'
b	4		b'24'
b	4		b'02'
b	4		b'10'
b	4		b'09'
b	5		b'07'
b	6		b'03'
b	6		b'01'
b	6		b'13'
b	5		b'13'
b	6		b'22'

Hình 24: Tách dữ liệu cột IMONTH

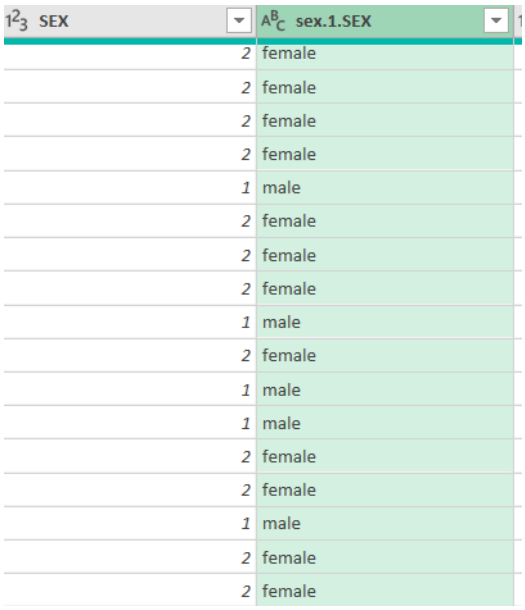
Sau khi đã tách dữ liệu của 3 cột ta thực hiện xóa các cột IMONTH.1,IMONTH.3 , IYEAR.1 ,IYEAR.3 ,IDAY.1 ,IDAY.3 bởi vì các cột này đều không có giá trị sử dụng . Thực hiện gộp các cột *IDAY.2 , IMONTH.2 ,IYEAR.2* với nhau bằng cách sử dụng *merge columns* với seperator là / , sau đó chuyển dữ liệu về dạng ngày tháng ta thu được kết quả như hình dưới đây :



IDATE	data
b'01092013'	09/01/2013
b'01192013'	19/01/2013
b'01192013'	19/01/2013
b'01112013'	11/01/2013
b'02062013'	06/02/2013
b'03272013'	27/03/2013
b'03222013'	22/03/2013
b'03042013'	04/03/2013
b'04242013'	24/04/2013
b'04242013'	24/04/2013
b'04022013'	02/04/2013
b'04102013'	10/04/2013
b'04092013'	09/04/2013
b'05072013'	07/05/2013
b'06032013'	03/06/2013
b'06012013'	01/06/2013

Hình 25: Merge quieres

- Giải mã hóa các cột giá trị khác
  Thực hiện thay thế các giá trị số thành các giá trị cụ thể ví dụ như ở cột **SEX** chỉ có 2 giá trị số là 1 và 2 , nhìn vào giá trị ta không thể biết được nó có nghĩa là gì do đó ta cần giải mã nó với 1 tương ứng là *male* , với 2 tương ứng là *female* . Thực hiện giải mã bằng cách thay thế giá trị (sử dụng replace values) hoặc megre quieres bảng gốc với bảng dữ liệu giải mã tương ứng . Tuy nhiên khi dữ liệu bị mã hóa có nhiều việc thay thế giá trị sẽ rất là tốn thời gian do đó ta thực hiện merge quieres thu được kết quả như hình dưới

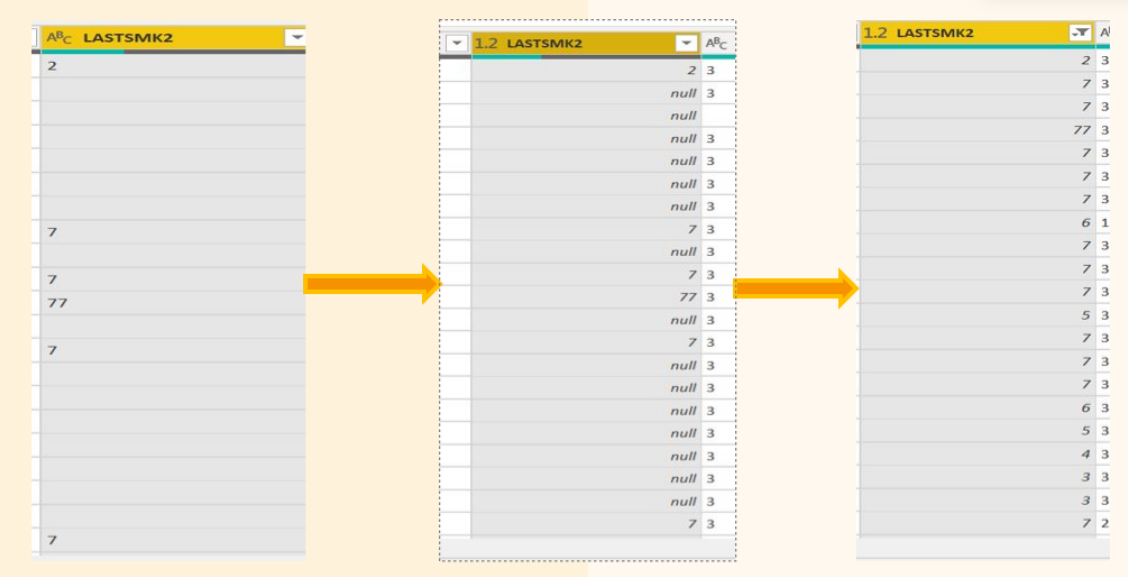


SEX	sex.1.SEX
2	female
2	female
2	female
2	female
1	male
2	female
2	female
2	female
1	male
2	female
1	male
1	male
2	female
2	female
1	male
2	female
2	female

Hình 26: Merge quieres

Xóa các giá trị *NULL* trong bảng

Thực hiện xóa các giá trị *NULL* ở mỗi bảng .Các giá trị này sẽ làm cho việc báo cáo thống kê phức tạp do đó ta cần thực thao tác xóa . Dưới đây là kết quả của quá trình xóa các giá trị null trong cột **LASTSMK2** :

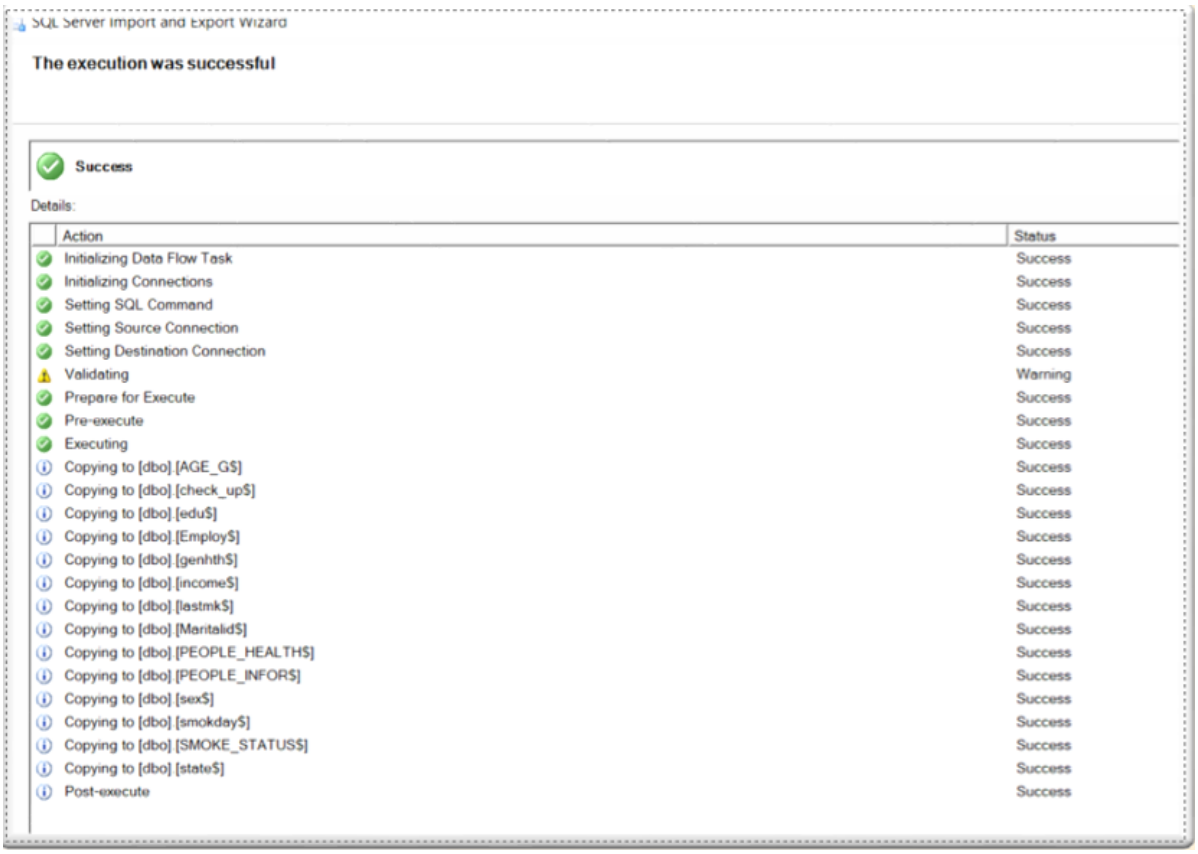


Hình 27: Xóa giá trị NULL cột LASTSMK2

Ngoài ra để tiện cho việc truy vấn dữ liệu ta có thể lưu dữ liệu vào cơ sở dữ liệu:

- Dữ liệu sau khi được xử lí đơn giản được lưu vào CSDL HealthCare, vùng dữ liệu được lưu này gọi là Stagging, là vùng đệm chứa dữ liệu trước khi đưa vào kho dữ liệu.
- Xây dựng kho dữ liệu bằng cách tạo các bảng trong cơ sở dữ liệu theo mô hình OLAP đã được thiết kế.

Đổ dữ liệu từ cơ sở dữ liệu vào kho dữ liệu bằng cách sử dụng SQL Server Import and Export Wizard trong SQL sever:



Hình 28: IMPORT DATA

- Các dữ liệu cố định, không thay đổi theo thời gian sẽ được chuyển trực tiếp từ Stagging vào Kho dữ liệu thông qua câu lệnh "Insert" như bên dưới :

```
/* đổ dữ liệu vào kho dữ liệu */
/* đổ dữ liệu vào bảng dim_AGE_G */
INSERT INTO dim_AGE_G SELECT * FROM AGE_G$
insert into dim_CHECKUP select * from check_up$
insert into dim_EMPLOY select * from Employ$
insert into dim_GENHLTH select * from genhth$
insert into dim_LASTMK select * from lastmk$
insert into dim_STATE select * from state$
/* đổ dữ liệu vào bảng [dim_PEOPLE_INFOR] */
INSERT INTO [dbo].[dim_PEOPLE_INFOR] (people_ID,sex ,Marinal_status ,income_value , emloy , age ,age_group)
select  a.people_ID , a.SEX , b.MARITAL , c.INCOME_VALUE,employ , a.Age , a.AGE_G
from PEOPLE_INFOR$ a    left join dim_MARITAL  b on  b.ID= a.MARITAL
                        left join dim_INCOME c on c.ID=a.INCOME
/* đổ dữ liệu vào bảng [dim_SMOKE_STATUS]  */
insert into [dbo].[dim_SMOKE_STATUS] ([status_IDT] ,[SMOKDAY] , [STOPSMK] , [SMOKE100] , [LASTMK] )
select [status_ID] , [SMOKDAY] , [STOPSMK] , [SMOKE100] , [LASTSMK]
from [dbo].[SMOKE_STATUS$]
```

Hình 29: Câu lệnh INSERT

- Đối với các dữ liệu được cập nhật định kì theo thời gian (tức là cứ sau một khoảngthời gian sẽ có dữ liệu) mới được thêm vào thì việc cứ mỗi lần phải viết lại câu lệnh "Insert" là rất mất thời gian. Thay vào đó, ta sẽ đưa câu lệnh "Insert" tạo thành một thủ tục để mỗi lần cập nhật dữ liệu ta chỉ cần gọi thủ thục một cách nhanh chóng:

```
/* xây dựng thủ tục đổ dữ liệu vào bảng dim_age_G */
create procedure insert_age_G
as
begin
    INSERT INTO dim_AGE_G SELECT * FROM AGE_G$
    where AGE_G$.AGE_G not in (select ID from   dim_AGE_G  )
end

/* xây dựng thủ tục đổ dữ liệu vào bảng dim_PEOPLE_INFOR2 */
create procedure insert_dim_PEOPLE_INFOR2
as
begin
    INSERT INTO [dbo].[dim_PEOPLE_INFOR] (people_ID,sex ,Marinal_status ,income_value , emloy , age ,age_group)

    select  a.people_ID , a.SEX , b.MARITAL , c.INCOME_VALUE,employ , a.Age , a.AGE_G

    from PEOPLE_INFOR$ a    left join dim_MARITAL  b on  b.ID= a.MARITAL

                        left join dim_INCOME c on c.ID=a.INCOME
    where a.people_ID not in (select people_ID from [dbo].[dim_PEOPLE_INFOR]  )
end
```

Hình 30: Thủ tục INSERT

Sau khi đã xây dựng xong các thủ tục ta thử kiểm tra lại kết quả :

```
execute insert_dim_PEOPLE_INFOR2
select * from dim_PEOPLE_INFOR
```

75 %

Results Messages

	people_ID	sex	Marinal_status	income_value	emloy	age	age_group
1	1	female	Divorced	$50,000 to less than $75,000	7	60	5
2	2	female	Married	$75,000 or more	1	50	4
3	3	female	Married	$75,000 or more	1	55	5
4	4	female	Married	$50,000 to less than $75,000	7	64	5
5	5	male	Married	$35,000 to less than $50,000	7	66	6
6	6	female	Divorced	$75,000 or more	1	49	4
7	7	female	Married	NULL	1	39	3
8	8	female	Windowed	$35,000 to less than $50,000	7	60	5
9	9	male	Married	$75,000 or more	7	50	4
10	10	female	Married	$20,000 to less than $25,000	5	68	6
11	11	male	Married	$75,000 or more	1	44	3
12	12	male	Married	Less than $10,000	4	57	5
13	13	female	Windowed	$50,000 to less than $75,000	7	73	6
14	14	female	Married	NULL	7	63	5
15	15	male	Married	NULL	8	65	6
16	16	female	Divorced	$75,000 or more	7	70	6
17	17	female	Divorced	$15,000 to less than $20,000	1	32	2

Hình 31: Kết quả

3.2.4 Facts

Chủ điểm phân tích của hệ thống được mô tả trong bảng "FACT_HEALTH_SURVEY" với khoá chính là SURVEY_ID và 4 khoá phụ bao gồm: PEOPLE_ID, REGION_ID, DATE_ID, STATUS_ID cùng với thuộc tính CHCCOPD.

STT	Thuộc tính	Tên thuộc tính	Kiểu
1	PK	SURVEY_ID	int
2	FK	PEOPLE_ID	int
3	FK	REGION_ID	int
4	FK	DATE_ID	nvarchar(255)
5	FK	STATUS_ID	int
6		HCCOPD	nvarchar(255)

Hình 32: FACT_HEALTH_SURVEY

3.2.5 Dimension

Theo chủ điểm phân tích FACT_HEALTH_SURVEY" ta có các chiều phân tích theo thời gian, tính trạng sức khoẻ, thông tin của người tham gia khảo sát, tình trạng hút thuốc và khu vực.

- Chiều về thời gian
  Chiều về thời gian bao gồm thời gian tham gia khảo sát với khoá chính là DATE_ID.

STT	Thuộc tính	Tên thuộc tính	Kiểu
1	PK	DATE_ID	nvarchar(255)
2		DATE_ID	datetime

Hình 33: OLAP DIM_DATE

- Chiều về tình trạng sức khoẻ của người tham gia khảo sát
  Có khoá chính là PEOPLE_ID và 7 thuộc tính lần lượt là: HEALTH_STATUS, HLTHPLN, PERSDOC, MEDCOST, HOW_LONG_CHECKUP, WEIGHT_KG, HEIGHT_CM.

STT	Thuộc tính	Tên thuộc tính	Kiểu
1	**PK**	PEOPLE_ID	int
2		HEALTH_STATUS	nvarchar(255)
3		HLTHPLN	nvarchar(255)
4		PERSDOC	nvarchar(255)
5		MEDCOST	nvarchar(255)
6		HOW_LONG_CHECKUP	nvarchar(255)
7		WEIGHT_KG	float
8		HEIGHT_CM	float

Hình 34: OLAP DIM_HEALTH

- Chiều về thông tin người tham gia khảo sát
  Được phân cấp thành 3 bảng như sau:

STT	Thuộc tính	Tên thuộc tính	Kiểu
1	**PK**	PEOPLE_ID	int
2	**FK**	SEX	nvarchar(255)
3	**FK**	EDUCA	int
4		EMPLOY	int
5		MARITAL_STATUS	nvarchar(255)
6		INCOME_VALUE	nvarchar(255)
7		AGE	int
8		AGE_GROUP	nvarchar(255)

Hình 35: OLAP DIM_PEOPLE_INFOR

STT	Thuộc tính	Tên thuộc tính	Kiểu
1	**PK**	EMPLOY	int
2		EMPLOY_STATUS	nvarchar(255)

Hình 36: OLAP DIM_EMPLOY

STT	Thuộc tính	Tên thuộc tính	Kiểu
1	PK	EDUCA	int
2		GRADE	nvarchar(255)

Hình 37: OLAP DIM_EDUCATION

- Chiều về tình trạng hút thuốc của người tham gia khảo sát
  Gồm một khoá chính STATUS_ID và 5 thuộc tính: SMOKEDAY, STOPSMK, SMOKE100, HOW_LONG_LAST_SMOKE, USENOW.

STT	Thuộc tính	Tên thuộc tính	Kiểu
1	PK	STATUS_ID	int
2		SMOKEDAY	nvarchar(255)
3		STOPSMK	nvarchar(255)
4		SMOKE100	nvarchar(255)
5		HOW_LONG_LAST_SMOKE	nvarchar(255)
6		USENOW	nvarchar(255)

Hình 38: OLAP DIM_SMOKE_STATUS

- Chiều về địa điểm khu vực khảo sát
  Gồm một khoá chính REGION_ID và thuộc tính STATE.

STT	Thuộc tính	Tên thuộc tính	Kiểu
1	PK	REGION_ID	int
2		STATE	nvarchar(255)

Hình 39: OLAP DIM_SMOKE_STATUS

Sau đây là phần thống kê dữ liệu theo các chiều:

- Chiều về thời gian

3
YEAR
2013
2014
2015

12
MONTH
1
2
3
4
5
6
7
8
9
10
11
12

4
QUARTER
1
2
3
4

- Chiều về tình trạng sức khoẻ người tham gia khảo sát

7
HEALTH STATUS
Excellent
Very good
Good
Fair
Poor
Not sure
Refused

2
HLTHPL
Yes
No

2
PERSDOC
Yes
No

2
MEDCOS
Yes
No

- Chiều về khu vực

6
STATE
Alabama
California
Massachusetts
New York
Texas
Washington

- Chiều về tình trạng hút thuốc của người tham gia khảo sát

8
HOW_LONG_LAST
Within the past month (less than 1 month ago)
Within the past 3 month (1 month but less than 3 month ago)
Within the past 6 month (3 months but less than 6 month ago)
Within the past year (6 months but less than 1 year ago)
Within the past 5 year (1 year but less than 5 years ago)
Within the past 10 year (5 years but less than 10 years ago)
10 years or more
Never smoked regularly

3
USENOW
Not at all
Every day
Some Days

3
SMOKDAY
Not at all
Every day
Some Days

2
SMOKE100
Yes
No

2
STOPSMK2
Yes
No

- Chiều về thông tin người tham gia khảo sát

2
SEX
Yes
No

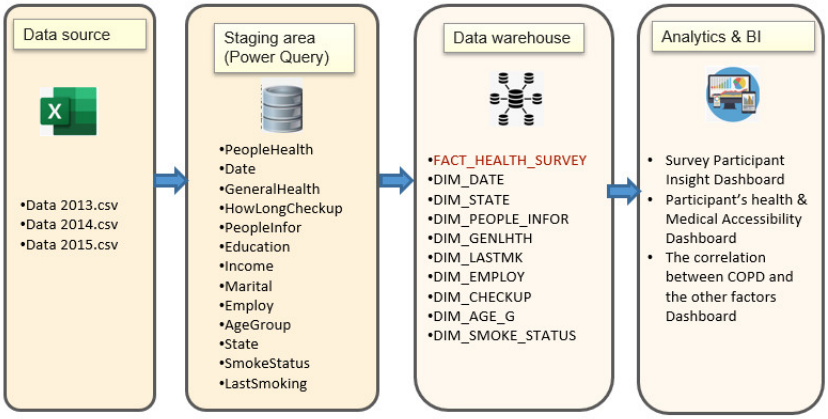
7
MARITAL_STATUS
Married
Divorced
Windowed
Separated
Never married
A member of an unmarried couple
Refused

7
GRADE
Never attended school or only kindergarten
Grades 1 through 8 (Elementary)
Grades 9 through 11 (Some high school)
Grades 12 or GED (High school graduate)
College 1 year to 3 years (Some college or technical school)
College 4 year or more (College graduate)
Refused

6
AGE
Age 18 to 24
Age 25 to 34
Age 35 to 44
Age 45 to 54
Age 55 to 64
Age 65 or older

9
EMPLOY_STATUS
Employed for wages
Self-employed
Out of work for 1 year or more
Out of work for less than 1 year
A homemaker
A student
Retired
Unable to work
Refused

8
INCOME_VALUE
Less than $10,000
$10,000 to less than $15,000
$15,000 to less than $20,000
$20,000 to less than $25,000
$25,000 to less than $35,000
$35,000 to less than $50,000
$50,000 to less than $75,000
$75,000 or more



Hình 40: Kiến trúc hệ thống OLAP Healthcare

3.2.6 Data model-ERD

Tầng dưới cùng là hệ CSDL quan hệ "DataCo_OLTP". Các công cụ đầu cuối, các tiện ích được dùng để đưa dữ liệu vào tầng dưới cùng từ hệ cơ sở dữ liệu hoạt động hoặc từ nguồn bên ngoài (file csv, excel...). Những công cụ và tiện ích này thực hiện việc loạibỏ dữ liệu thừa, làm sạch dữ liệu, chuyển đổi dữ liệu, cập nhật dữ liệu. Dữ liệu được lưu trữ vào 13 bảng:

- PEOPLEHEALTH
- DATE
- GENARALHEALTH
- HOWLONGCHECKUP
- PEOPLEINFOR
- EDUCATION
- INCOME
- MARITAL
- EMPLOY
- AGEGROUP
- STATE
- SMOKESTATUS
- LASTSMOKING

Tầng giữa là Data warehouse "DataCo_OLAP" được cài đặt dùng mô hình quan hệ OLAP. Dữ liệu được đổ từ mô hình OLTP của hệ CSDL hoạt động sang mô hình OLAP của Data warehouse, bao gồm các bảng:

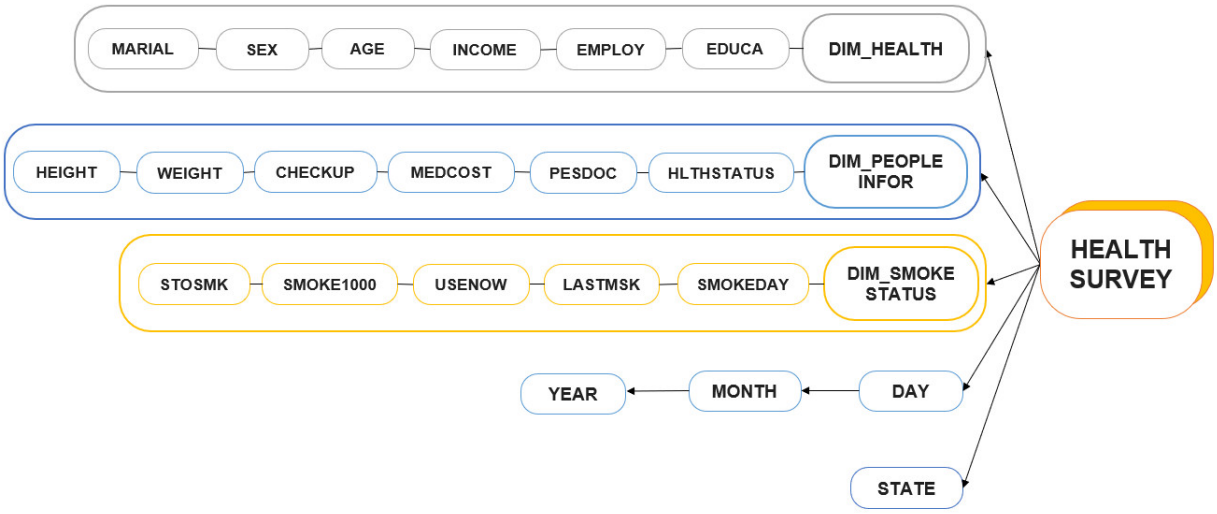
- FACT_HEALTH_SURVEY
- DIM_HEALTH
- DIM_DATE
- DIM_REGION
- DIM_PEOPLE_INFOR
- DIM_EDUCATION
- DIM_EMPLOY
- DIM_SMOKE_STATUS

Tầng trên cùng là tầng người dùng cuối, gồm các câu truy vấn và các công cụ làm báo cáo, phân tích, công cụ khai thác dữ liệu về:

- Survey Participant Insight Dashboard: thông tin chi tiết người tham gia khảo sát.
- Participant's health & Medical Accessibility Dashboard: tổng quan về khả năng tiếp cập y tế và sức khỏe người tham gia khảo sát.
- The correlation between COPD and the other factors Dashboard: mối tương quan giữa COPD và các yếu tố khác.

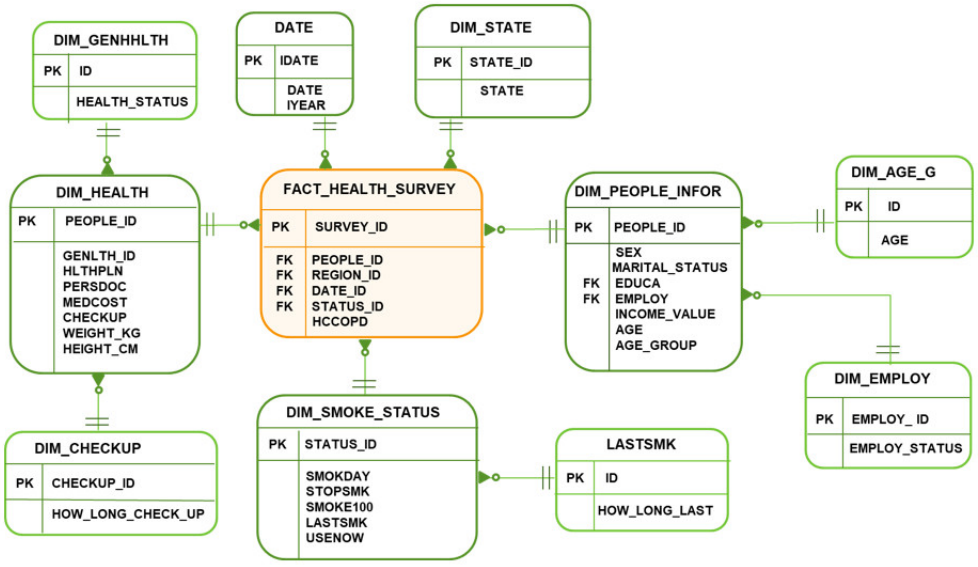
3.2.7 Data model OLAP

Mô hình logic



Hình 41: Mô hình logic của hệ thống OLAP

Mô hình quan hệ



Hình 42: Mô hình quan hệ của hệ thống OLAP

3.3 Xây dựng hệ thống

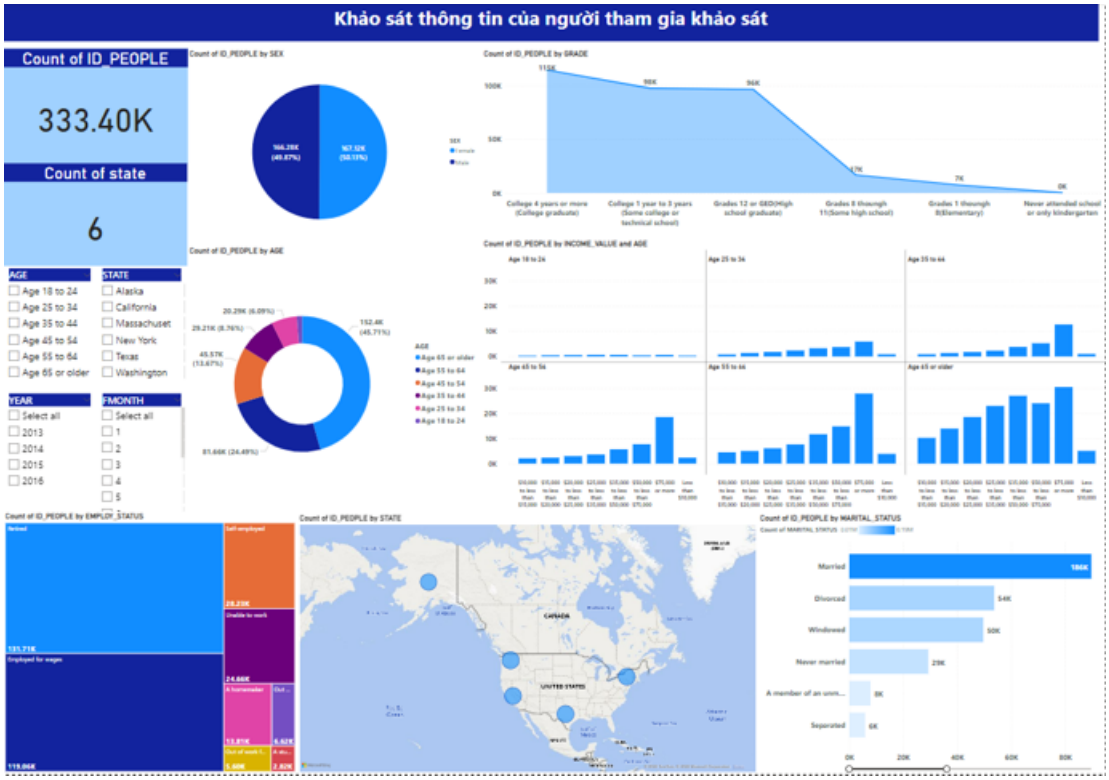
3.3.1 Xây dựng Dashboard

1. **Khảo sát thông tin của những người tham gia thực hiện khảo sát**
   Dashboard thể hiện tổng quan nhất về những người tham gia thực hiện khảo sát này với các yếu tố như độ tuổi, giới tính, trình độ học vấn, thu nhập,... thông qua

   – Cart thể hiện tổng số người tham gia và tất cả các khu vực.

   – Slicer theo các thuộc tính age, state, year, fmonth để có thể linh hoạt lọc ra các thông tin theo mong muốn.

   – Một số biểu đồ như Pie chart, Donut chart, map, treemap,...



Hình 43: Khảo sát thông tin của những người tham gia thực hiện khảo sát

Từ Dashboard này ta có được thông tin phân tích như sau

– Có tổng cộng tất cả là 333.4 nghìn người tham gia thực hiện khảo sát đến từ 6 bang được phân thành 6 nhóm tuổi chính từ 18 đến 24 tuổi, từ 25 đến 34 tuổi, từ 35 đến 44, từ 45 đến 54, từ 55 đến 64 tuổi và từ 65 tuổi trở lên.

– Về giới tính của người tham gia khảo sát thì tỷ lệ giữa nam và nữ là xấp xỉ nhau với nữ là 50,13% và nam là 49,87%.

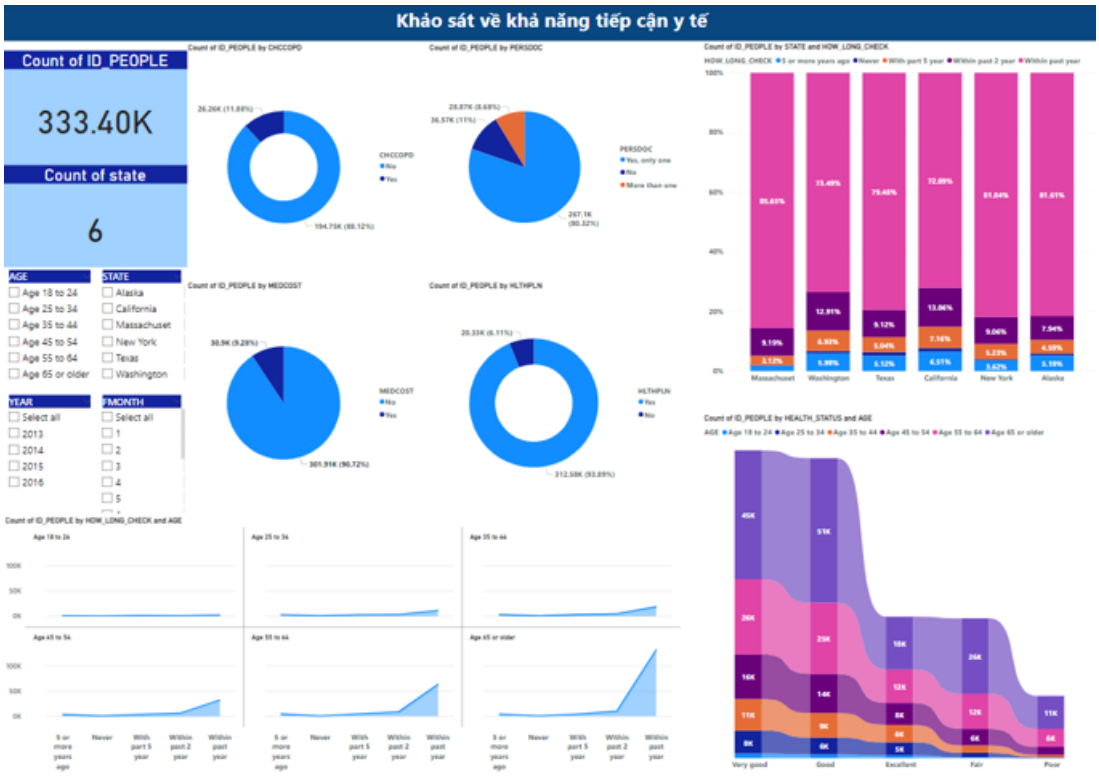
- Về độ tuổi thì ta thấy nhóm tuổi tham gia thực hiện nhiều nhất là từ 65 tuổi trở lên chiếm 45,71% trên tổng tất cả các nhóm tuổi, tiếp đến là nhóm từ 55 đến 64 tuổi chiếm 24,49% và tỷ lệ này có xu hướng giảm dần theo độ tuổi, thấp nhất là nhóm từ 18 đến 24 tuổi với 1,28%.
- Về trình độ học vấn thì số người đã tốt nghiệp cao đẳng,đại học đa số chiếm phần lớn và xếp sa lần lượt là đang học cao đẳng, đại học và tốt nghiệp lớp 12.
- Về thu nhập thì ở nhóm tuổi từ 18 đến 24 tuổi mức thu nhập được trải đều còn với các nhóm tuổi còn lại thì mức thu nhập hầu hết là hơn $ 75000, riêng với nhóm từ 65 tuổi trở lên thì tỷ lệ mức thu nhập từ dưới $ 75000 nhiều hơn so với các bảng còn lại.
- Về tình trạng việc làm vì chủ yếu người tham gia khảo sát là trên 65 tuổi nên đa số đã về hưu, theo sau đó là những người làm công ăn lương cũng chiếm khá đông.
- Về số lượng người tham gia thực hiện theo vị trí địa lý ta thấy tỷ lệ người tham gia từ các bang là tương đương nhau.
- Về tình trạng hôn nhân thì chủ yếu là những người đã kết hôn với hơn 186 nghìn người trên tổng số hơn 333 nghìn người chiếm hơn 50%.

Qua đó, ta thấy khảo sát này được thực hiện đa số ở những người trung tuổi nên hầu hết đã có việc làm hoặc đã về hưu và tuy họ ở các khu vực khác nhau nhưng đời sống đều ở mức ổn định.

2. Khảo sát về khả năng tiếp cận y tế

Khả năng tiếp cận y tế tùy thuộc vào hoàn cảnh của mỗi người nên việc xây dựng Dashboard này để thể hiện các vấn đề về việc khám sức khỏe với

- Cart thể hiện tổng số người tham gia và tất cả các khu vực.
- Slicer theo các thuộc tính age, state, year, fmonth để có thể linh hoạt lọc ra các thông tin theo mong muốn.
- Một số biểu đồ như Pie chart, Donut chart, Area chart, 100% Stacked column chart,...



Hình 44: Khảo sát về khả năng tiếp cận y tế

Từ Dashboard này ta có được thông tin phân tích như sau

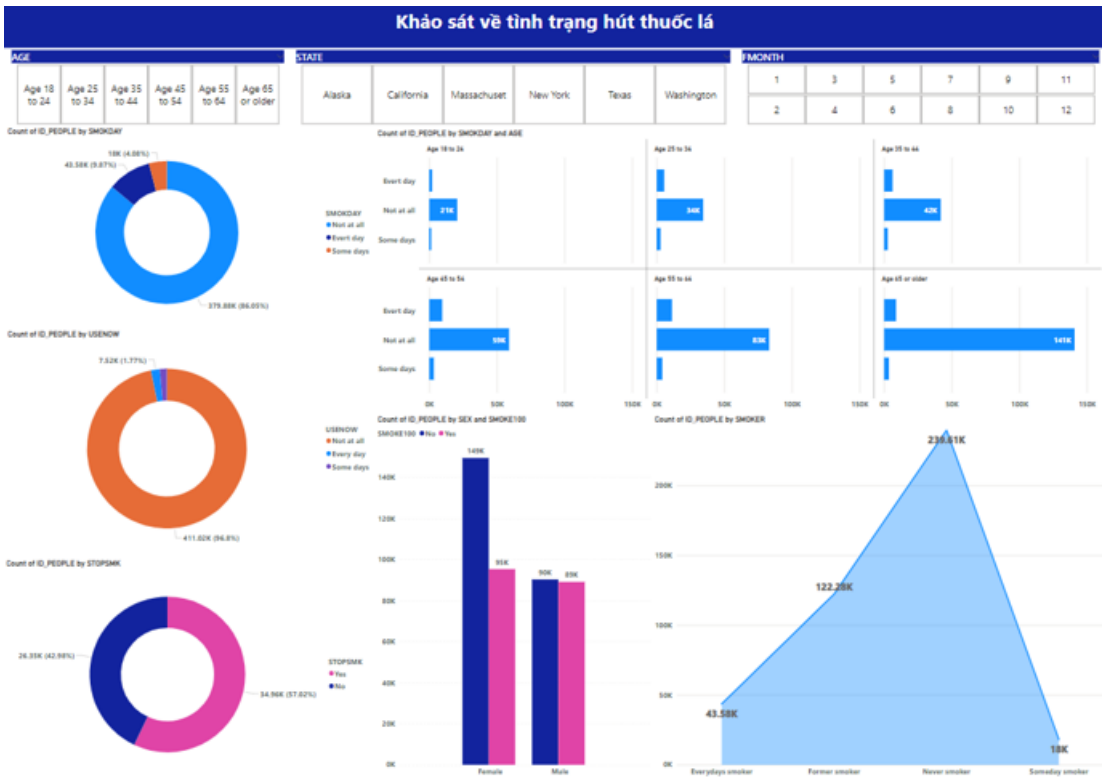
- Tỷ lệ người từng mắc bệnh tắc nghẽn phổi mãn tính là 11,88%.
- Số người có bác sĩ riêng chiếm khá ca trong đó có đến 80,32% là có một bác sĩ riêng và 8,68% là nhiều hơn một bác sĩ.
- Trong đó cũng có những người chưa có đủ điều kiện đi thăm khám bác sĩ cần thiết trong vòng 12 tháng qua chiếm 9,28%.
- Số người tham gia bảo hiểm y tế rất cao chiếm gần 94%.
- Về lần cuối đi khám sức khỏe, ta thấy mọi người thường xuyên đi khám sức khỏe và nhóm từ 65 tuổi trở lên có tỷ lệ đi khám sức khỏe lớn nhất do càng lớn tuổi thì sẽ càng quan tâm chăm sóc sức khỏe nhiều hơn, ngược lại nhóm từ 18 đến 24 tuổi chiếm tỷ lệ thấp nhất.
- Xét về từng khu vực thì thấy Massachuset có tỷ lệ thăm khám đều đặn hơn so với các khu vực còn lại.
- Về tình trạng sức khỏe theo độ tuổi thì từ 18 đến 24 tuổi trạng thái tốt hơn hẳn, từ 25 đến 64 tuổi có sức khỏe tốt chiếm phần lớn và trên 65 tuổi trở lên số người có tình trạng không ổn chiếm nhiều hơn so với các nhóm tuổi khác, đó cũng là điều dễ hiểu.

Qua đó, ta thấy về chất lượng cuộc của những người này khá ổn với đa số người có bác sĩ riêng và có tham gia bảo hiểm y tế, về việc đi khám sức khỏe cũng được diễn ra thường xuyên.

3. Khảo sát về tình trạng hút thuốc lá

Dashboard thể hiện tình trạng hút thuốc lá theo các yếu tố như độ tuổi, tần suất sử dụng,...

- Slicer theo các thuộc tính age, state, fmonth để có thể linh hoạt lọc ra các thông tin theo mong muốn.
- Một số biểu đồ như Pie chart, Donut chart, Area chart, Clustered column chart,...



Hình 45: Khảo sát về tình trạng hút thuốc lá

Từ Dashboard này ta có được thông tin phân tích như sau

- Số lượng người hút thuốc mỗi ngày chiếm 9,87% còn lại là số lượng người chưa hút thuốc hoặc hút ít.
- Thói quen hút thuốc theo nhóm tuổi: số lượng người hút thuốc mỗi ngày ở nhóm từ 45 đến 54 tuổi và từ 55 đến 64 tuổi là đông hơn so với các nhóm còn lại.
- Số lượng người cố gắng bỏ thuốc chiếm 57,02% so với những người chưa có ý định đó.
- Ở nữ thì việc hút ít hơn 100 điếu có số lượng nhiều hơn hẳn so với việc hút nhiều hơn 100 điếu. Còn ở nam thì việc hút ít hơn hay nhiều hơn 100 điếu là ngang nhau và ngang với số người nữ hút nhiều hơn 100 điếu.

– Cuối cùng là số lượng người hút thuốc theo tần suất thì ở đây ngoài việc không sử dụng là chủ yếu thì tiếp đến là người đã từng sử dụng chiếm khoảng $\frac{1}{3}$ tổng người tham gia, sau đó mới đến người hút hàng ngày và hút ít.

Qua đó, ta thấy trong những người tham gia thực hiện khảo sát này thì số người sử dụng thuốc lá là không nhiều.

3.3.2 Bài học tổng kết

Sau khi cùng nhau tìm hiểu và hoàn thành bài tập nhóm này, nhóm chúng em đã rút ra được một số bài học như sau

1. Cần tìm hiểu kĩ về chuyên môn, nghiệp vụ và các quá trình liên quan trước khi phân tích thiết kế và xây dựng Datawarehouse
2. Kích thước dữ liệu lớn cần phải import và ETL sao cho hợp lý
3. Dữ liệu trong thực tế không bao giờ hoàn hảo, luôn có các trường cần xử lý
4. Data model phải có quan hệ giữa các bảng, thông tin như khoá chính, kiểu dữ liệu,...
5. Dashboard phải bám sát nghiệp vụ, đảm bảo tính đa dạng, trực quan, có các thành phần như: footer, header, slicer,...
6. Dashboard phải mang tính phân tích, so sánh, không quá tập trung vào thống kê, miêu tả.

Kết luận

Kết luận

Báo cáo đã đạt được mục tiêu đề ra

Báo cáo đã nghiên cứu tìm hiểu tổng quan về Data Warehouse với những kiến thức nền tảng về kho dữ liệu, kinh doanh thông minh, phân tích dữ liệu và phân tích kinh doanh

Kết quả của báo cáo

1. Trình bày được những kiến thức nền tảng cơ bản của Kho dữ liệu và Kinh doanh thông minh.
2. Xây dựng được kiến trúc của kho dữ liệu của một bộ dữ liệu thực tế.
3. Phân tích và thiết kế được các mô hình của hệ thống mới.
4. Làm chủ được công cụ trực quan hóa dữ liệu Power Bi, Google Data Studio, công cụ Exel, Power Query trong phân tích và xử lý dữ liệu.

Kỹ năng đạt được

1. Bước đầu biết tìm kiếm, đọc, dịch tài liệu chuyên ngành liên quan đến nội dung báo cáo.
2. Biết tổng hợp các kiến thức đã học và kiến thức trong tài liệu tham khảo để viết báo cáo.
3. Biết ứng dụng các kiến thức có được vào một vấn đề thực tế.

Tài liệu

[1] Bài giảng kho dữ liệu và kinh doanh thông minh, Viện Toán ứng dụng và Tin học, Nguyễn Danh Tú, 2020

[2] Kênh Youtube Học Excel cơ bản

[3] Data Analysis Methods and Techniques, The Datapine Blog. Business intelligence (BI), Search Business Analytics.

[4] Business Intelligence: What It Is, How It Works, Its Importance, Examples, & Tools, Tableau.

[5] What is a Business Analyst? , International Institute Of Business Analysis.