TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC

----------ଌ----------

Báo cáo bài tập lớn môn Hệ Hỗ Trợ Quyết Định

Chủ đề: Xây dựng Data WareHouse và phân tích BI về vấn đề Health Care

Giảng viên hướng dẫn: TS. Vũ Thành Nam

Nhóm thực hiện: Nhóm 4

Thành viên nhóm:

Hoàng Phương Cúc	20185332
Nguyễn Ngọc Thìn	20185408
Nguyễn Thị Quý	20185396

01/05/2021

Phân công nhiệm vụ

Sinh viên	MSSV	Công việc
Hoàng Phương Cúc	20185332	+ Tổng hợp báo cáo + Lí thuyết Dashboard + Xử lí ETL + Xây dựng OLAP
Nguyễn Ngọc Thìn	20185408	+ Lí thuyết Datawarehouse + Xử lí ETL + Xây dựng OLTP + Slide thuyết trình
Nguyễn Thị Quý	20185396	+ Lí thuyết phần BI + Xác định bài toán + Xử lí ETL + Báo cáo Dashboard

MỤC LỤC

LỜI MỞ ĐẦU

Đi theo dòng của 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. Chính vì thế đối với các doanh nghiệp lớn, tập đoàn trong và ngoài nước, việc thiết kế và xây dựng một hệ thống quản lý khối lượng lớn dữ liệu để truy xuất dễ dàng đồng thời phải đảm bảo tính khách quan đúng đắn của dữ liệu là vô cùng cần thiết. Chắc hẳn đối với bộ phận phân tích và phát triển kinh doanh trong mỗi doanh nghiệp, khái niệm kho dữ liệu (data warehouse), dữ liệu lớn (big data) tại các công ty công nghệ, cơ sở dữ liệu (database) ở các công ty lập trình… không còn quá xa lạ nữa. 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.

PHẦN I: CƠ SỞ LÝ THUYẾT

1. Tổng quan về Data WareHouse

1.1. Khái niệm

- Data WareHouse (Kho dữ liệu) là tập hợp của các cơ sở dữ liệu tích hợp, hướng chủ đề, được thiết kế để hỗ trợ cho chức năng trợ giúp quyết định mà mỗi đơn vị dữ liệu đều liên quan tới một khoảng thời gian cụ thể.
- Những dữ liệu trong kho dữ liệu đã được tổ chức dưới dạng sẵn sàng cho quá trình phân tích và chúng đều mang những tiềm năng (hiện tại và lịch sử) mà các nhà lãnh đạo doanh nghiệp quan tâm.
- Dữ liệu đầu vào có thể là tệp .cvs, bảng tính, cơ sở dữ liệu quan hệ hoặc không quan hệ, ngoài ra có thể lấy dữ liệu bằng API.
- Lợi ích mà kho dữ liệu mang lại:
  + Tạo ra những quyết định có ảnh hưởng lớn.
  + Công việc kinh doanh trở nên thông minh hơn.
  + Dịch vụ khách hàng được nâng cao.
  + Tái sáng tạo những tiến trình kinh doanh.
  + Tái sáng tạo hệ thống thông tin.

1.2. Đặc điểm của Data Warehouse

Kho dữ liệu là một tập hợp dữ liệu có những tính chất sau:

- Hướng chủ đề:
  + Được tổ chức quanh các chủ đề: customer, product, sales, …
  + Tập trung vào việc mô hình và phân tích dữ liệu cho việc ra quyết định chứ không xử lý các giao dịch hay tác nghiệp hàng ngày.
  + Cung cấp một góc nhìn đơn giản và xúc tích quanh một chủ đề cụ thể bằng cách làm sạch dữ liệu, loại bỏ dữ liệu không hữu dụng trong tiến trình hỗ trợ quyết định.
- Tính tích hợp: Cách dữ liệu được trích xuất và chuyển đổi là thống nhất, bất kể nguồn gốc của dữ liệu đó từ đâu mà có.
- Dữ liệu gắn với thời gian và có tính lịch sử: Dữ liệu được sắp xếp theo các khoảng thời gian (hàng tuần, hàng tháng, hàng năm, …).
- Dữ liệu chỉ đọc
- Dữ liệu không biến động: Kho dữ liệu không được cập nhật theo thời gian thực. Nó được cập nhật định kỳ thông qua việc tải dữ liệu lên, bảo vệ dữ liệu khỏi ảnh hưởng của sự thay đổi nhất thời.
- Dữ liệu được tổng hợp và chi tiết: Sử dụng kho dữ liệu giúp việc tạo báo cáo, chạy các truy vấn đột xuất và trích xuất các luồng dữ liệu trở nên đơn giản hơn.

Từ những đặc điểm trên ta có những thuật ngữ liên quan tới Data WareHouse:

- DataMart (DM): Kho dữ liệu chủ đề

- ODS: Cung cấp dữ liệu hiện tại, hỗ trợ cho những quyết định ngắn hạn
- EDW: Kho dữ liệu quy mô lớn cho toàn doanh nghiệp
- MetaData: Siêu dữ liệu

Lợi ích mà Data WareHouse là đưa thông tin và những đánh giá tổng quan, kỹ lưỡng đến những nhà lãnh đạo doanh nghiệp một cách kịp thời. Điều này giúp các doanh nghiệp có thể bắt kịp với tốc độ tăng trưởng và thay đổi của thị trường khi nền cách mạng công nghiệp đang dần chuyển đổi sang kỹ thuật số.

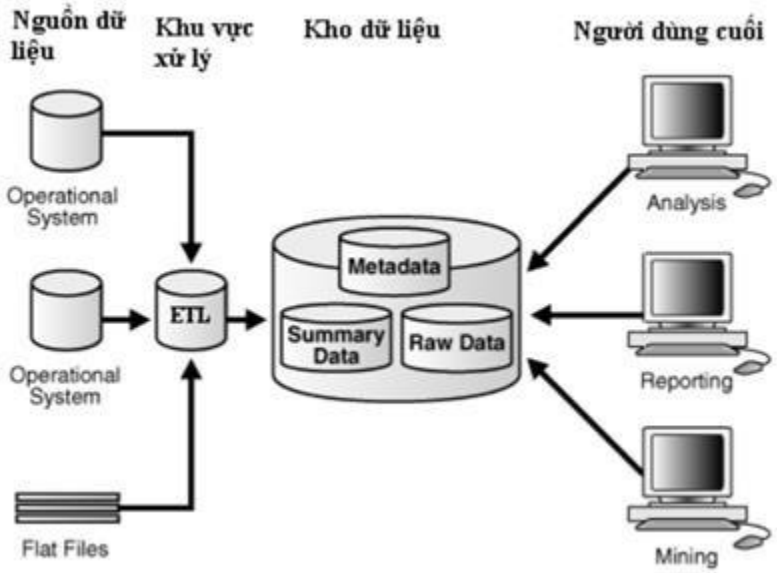
1.3. Cấu trúc lưu giữ DW

Mô hình kiến trúc của kho dữ liệu cơ bản gồm 3 thành phần:

+ Dữ liệu nguồn

+ Khu vực xử lý

+ Kho dữ liệu



1.3.1. Nguồn dữ liệu

- Dữ liệu có thể đến từ nhiều nguồn khác nhau:
  + Các hệ thống tác nghiệp
  + Hệ thống kế thừa
  + Các nguồn dữ liệu bên ngoài

1.3.2. Các công cụ truy vấn, tạo báo cáo, phân tích dữ liệu

- Công cụ tạo báo cáo và câu hỏi truy vấn (Report) -
  Công cụ phân tích trực tuyến (OLAP):

+ Dữ liệu phát sinh từ các hoạt động hằng ngày được thu thập, xử lý để phục vụ công việc cụ thể của một tổ chức thường được gọi là dữ liệu tác nghiệp và hoạt động thu thập xử lý loại dữ liệu này được gọi là xử lý giao dịch trực tuyến (OLTP).

+ Dữ liệu tại các CSDL tác nghiệp được lấy từ nhiều nguồn khác nhau nên dễ bị nhiễu, hỗn tạp dẫn đến dữ liệu không được sạch, không toàn vẹn và dễ bị trùng nhau. Do đó việc kiểm tra dữ liệu, làm sạch dữ liệu phải được tiến hành ngay tại đây nhằm đảm bảo tính toàn vẹn, tính đúng đắn, tính khách quan và tính nhất quán dữ liệu trước khi ta đưa dữ liệu đó vào kho đích. Quá trình thường được diễn ra theo thứ tự Extract – Transform – Load (ETL), ngoài ra có thể xử lí dữ liệu theo thứ tự Extract – Load – Transform (ELT).

- Công cụ phân tích, tìm kiếm nâng cao (Data Mining).

1.3.3. Kho dữ liệu

- Cơ sở dữ liệu của kho dữ liệu
- Siêu dữ liệu (Metadata)
- Kho dữ liệu chủ đề (DataMart)
- Bảng sự kiện tổng hợp (Fact)

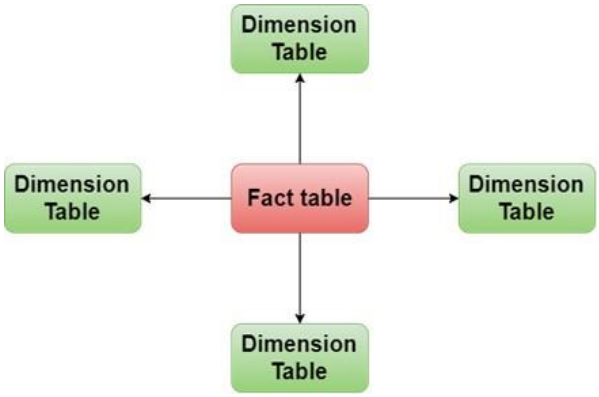
1.4. Mô hình dữ liệu DW

1.4.1. Mô hình vật lí

Mô hình quan hệ

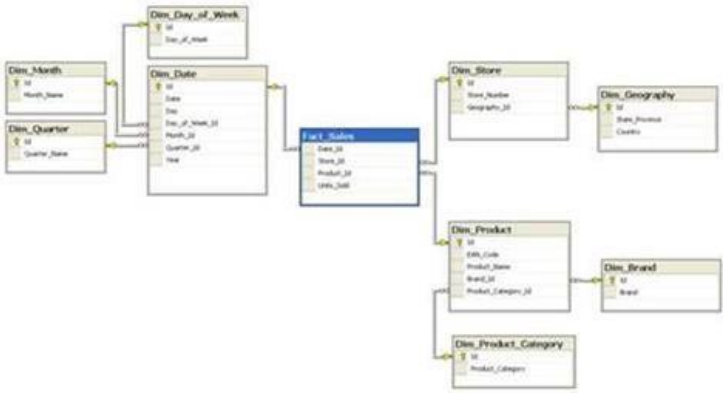
- Mục đích: chuyển từ mô hình nhiều chiều sang mô hình quan hệ để có thể giảm mất mát thông tin ngữ nghĩa, dễ dàng bảo trì và chuyển đổi hiệu quả từ các truy vấn đa chiều (điểm cốt lõi của OLAP).

• Lược đồ hình sao (Star Schema)



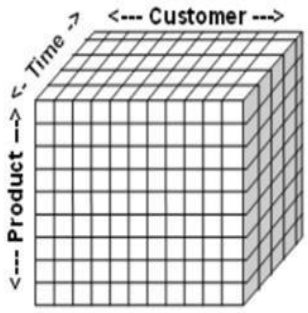
Ý tưởng: Sử dụng lược đồ phi chuẩn cho tất cả các chiều.

• Lược đồ hình bông tuyết (Snowflake – schema)



Ý tưởng: Sử dụng 1 bảng cho một mức phân lớp

- Lược đồ kết hợp



Là sự kết hợp giữa lược đồ hình sao và lược đồ bông tuyết rơi.

Các mô hình vật lí

- MOLAP
- ROLAP
- HOLAP
- DOLAP

1.4.2. Mô hình dữ liệu đa chiều

Bản chất đa chiều của các câu hỏi trong nghiệp vụ được phản ánh trong thực tế chẳng hạn như: những người quản lí thị trường không được thỏa mãn với câu hỏi theo một chiều đơn giản, thay vào đó là những câu hỏi phức tạp. Một cách để quan sát một mô hình dữ liệu nhiều chiều là nhìn nó như một hình khối.

Các thành phần chính bao gồm:

+ Các dữ kiện (Facts)

+ Các chiều (Dimensions)

+ Các khối đa chiều (Cubes)

✦ Bảng sự kiện

Bảng sự kiện điển hình có hai kiểu cột, chúng chứa đựng những sự kiện số (thường gọi là thước đo), và chứa khóa của các bảng dimension. Bảng sự kiện chứa đựng những sự kiện mức chi tiết hoặc những sự kiện đã được tổng hợp. Bảng sự kiện chứa sự kiện tổng hợp thường được gọi là những bảng tóm tắt. Bảng sự kiện thông thường chứa đựng những sự kiện với cùng mức của sự tổng hợp. Tuy nhiên hầu hết những sự kiện liên kết tất cả các chiều, nó có thể liên kết với một số chiều hoặc không liên kết.

✦ Bảng chiều

Các chiều là cách mô tả chủng loại mà theo đó các dữ liệu số trong khối được phân chia để phân tích. Khi xác định một chiều, chọn một hoặc nhiều cột của một trong các bảng liên kết (bảng chiều). Nếu ta chọn các cột phức tạp thì tất cả cần có quan hệ với nhau, chẳng hạn các giá trị của chúng có thể được tổ chức theo hệ thống phân cấp đơn. Để xác định hệ thống phân cấp, sắp xếp các cột từ chung nhất tới cụ thể nhất. Ví dụ: một chiều thời gian được tạo ra từ các cột năm, quý, tháng, ngày.

Mỗi cột trong chiều góp phần vào một cấp độ cho chiều. Các cấp độ được sắp đặt theo nét riêng biệt và được tổ chức trong hệ thống cấp bậc mà nó thừa nhận các con đường hợp logic cho việc đào sâu.

1.5. Tổ chức dữ liệu vật lý

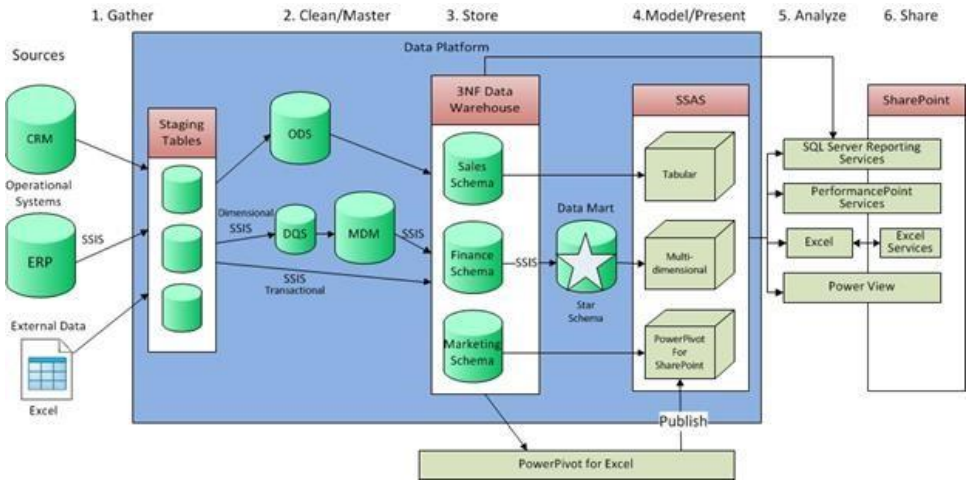
1.5.1. Phân vùng

Phân vùng (partition) là kĩ thuật được sử dụng trong kho dữ liệu nhằm tối ưu hiệu suất truy vấn bằng cách cho phép người thiết kế phân vùng các vùng nhớ để chứa dữ liệu thoả mãn những yêu cầu do người thiết kế đặt ra.

1.5.2. Chỉ mục

Đánh chỉ mục (Index) là kỹ thuật phổ biến nhằm tăng hiệu suất các truy vấn dữ liệu. Chuyên gia thiết kế sẽ chọn trường phù hợp của một bảng để đánh chỉ số, khi đó trường chỉ số đó sẽ được lưu ra một bảng tham chiếu, được sắp xếp sẵn. Khi có truy vấn dữ liệu, thời gian truy vấn sẽ giảm do dữ liệu cần truy vấn đã được sắp xếp từ trước.

1.6. Quy trình xử lý luồng dữ liệu



Quy trình xử lý luồng dữ liệu ✦

Có 6 bước trong xử lý luồng dữ liệu trong Data Warehouse:

Bước 1: Nhập dữ liệu hoặc lấy dữ liệu từ nguồn.

Bước 2: Chuẩn bị dữ liệu, loại bỏ dữ liệu xấu và chuyển đổi dữ liệu.

Bước 3: Lưu dữ liệu vào bộ nhớ.

Bước 4: Trình bày thông tin chi tiết về dữ liệu đó vào biểu diễn trực quan.

Bước 5: Phân tích dữ liệu được lưu trữ để rút ra thông tin chi tiết.

Bước 6: Cung cấp chia sẻ thông tin, kết quả cuối cùng thu được sau quá trình phân tích dữ liệu.

1.7. Kho dữ liệu hiện nay và xu hướng trong tương lai

1.7.1. Kho dữ liệu hiện nay

Quản trị doanh nghiệp thông minh.

Quản lý mối quan hệ khách hàng.

Khai phá dữ liệu.

Quản lý dữ liệu chủ.

Tích hợp dữ liệu khách hàng.

1.7.2. Xu hướng trong tương lai của kho dữ liệu

Dữ liệu phi cấu trúc.

Tìm kiếm.

Kiến trúc hướng dịch vụ.

Kho dữ liệu thời gian thực

2. Tổng quan về Business Intelligence (BI)

2.1. Khái niệm BI

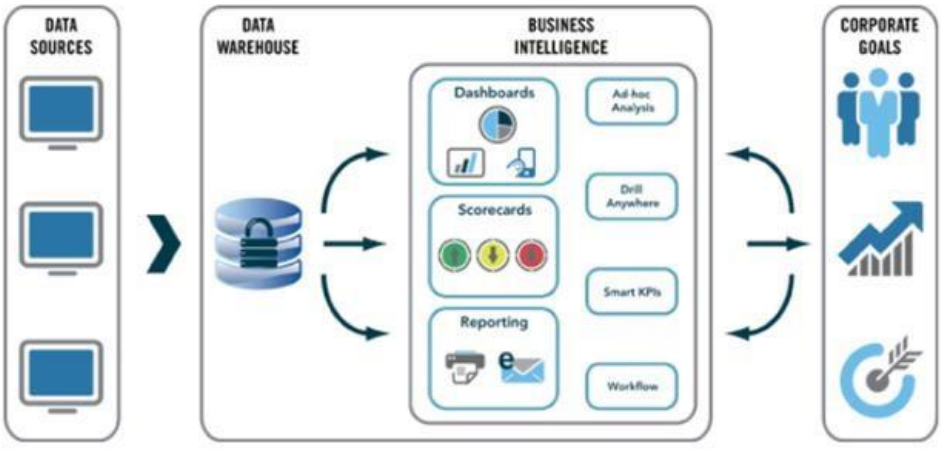
Theo Solomon Negash và Paul Gray, trí thông minh kinh doanh (BI) có thể được định nghĩa là các hệ thống kết hợp:

- Thu thập dữ liệu
- Lưu trữ dữ liệu • Quản lý kiến thức

với phân tích để đánh giá thông tin phức tạp của công ty và cạnh tranh để trình bày cho các nhà hoạch định và ra quyết định, với mục tiêu cải thiện tính kịp thời và chất lượng của đầu vào cho quá trình ra quyết định.

Theo Forrester Research , kinh doanh thông minh là "một tập hợp các phương pháp luận, quy trình, kiến trúc và công nghệ giúp chuyển đổi dữ liệu thô thành thông tin hữu ích và có ý nghĩa được sử dụng để cho phép hiểu rõ hơn về chiến lược, chiến thuật và hoạt động và ra quyết định.

Business Intelligence (gọi tắt là 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 kinh doanh phù hợp.



2.2. Lịch sử hình thành và phát triển

Việc sử dụng thuật ngữ kinh doanh thông minh sớm nhất được biết đến là trong cuốn Cyclopaedia of Commercial and Business Anecdotes (1865) của Richard Millar Devens . Devens đã sử dụng thuật ngữ này để mô tả cách chủ ngân hàng Sir Henry Furnese thu được lợi nhuận bằng cách tiếp nhận và hành động dựa trên thông tin về môi trường của anh ta, trước các đối thủ cạnh tranh của anh ta.

Kinh doanh thông minh như cách hiểu ngày nay được cho là đã phát triển từ hệ thống hỗ trợ quyết định (DSS) bắt đầu từ những năm 1960 và phát triển trong suốt giữa những năm 1980. DSS bắt nguồn từ các mô hình có sự hỗ trợ của máy tính được tạo ra để hỗ trợ việc ra quyết định và lập kế hoạch *Quá trình phát triển:*

+ 1970: Hệ thống thông tin quản lý (MIS): chức năng báo cáo, chủ yếu là static/periodic report
+ 1980: Hệ thống thông tin điều hành (Executive Information Systems): hỗ trợ người dùng lãnh đạo và quản lý cấp cao, có thêm các chức năng báo cáo động, phân tích (phân tích xu hướng, dự báo)
+ 1990: Kinh doanh thông minh (BI): với các công cụ, công nghệ như OLAP, CSDL nhiều chiều - multidimensional
+ 2005: Bổ sung thêm dashboard, portal, AI, Data/text minning

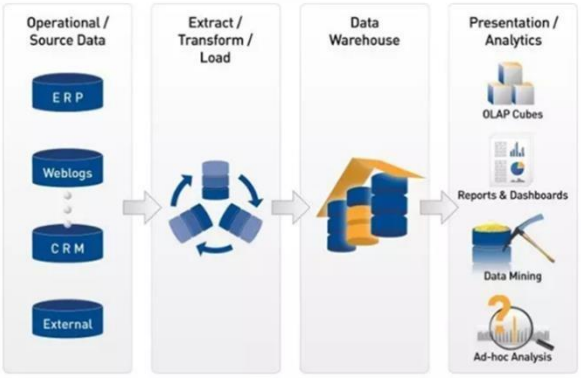
2.3. Lợi ích của BI

Kinh doanh thông minh có thể được áp dụng cho các mục đích kinh doanh sau:

+ Các thước đo hiệu suất và điểm chuẩn thông báo cho các nhà lãnh đạo doanh nghiệp về tiến độ hướng tới các mục tiêu kinh doanh (quản lý quy trình kinh doanh).
+ Analytics định lượng các quy trình để một doanh tăng tốc và cải thiện việc ra quyết định tối ưu và thực hiện khám phá kiến thức kinh doanh. Analytics có thể khác nhau như liên quan đến khai thác dữ liệu, khai thác quá trình, phân tích thống kê, phân tích dự báo, xây dựng mô hình dự báo, xây dựng mô hình quy trình kinh doanh, dòng dữ liệu, xử lý sự kiện phức tạp, và phân tích quy tắc.
+ Báo cáo kinh doanh có thể sử dụng dữ liệu BI để thông báo chiến lược. Báo cáo kinh doanh có thể liên quan đến trang tổng quan, trực quan hóa dữ liệu, hệ thống thông tin điều hành và / hoặc OLAP.
+ BI có thể tạo điều kiện hợp tác cả bên trong và bên ngoài doanh nghiệp bằng cách cho phép chia sẻ dữ liệu và trao đổi dữ liệu điện tử.

+ Quản lý tri thức liên quan đến việc tạo ra, phân phối, sử dụng và quản lý thông tin kinh doanh và tri thức kinh doanh nói chung. Quản lý tri thức dẫn đến quản lý học tập và tuân thủ quy định.

2.4. Các thành phần chính của BI



Data Sources

+ Là cơ sở dữ liệu thô (thường là cơ sở dữ liệu quan hệ) đến từ nhiều nguồn khác nhau như các ứng dụng business như Human Resource Management (HRM), Customer relationship management (CRM), phần mềm bán hàng, website thương mại điện tử…
+ Có thể là bất kỳ một hệ quản trị cơ sở dữ liệu như MySQL, Oracle, MSSQL, DB2, …
+ Thường được thiết kế theo mô hình cơ sở dữ liệu quan hệ nhưng cũng có thể là dữ liệu lớn, dữ liệu phi quan hệ (như mạng xã hội, NoSQL)

Data Warehouse

+ Là cơ sở dữ liệu được thiết kế theo mô hình khác với CSDL OLTP thông thường (Online Transaction Processings – OLTP là thiết kế CSDL dành cho việc đọc ghi thường xuyên, lượng dữ liệu cho mỗi lần đọc ghi ít) và là nơi lưu trữ dữ liệu lâu dài của tổ chức.
+ Dữ liệu của DWH chỉ có thể đọc, không được sử dụng để ghi hay update bởi ứng dụng thông thường, nó chỉ được cập nhật/ghi bởi công cụ ETL (Extract Transform Load), công cụ chuyển đổi dữ liệu từ Data Sources vào Data Warehouse.

Integrating Server

Chịu trách nhiệm trung gian vận hành công cụ ETL để chuyển đổi dữ liệu từ Data Sources vào Data Warehouse.

Analysis Server

+ Chịu trách nhiệm thực thi các cube được thiết kế dựa trên các chiều dữ liệu và tri thức nghiệp vụ

+ Cube chịu trách nhiệm nhận dữ liệu đầu vào từ DWH và thực thi theo nghiệp vụ định nghĩa sẵn để trả về kết quả. **Reporting Server**

+ Thực thi các report với output nhận được từ Analysis Server.
+ Nơi quản trị tập trung các report trên nền web, các report này có thể được attach vào ứng dụng web, hay application.

Data Mining

+ Là quá trình trích xuất thông tin dữ liệu đã qua xử lý (phù hợp với yêu cầu riêng của doanh nghiệp) từ Data Warehouse rồi kết hợp với các thuật toán để đưa ra (hoặc dự đoán) các quyết định có lợi cho việc kinh doanh của doanh nghiệp.
+ Đây là một quá trình quan trọng trong BI, thông thường một doanh nghiệp muốn sử dụng giải pháp BI thường kèm theo về Data Mining.

Data Presentation

Tạo ra các báo cáo, biểu đồ từ quá trình data mining để phục vụ cho nhu cầu của người dùng cuối.

2.5. Các công cụ của BI

Một số yếu tố của kinh doanh thông minh:

+ Tổng hợp và phân bổ đa chiều
+ Chuẩn hóa , gắn thẻ và chuẩn hóa
+ Báo cáo thời gian thực với cảnh báo phân tích
+ Một phương pháp giao tiếp với các nguồn dữ liệu phi cấu trúc
+ Hợp nhất nhóm, lập ngân sách và dự báo luân phiên
+ Suy luận thống kê và mô phỏng xác suất
+ Tối ưu hóa các chỉ số hiệu suất chính
+ Kiểm soát phiên bản và quản lý quy trình +
Mở quản lý mặt hàng

BI kết hợp một bộ lớn các ứng dụng phân tích bao gồm cả phân tích và truy vấn đặc thù (ad hoc), báo cáo doanh nghiệp, xử lý phân tích trực tuyến (OLAP) và Location Intelligence (LI).

Công nghệ BI cũng bao gồm phần mềm trực quan hóa dữ liệu phục vụ việc thiết kế các sơ đồ và các đồ họa thông tin, cũng như các công cụ sử dụng cho việc xây dựng các bảng điều khiển (dashboard) và các thẻ điểm hiệu suất hiển thị các dữ liệu được trực quan hóa trên các chiều kinh doanh và các KPI theo cách dễ dàng nắm bắt.

BI có thể cũng kết hợp các hình thức phân tích tiên tiến như khai thác dữ liệu, phân tích dự đoán, khai thác chữ (Text Mining), phân tích thống kê và phân tích dữ liệu lớn.

2.6. Xu hướng xây dựng BI

Bên cạnh các nhà quản lý BI, nhóm ứng dụng BI nhìn chung bao gồm các kiến trúc sư BI, các nhà phát triển, phân tích nghiệp vụ và các chuyên gia quản lý dữ liệu BI. Những người sử dụng nghiệp vụ cũng tham gia nhóm dự án, họ đại diện cho phía nghiệp vụ và có vai trò đảm bảo các yêu cầu nghiệp vụ cần thiết được đáp ứng trong quá trình phát triển BI.

Để hỗ trợ việc này, ngày càng nhiều tổ chức đang thay thế mô hình phát triển kiểu thác nước thành Agile BI và các cách tiếp cận data warehouse sử dụng kỹ thuật phát triển phần mềm Agile để chia nhỏ dự án BI thành các phần nhỏ và phát hành các chức năng cho phân tích nghiệp vụ trên cơ sở lặp và nâng cấp dần. Làm như vậy cho phép các doanh nghiệp có thể đưa các tính năng của BI vào thực tiễn nhanh hơn và làm mịn hoặc điều chỉnh các kế hoạch phát triển khi có các thay đổi nghiệp vụ cần hoặc xuất hiện các yêu cầu mới và có ưu tiên cao hơn các vấn đề cũ.

2.7. BI với dữ liệu lớn

Các nền tảng BI càng ngày càng được sử dụng như các giao diện đầu cuối cho các hệ thống dữ liệu lớn. Phần mềm BI hiện đại thường phục vụ các hệ thống phía sau (back end), cho phép chúng có thể kết nối đến một loạt các nguồn dữ liệu khác nhau. Cùng với giao diện người dùng đơn giản, cho phép các công cụ tích hợp tốt với các hệ thống dữ liệu lớn. Người dùng có thể kết nối đến một loạt nguồn dữ liệu, bao gồm các hệ thống Hadoop, các CSDL NoSQL, các nền tảng đám mây và nhiều các data warehouse thông thường khác, và có thể phát triển khung nhìn thống nhất cho các dữ liệu khác nhau.

BI thường được dùng để cung cấp giao diện cuối cùng đơn giản, trực quan nhằm cung cấp thông tin cho những người dùng sử dụng thông thường hơn là cách tiếp cận thường thấy của việc cung cấp cho các chuyên gia dữ liệu hay chuyên gia công nghệ.

2.8. Yếu tố chi phối sự thành công của một dự án BI

Để triển khai thành công một dự án BI, một trong những điều kiện tiên quyết là đội ngũ dự án phải am hiểu rõ nghiệp vụ và các sản phẩm đầu ra cho dự án. Đội ngũ phát triển phải có kiến thức, kinh nghiệm về phương thức thiết kế, tổ chức dữ liệu cho DWH. Việc làm chủ các nguồn dữ liệu và kiểm soát các công cụ ETL cũng đóng vai trò hết sức quan trọng. Bên cạnh đó, đội ngũ phát triển cũng cần am hiểu các công cụ của BI để có thể thiết kế và xây dựng hệ thống nhanh chóng, dễ dùng và hiệu quả.

2.9. Doanh nghiệp với BI

Doanh nghiệp có thể sử dụng trí tuệ kinh doanh để hỗ trợ một loạt các quyết định kinh doanh từ hoạt động đến chiến lược. Các quyết định vận hành cơ bản bao gồm định vị hoặc định giá sản phẩm . Các quyết định kinh doanh chiến lược liên quan đến các ưu tiên, mục tiêu và định hướng ở cấp độ rộng nhất. Trong mọi trường hợp, BI có hiệu quả nhất khi nó kết hợp dữ liệu có được từ thị trường mà công ty hoạt động (dữ liệu bên ngoài) với dữ liệu từ các nguồn nội bộ của công ty như dữ liệu tài chính và hoạt động (dữ liệu nội bộ). Khi kết hợp, dữ liệu bên ngoài và bên trong có thể cung cấp một bức tranh hoàn chỉnh, trên thực tế, tạo ra một "trí thông minh" mà không thể bắt nguồn từ bất kỳ tập hợp dữ liệu đơn lẻ nào.

Trong số vô số cách sử dụng, các công cụ thông minh kinh doanh cho phép các tổ chức hiểu rõ hơn về các thị trường mới, đánh giá nhu cầu và tính phù hợp của các sản phẩm và dịch vụ đối với các phân khúc thị trường khác nhau và đánh giá tác động của các nỗ lực tiếp thị.

3. Dashboard

3.1. Dashboard là gì ?

Dashboard là một bảng điều khiển kỹ thuật số (digital control), hay một giao diện số dùng để thu thập và tổng hợp dữ liệu của toàn bộ tổ chức. Nó không chỉ cung cấp các dữ liệu chuyên sâu trong quá trình sản xuất kinh doanh, đồng thời còn đưa ra một cái nhìn tổng quát về năng xuất của các bộ phận, các xu hướng, các hoạt động, các chỉ số KPI (Key Perormance Indicator – chỉ số đánh giá thực hiện công việc).

Hay nói cách khác, Dashboard là bảng thông tin tổng hợp kết hợp nhiều Báo cáo (Report) trong một màn hình hiển thị.

Dashboard thường được đưa ra bởi các chuyên gia nhằm tìm ra xu hướng hỗ trợ việc ra các quyết định hoạt động của tổ chức sao cho hiệu quả.

Xây dựng Dashboard thường tập trung đi trả lời những câu hỏi kinh doanh, trả lời các câu hỏi quan trọng về doanh nghiệp. Bởi vậy Dashboard phân tích dữ liệu nhanh, tập trung và hiệu quả.

3.2. Lợi ích của việc xây dựng Dashboard

+ Trực quan và sinh động vì chủ yếu gồm các biểu đồ, đồ thị và hình ảnh giúp các nhà phân tích có thể tìm ra vấn đề một cách nhanh chóng.
+ Giảm áp lực cho người trình bày khi đọc vì báo cáo thường chỉ tóm gọn trong một màn hình trình chiếu.
+ Trình bày thông tin tổng quan mang tính hỗ trợ đưa ra hành động, quyết định.
+ Linh hoạt, dễ dàng cho phép người dùng tương tác để lựa chọn các phương án, chỉ tiêu khác nhau từ tổng quan đến chi tiết một cách nhanh chóng và kịp thời cho việc đưa ra quyết định.
+ Tính tự động hóa, cập nhập dữ liệu dễ dàng, tiện lợi giúp tiết kiệm thời gian.

PHẦN II: BÀI TẬP THỰC HÀNH

1. Đặt vấn đề

Trước sự phát triển không ngừng của xã hội làm cho cuộc sống của người dân dần được cải thiện hơn, trong đó vấn đề chăm sóc sức khỏe được đặt lên hàng đầu. Các nhà điều hành đã dựa trên các dự liệu cụ thể nhằm đánh giá tình trạng sức khỏe của người dân Mỹ trong giai đoạn 5 năm từ năm 2010 đến 2015 muốn cải thiện về cơ sở vật chất, cũng như dịch vụ chăm sóc sức khỏe tốt nhất cho người dân Mỹ. Cuộc khảo sát được thực hiện trên tất cả các bang trên đất nước Mỹ cho thấy được tình trạng chung về sức khỏe cơ bản để từ đó tăng chi phí hỗ trợ y tế cộng đồng đến từng bang một cách hiệu quả và hợp lý. Trước vấn đề này, nhóm chúng em sẽ phân tích vấn đề sức khỏe trên bang Alabamas để tìm hiểu tình trạng sức khỏe hiện tại cũng như cơ sở hạ tầng, dịch vụ y tế hỗ trợ việc đưa ra quyết định của các nhà điều hành.

2. Dữ liệu

2.1. Nguồn dữ liệu

- Dữ liệu ban đầu lấy từ BRFSS – hệ thống khảo sát các yếu tố rủi ro hành vi của sức khỏe người dân trong 50 bang ở Mỹ và các vùng lãnh thổ ngoài Mỹ. Các yếu tố được BRFSS đánh giá bao gồm sử dụng thuốc lá, các dịch vụ chăm sóc sức khỏe, kiến thức hoặc phòng ngừa HIV/AIDS, hoạt động thể chất và tiêu thụ trái cây và rau quả. Dữ liệu thu thập từ mẫu ngẫu nhiên với gần 500 000 người từ năm 2011 – 2015 thông qua các cuộc khảo sát trên điện thoại.
- Link dữ liệu: BehavioralRiskFactorSurveillanceSystem | Kaggle.

2.2 Mô tả dữ liệu

- Dữ liệu gồm:
  + 5 file dữ liệu tương ứng với từng năm 2011, 2012, 2013, 2014, 2015.
  + 1 file mô tả thuộc tính các cột có trong file.
- Kích thước: 2.68GB
- Mỗi file bao gồm 1758 cột tương ứng với 1758 thuộc tính.
- Số bản ghi: khoảng 500 000 bản ghi / file.
- Dữ liệu chủ yếu là dạng có cấu trúc.

3. Yêu cầu phân tích

Dựa trên bộ dữ liệu thực tế trên, nhóm chúng em đặt ra 2 yêu cầu phân tích cho bài toán như sau:

- Đánh giá mức độ độc hại từ việc uống rượu, sử dụng thuốc lá và mức độ tiếp xúc khói thuốc thụ động đã tác động đến bệnh viêm phổi của công dân ở bang Alabama, Mỹ.
- Theo dõi tình trạng mắc bệnh tiểu đường và tình trạng thừa cân, béo phì đang diễn ra có xu hướng tăng của công dân đang sinh sống tại bang Alabama, Mỹ.

Từ nhưng yêu cầu trên chúng em xin đặt ra các tiêu chí để phân tích bao gồm:

- Phân tích tỷ lệ mắc bệnh tiểu đường dựa trên các tiêu chí về độ tuổi, chỉ số cơ thể, theo chủng tộc…

- Phân tích tỷ lệ mắc bệnh viêm phổi theo độ tuổi,… và trình trạng hút thuốc, tập thể dục ảnh hưởng đến sức khỏe của người dân Alabamas.

4. Xây dựng kho dữ liệu

4.1. Quá trình ETL

- Giai đoạn trích xuất – Extract

  + Trích xuất khoảng 39260 mẫu dữ liệu từ 5 file tương ứng với 5 năm 2011, 2012, 2013, 2014, 2015.

  + Lựa chọn ra 35 cột thuộc tính phù hợp với yêu cầu bài toán phân tích ban đầu từ mỗi file.

- Giai đoạn chuyển đổi – Transform

  Bước 1: Làm sạch dữ liệu, thống nhất một kiểu dữ liệu cho mỗi cột thuộc tính

	A	B	C	D	E	F	G	H	I
1	IMONTH	IDAY	IYEAR	GENHLTH	PHYSHLTH	MENTHLTH	EXERANY2	CHECKUP1	POORHLTH
2	b'01'	b'20'	b'2011'	4	88	30	2	1	88
3	b'01'	b'14'	b'2011'	4	12	4	2	1	4
4	b'01'	b'06'	b'2011'	2	88	3	1	1	88
5	b'02'	b'01'	b'2011'	3	88	88		7	
6	b'02'	b'01'	b'2011'	5	25	15	1	1	25
7	b'01'	b'06'	b'2011'	2	88	88	1	1	
8	b'02'	b'01'	b'2011'	3	88	88		1	
9	b'01'	b'06'	b'2011'	5	30	30	2	1	30
10	b'01'	b'06'	b'2011'	3	1	2		1	88
11	b'02'	b'01'	b'2011'	4	10	88		1	88
12	b'02'	b'01'	b'2011'	1	88	3		1	3
13	b'01'	b'06'	b'2011'	2	88	88	2	1	
14	b'02'	b'01'	b'2011'	3	88	88		1	
15	b'05'	b'31'	b'2011'	4	28	88		1	88
16	b'04'	b'30'	b'2011'	3	88	88	2	1	

Dữ liệu được lựa chọn

	A	B	C	D	E	F
1	IYEAR	IDAY	IMONTH	PHYSHLTH	GENHLTH	POORHLTH
2	2011	20	1	0	Fair	0
3	2011	14	1	12	Fair	4
4	2011	6	1	0	Very good	0
5	2011	1	2	0	Good	0
6	2011	1	2	25	Poor	25
7	2011	6	1	0	Very good	0
8	2011	1	2	0	Good	0
9	2011	6	1	30	Poor	30
10	2011	6	1	1	Good	0
11	2011	1	2	10	Fair	0
12	2011	1	2	0	Excellent	3
13	2011	6	1	0	Very good	0
14	2011	1	2	0	Good	0
15	2011	31	5	28	Fair	0
16	2011	30	4	0	Good	0
17	2011	30	4	0	Good	0

Đã làm sạch và thống nhất kiểu dữ liệu

Bước 2: Chia dữ liệu đã được xử lí thành từng bảng và thêm các cột ID tương ứng cho từng bảng

ID_PEOPLE	SEX	EMPLOY	INCOME	EDUCA	AGE	MARITAL	PRACE	HEIGHT	WEIGHT
P1	Female	Employed for wages	20.000 – 25.000$	Graduate High School	61	Married	White	152	43,09
P2	Male	Unable to work	25.000 – 35 000$	Did not graduate High School	32	Single	White	175	10,41
P3	Female	Employed for wages	25.000 – 35 000$	Graduated from College or Technical School	9	Widowed	White	168	47,63
P4	Female	Employed for wages		Graduate High School	9	Married	White	165	70,31
P5	Female	Unable to work	<10.000$	Did not graduate High School	54	Seperated	Black or African American	155	10,33
P6	Male	Retired		Graduate High School	60	Married	Black or African American	183	98,88
P7	Female	Retired			9	Widowed	White	157	
P8	Male	Unable to work		Graduate High School	51	Divorced	White	173	92,99
P9	Male	Self-employed	15.000 – 20.000$	Graduate High School	26	Single	Black or African American	178	63,50
P10	Female	Retired	35 000 – 50 000$	Graduated from College or Technical School	68	Widowed	White	178	
P11	Female	Out of work for less than 1 year	<10.000$	Did not graduate High School	42	Single	Black or African American	152	92,08
P12	Female	Retired		Graduated from College or Technical School	90	Widowed	White	175	65,77
P13	Female	A homemaker		Graduate High School	70	Married	White	170	72,57
P14	Male	Retired	25.000 – 35 000$	Graduate High School	80	Divorced	White	183	90,72
P15	Female	Retired	25.000 – 35 000$	Graduate High School	85	Widowed	White	160	56,70
P16	Female	Employed for wages	25.000 – 35 000$	Graduate High School	67	Single	Black or African American	170	92,53
P17	Female	A homemaker		Attended College or Technical School	29	Married	White	170	90,72
P18	Female	Employed for wages		Attended College or Technical School	43	Single	Black or African American	168	11,40
P19	Male	Employed for wages	>75 000$	Attended College or Technical School	46	Married	White	183	13,54
P20	Female	Self-employed	25.000 – 35 000$	Attended College or Technical School	51	Married	White	160	67,13
P21	Female	Self-employed		Graduate High School	53	Married	White	165	
P22	Male	Out of work for less than 1 year	15.000 – 20.000$	Graduate High School	19	Single	Others	180	72,57
P23	Female	Employed for wages	>75 000$	Graduated from College or Technical School	30	Married	White	155	61,23
P24	Female	Unable to work	<10.000$	Graduate High School	52	Divorced	Black or African American	160	95,25
P25	Female	Employed for wages	>75 000$	Graduated from College or Technical School	41	Married	White	170	91,63
P26	Female	Employed for wages	50 000 – 75 000$	Graduated from College or Technical School	35	Married	White	175	99,79
P27	Female	Unable to work	<10.000$	Did not graduate High School	49	Divorced	Black or African American	165	10,86
P28	Female	A homemaker	35 000 – 50 000$	Graduate High School	32	Married	White	163	

Bảng PEOPLE

ID_TIME	ID_PEOPLE	IYEAR	IDATE
T1	P1	2011	01202011
T2	P2	2011	01142011
T3	P3	2011	01062011
T4	P4	2011	02012011
T5	P5	2011	02012011
T6	P6	2011	01062011
T7	P7	2011	02012011
T8	P8	2011	01062011
T9	P9	2011	01062011
T10	P10	2011	02012011
T11	P11	2011	02012011
T12	P12	2011	01062011
T13	P13	2011	02012011
T14	P14	2011	05312011
T15	P15	2011	04302011
T16	P16	2011	04302011
T17	P17	2011	06302011
T18	P18	2011	06302011
T19	P19	2011	06302011
T20	P20	2011	08292011
T21	P21	2011	08302011
T22	P22	2011	10312011
T23	P23	2011	12302011
T24	P24	2011	12302011
T25	P25	2011	11302011
T26	P26	2011	11302011
T27	P27	2011	12302011
T28	P28	2011	05312011

Bảng TIME

ID_PEOPLE	PDIABTST	PREDIAB	INSULIN	DOCTDIAB	DIABEDU	DIABAGE
P32731	Yes	3				
P32732	Yes	No				
P32733			Yes	4	1	32
P32734			No	4	2	72
P32735	Yes	3				
P32736	Yes	3				
P32737			No	2	2	63
P32738	No	3				
P32739	Yes	3				
P32740	No	3				
P32741			No	6	2	40
P32742			Yes	3	2	62
P32743	Yes	3				
P32744	Yes	Yes				
P32745			Yes	3	1	46
P32746	Yes	3				
P32747	Yes	3				
P32748			Yes	2	1	51
P32749	No	3				
P32750	Yes	3				
P32751	No	3				
P32752	Yes	3				
P32753	Yes	3				
P32754	No	3				
P32755	No	3				
P32756	Yes	3				
P32757	Yes	3				
P32758	Yes	3				

Bảng DIABETE

ID_PEOPLE	HLTHPLN	PERSDOC	CHECKUP	EXERANY	MEDCOST
P1	Yes	>1	Less than 12 months	No	No
P2	Yes	1	Less than 12 months	No	Yes
P3	Yes	1	Less than 12 months	Yes	No
P4	Yes	1			No
P5	Yes	1	Less than 12 months	Yes	No
P6	Yes	1	Less than 12 months	Yes	No
P7	Yes	1	Less than 12 months		No
P8	No	1	Less than 12 months	No	Yes
P9	No	No	Less than 12 months		No
P10	Yes	1	Less than 12 months		No
P11	No	1	Less than 12 months		Yes
P12	Yes	1	Less than 12 months	No	No
P13	Yes	1	Less than 12 months		No
P14	Yes	1	Less than 12 months		No
P15	Yes	1	Less than 12 months	No	No
P16	Yes	1	Less than 12 months	No	No
P17	No	No	More 5 years	Yes	Yes
P18	No	>1	Less than 12 months		Yes
P19	Yes	1	Less than 12 months	Yes	No
P20	Yes	1	Less than 12 months		No
P21	Yes	1	Less than 12 months	Yes	No
P22	No	No	2 - 5 years	Yes	Yes
P23	No	1	Less than 12 months	Yes	No
P24	No	1	Less than 12 months		Yes
P25	Yes	1	2 - 5 years		No
P26	Yes	1	1 - 2 years		No
P27	No	1	Less than 12 months		Yes
P28	Yes	No	Less than 12 months	Yes	No

Bảng HEALTH CARE ACCESS

ID_TIME	GENHLTH	PHYSHLTH	MENTHLTH	POORHLTH
T1	Fair	0	30	0
T2	Fair	12	4	4
T3	Very good	0	3	0
T4	Good	0	0	
T5	Poor	25	15	25
T6	Very good	0	0	
T7	Good	0	0	
T8	Poor	30	30	30
T9	Good	1	2	0
T10	Fair	10	0	0
T11	Excellent	0	3	3
T12	Very good	0	0	
T13	Good	0	0	
T14	Fair	28	0	0
T15	Good	0	0	
T16	Good	0	0	
T17	Good	20	30	15
T18	Good	0	0	
T19	Fair	0	0	
T20	Fair	10	5	7
T21	Fair	14	14	14
T22	Good	0	2	0
T23	Very good	0	0	
T24	Poor	30	10	30
T25	Excellent	0	3	0
T26	Excellent	0	15	0
T27	Fair	10	15	15
T28	Very good	5	0	0

Bảng HEALTH STATUS

ID_PEOPLE	ID_USETO	CHCCOPD	DIABETE
P1	U1	No	No
P2	U2	No	No
P3	U3	Yes	No
P4	U4	No	No
P5	U5	Yes	No
P6	U6	No	Yes
P7	U7	No	No
P8	U8	No	No
P9	U9	No	No
P10	U10	No	No
P11	U11	No	No
P12	U12	No	No
P13	U13	No	No
P14	U14	No	No
P15	U15	No	No
P16	U16	No	No
P17	U17	Yes	No
P18	U18	No	No
P19	U19	No	Yes
P20	U20	No	Yes
P21	U21	No	No
P22	U22	No	No
P23	U23	Yes	No
P24	U24	Yes	No
P25	U25	No	No
P26	U26	No	No
P27	U27	No	No
P28	U28	No	No

Bảng DISEASE

ID_TIME	ID_PEOPLE	SMOKDAY	STOPSMK	LASTSMK	USENOW
T1	P1	Everyday	Yes		Not at all
T2	P2	Everyday	Yes		Not at all
T3	P3	Not at all		>10 years	Not at all
T4	P4				Not at all
T5	P5	Not at all		>10 years	Not at all
T6	P6	Not at all		>10 years	Not at all
T7	P7				Not at all
T8	P8	Some days	No		Not at all
T9	P9				Not at all
T10	P10				Not at all
T11	P11	Not at all		>10 years	Not at all
T12	P12	Not at all		>10 years	Not at all
T13	P13				Not at all
T14	P14	Not at all		>10 years	Not at all
T15	P15				Not at all
T16	P16	Everyday	No		Not at all
T17	P17	Everyday	Yes		Not at all
T18	P18				Not at all
T19	P19	Not at all		1-5 years	Not at all
T20	P20	Everyday	Yes		Not at all
T21	P21				Not at all
T22	P22				Not at all
T23	P23				Not at all
T24	P24	Some days	Yes		Not at all
T25	P25				Not at all
T26	P26	Everyday	Yes		Not at all
T27	P27				Not at all
T28	P28				Not at all

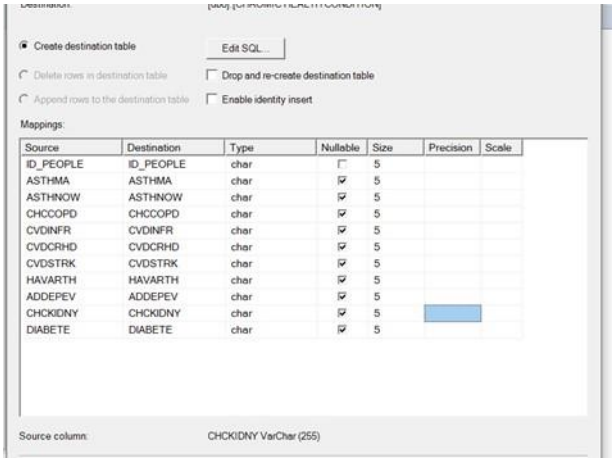
Bảng USE TOXIC

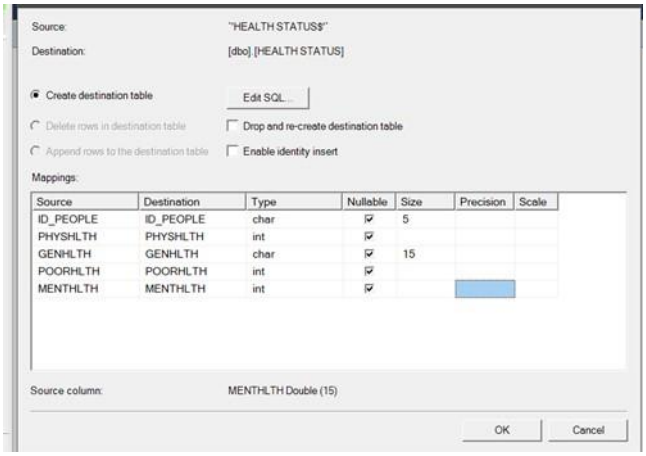
ID_USETOXIC
U1
U2
U3
U4
U5
U6
U7
U8
U9
U10
U11
U12
U13
U14
U15
U16
U17
U18
U19
U20
U21
U22
U23
U24
U25
U26
U27
U28

Bảng CHCCOPD

- Giai đoạn tải – Load
  + Sử dụng SQL Server để lưu trữ dữ liệu.

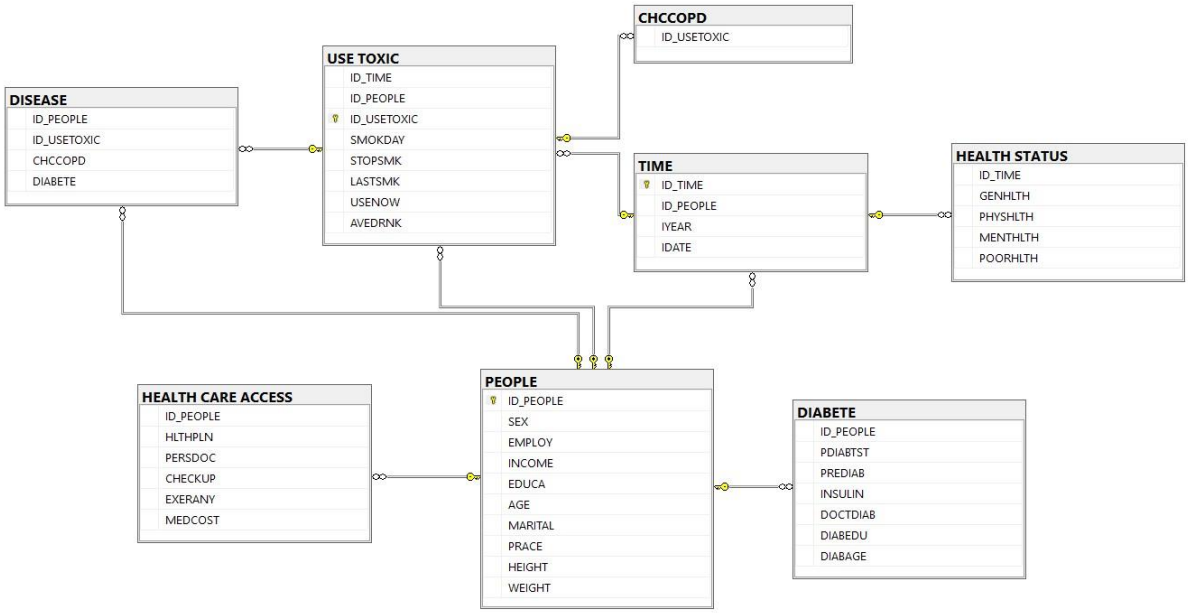
+ Thiết lập các khóa chính, khóa ngoại thể hiện mối quan hệ ràng buộc của từng bảng với nhau





4.2. Xây dựng DW

• Mô hình logic – Mô hình ERD



- Gồm 8 thực thể:

PEOPLE: Người tham gia khảo sát

Tên	Mô tả
SEX	Giới tính
EMPLOY	Nghề nghiệp
INCOME	Mức lương

EDUCE	Trình độ học thức
AGE	Tuổi
MARITAL	Tình trạng hôn nhân
PRACE	Chủng tộc
HEIGHT	Chiều cao
WEIGHT	Cân nặng

TIME: Thời gian tham gia khảo sát

IYEAR	Năm tham gia khảo sát
IDATE	Ngày tháng năm tham gia khảo sát

HEALTH STATUS: trạng thái sức khỏe

GENHLTH	Trạng thái sức khỏe hiện tại
PHYSHLTH	Số ngày thể chất không khỏe trong 30 ngày qua
MENTHLTH	Số ngày sức khỏe tinh thần không tốt
POORHLTH	Số ngày sức khỏe chung không tốt trong 30 ngày

HEALTH CARE ACCESS: tiếp cận chăm sóc sức khỏe

HLTHPLN	Bảo hiểm chăm sóc sức khỏe
PERSDOC	Số lượng bác sĩ riêng
CHECKUP	Lần cuối kiểm tra định kỳ sức khỏe cách đây bao lâu
EXERANY	Tập thể dục trong vòng 30 ngày
MEDCOST	Không uống thuốc vì không đủ tiền

DISEASE: Bệnh

CHCCOPD	Bệnh viêm phổi
DIABETE	Bệnh tiểu đường

DIABETE: Bệnh tiểu đường

PDIABTST	Đã đo lượng đường có trong máu trong 3 năm qua chưa
PREDIAB	Bệnh đái tháo đường
INSULIN	Có đang dùng Insulin không?
DOCTDIAB	Kiểm tra tiểu đường trong 12 tháng qua?
DIABEDU	Tham gia lớp học tiểu đường
DIABAGE	Tuổi bị tiểu đường

USE TOXIC: Thói quen xấu ảnh hưởng tới phổi

SMOKDAY	Tần suất hút thuốc
STOPSMK	Đã ngừng hút thuốc trong 12 tháng qua
LASTSMK	Thời gian hút thuốc cuối cùng cách khoảng bao lâu
USENOW	Sử dụng thuốc lá không khói

AVEDRNK	Số ngày uống rượu trong 30 ngày qua

CHCCOPD: Bệnh viêm phổi

- Một số truy vấn để kiểm tra tính đúng đắn và chính xác, khách quan của dữ liệu

```
select count(dbo.PEOPLE.ID_PEOPLE) as SUM_PEOPLE_SMOKDAY
from dbo.PEOPLE inner join dbo.[USE TOXIC]
on dbo.PEOPLE.ID_PEOPLE = dbo.[USE TOXIC].ID_PEOPLE
and SMOKDAY = 'Everyday';
```

	SUM_PEOPLE_SMOKDAY
1	5132

Những người hút thuốc lá mỗi ngày

```
select count(dbo.PEOPLE.ID_PEOPLE) as SUM_PEOPLE_INSULIN
from dbo.PEOPLE inner join dbo.DIABETE
on dbo.PEOPLE.ID_PEOPLE = dbo.DIABETE.ID_PEOPLE
and INSULIN  = 'yes';
```

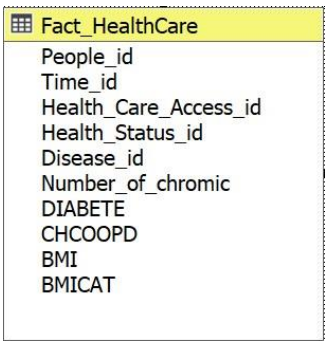
	SUM_PEOPLE_INSULIN
1	837

Những người dùng Insulin

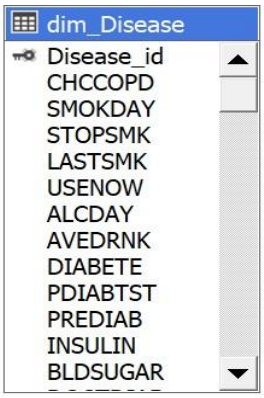
• Xây dựng mô hình vật lí kiểu quan hệ

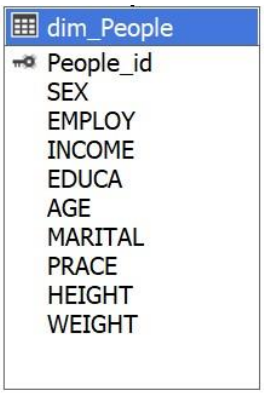
- Bảng Fact được xây dựng gồm có:
  - ✦ 5 dimension tables
    + dim_PEOPLE: lưu trữ thông tin về người tham gia khảo sát
    + dim_TIME: lưu trữ thông tin về thời gian phỏng vấn qua điện thoại
    + dim_HEALTH_CARE_ACCESS: lưu trữ thông tin về việc truy cập sử dụng các dịch vụ sức khỏe
    + dim_HEALTH_STATUS: lưu trữ thông tin về tình trạng sức khỏe hiện tại và việc sử dụng các sản phẩm có hại cho sức khỏe
    + dim_DISEASE: lưu trữ thông tin về 2 bệnh là viêm phổi và tiểu đường
  - ✦ 4 measures là:
    + Number_of_chromic: Số bệnh mà người tham gia mắc phải
    + DIABETE: Người tham gia khảo sát có mắc bệnh tiểu đường hay không
    + CHCOOPD: Người tham gia khảo sát có bị viêm phổi hay không
    + BMI: Chỉ số đánh giá tình trạng cơ thể của người tham gia khảo sát

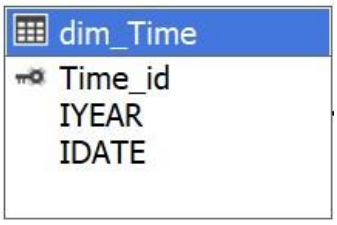
Bảng Fact:

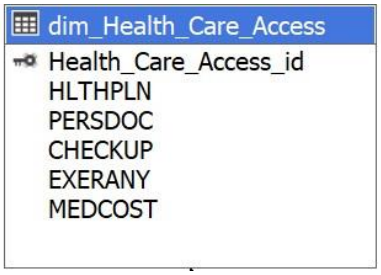


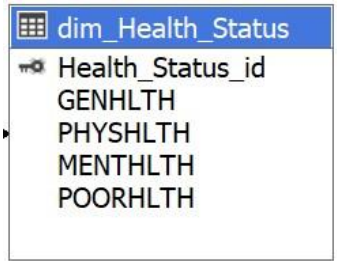
Bảng dim:



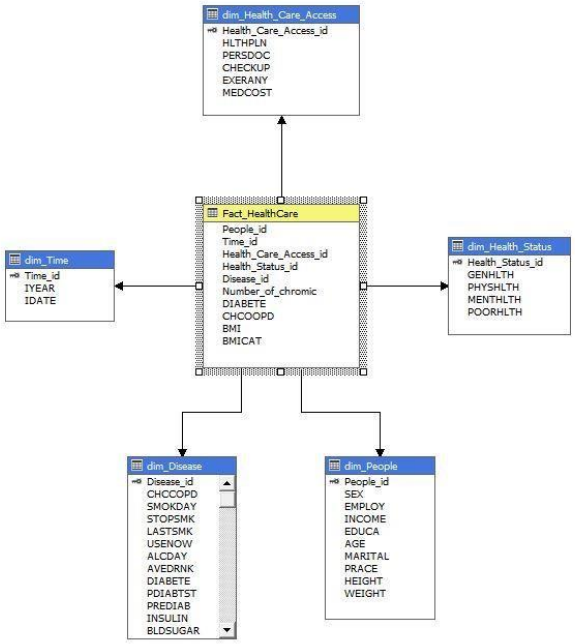








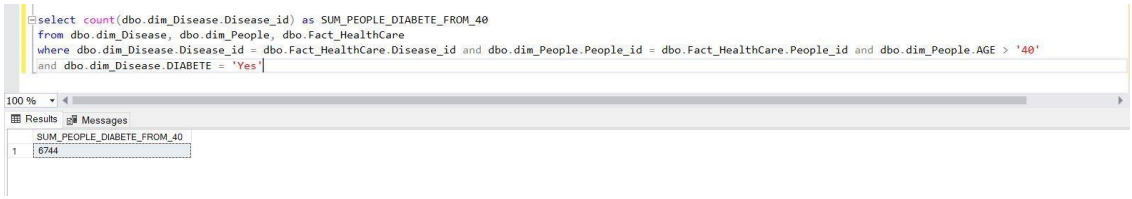
Lược đồ Star:



Một số truy vấn vào OLAP để kiểm tra thông tin

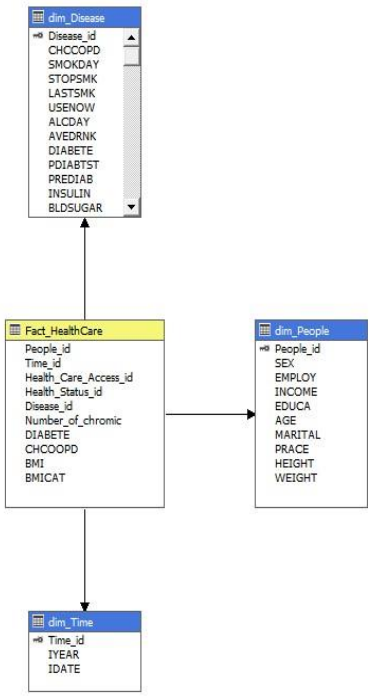
- Số lượng người tham gia khảo sát mắc bệnh viêm phổi

- Số người trên 40 tuổi tham gia khảo sát bị mắc bệnh tiểu đường

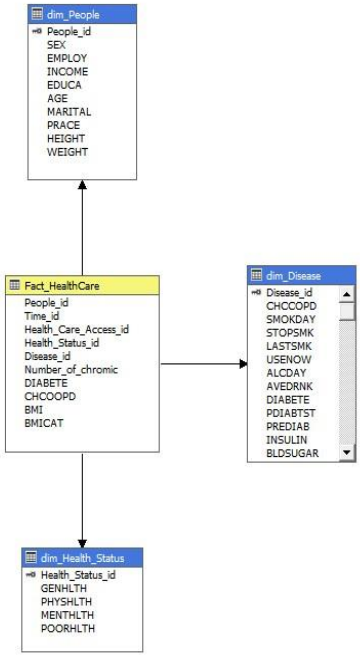


Đứng trên lập trường của người truy vấn thông tin từ dữ liệu, nhóm em xây dựng thêm 3 cube nhỏ mà nhóm nghĩ đây là dữ liệu được quan tâm và truy vấn phổ biến nhất

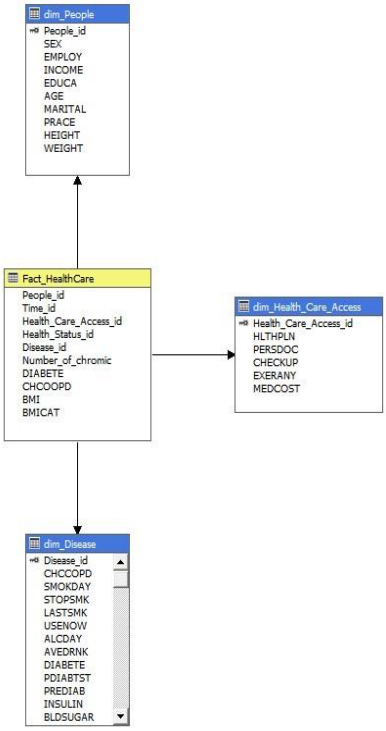
Cube 1: Những người tham gia khảo sát qua từng năm bị mắc bệnh nằm nhiều nhất trong tầm tuổi nào, tỷ lệ giới tính, hoàn cảnh sống của họ ra sao



Cube 2: Những người tham gia khảo sát đang có tình trạng sức khỏe như thế nào



Cube 3: Những người tham gia khảo sát họ có thực sự quan tâm đến sức khỏe bản thân không, điều kiện chăm sóc bởi dịch vụ y tế của từng người như thế nào



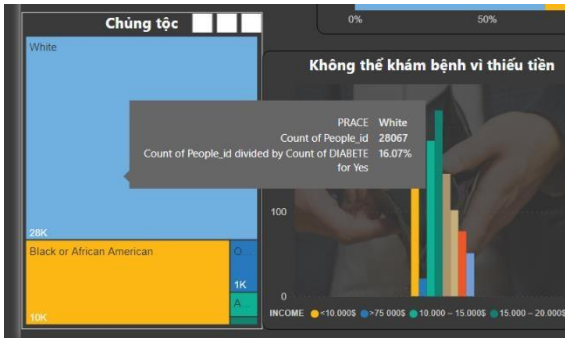
5. Xây dựng Dashboard

Một nghiên cứu tại nước Mĩ cho thấy, tỷ lệ người mắc bệnh tiểu đường và viêm phổi ở Mĩ ngày càng gia tăng. Riêng trong năm 2010, đã có khoảng 18,8 triệu người Mĩ được chuẩn đoán mắc bệnh tiểu đường. Tỉ lệ mắc bệnh ở các bang ở Mĩ là rất cao, đặc biệt ở bang Alabama. Vì thế bài toán đặt ra là phân tích tỉ lệ mắc bệnh tiểu đường và viêm phổi của người dân Alabama qua các khía cạnh về chăm sóc sức khỏe nhằm đưa ra các quyết định y tế nhằm giảm tỉ lệ mắc bệnh ở người dân.

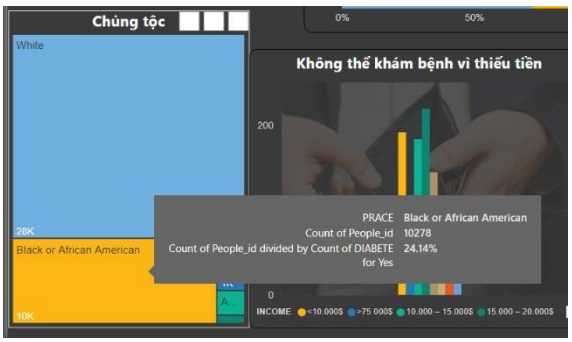
Từ đó, nhằm đưa ra các phân tích hỗ trợ quyết định hiệu quả, chúng em sẽ thể hiện qua 2 bảng dashboard.

✞ Dashboard phân tích tỉ lệ mắc bệnh tiểu đường của người dân Alabama

- Về thiết kế: Dashboard gồm 2 lát cắt theo năm, theo độ tuổi, và các con số quan trọng để nắm bắt dữ liệu tổng quan nhất. Đó là số lượng người tham gia khảo sát, số người mắc bệnh từ đó đưa ra tỉ lệ mắc bệnh.
- Với dữ liệu ban đầu gồm cả 5 năm, ta thấy tỉ lệ mắc bệnh trung bình là 18.1% và thường tập trung vào độ tuổi trên 35 tuổi, đặc biệt độ tuổi trên 65 tuổi chiếm đến gần 50%. Con số này hiển nhiên, bởi bệnh tiểu đường thường xảy ra ở người lớn tuổi.
- Nhìn vào biểu đồ chủng tộc, ta thấy về mật độ dân cư, người da trắng chiếm tỉ trọng cao nhất, sau đó là người da màu (người Mĩ gốc Phi), rồi đến những chủng tộc khác. Tuy nhiên khi nhìn vào tỉ lệ mắc bệnh trong những người tham gia khảo sát:



Tỉ lệ mắc bệnh ở người da trắng: 16.07%



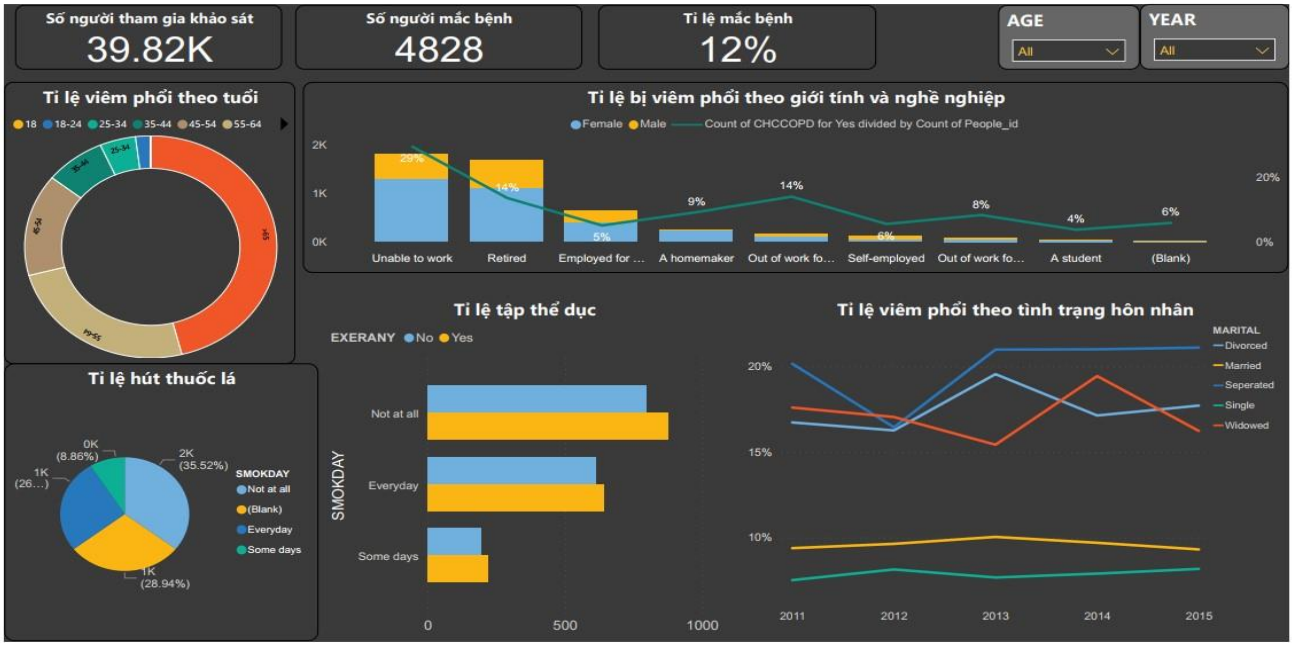
Tỉ lệ mắc bệnh ở người da đen 24.14%

Ta thấy tỉ lệ mắc bệnh tiểu đường ở người da màu cao hơn và lên đến 24%. Tại sao tỉ lệ mắc bệnh ở người da màu lại cao vượt trội như thế? Một phần đó là do vấn đề phân biệt chủng tộc Mĩ. Mặc dù phong trào dân quyền đã chấm dứt được luật phân biệt chủng tộc tồn tại trong suốt một thế kỉ tại Mỹ, thế nhưng tình trạng phân biệt chủng tộc đối với người da màu thực sự là chưa bao giờ hết đặc biệt là người Mĩ da màu và người Mĩ gốc Phi. Không khó qua sát thấy phần lớn người Mĩ gốc Phi thường làm những công việc tay chân hơn, sống tập trung ở những khu nhà nghèo khó và ô nhiễm. • Nhìn vào biểu đồ tỉ lệ mắc bệnh dựa trên thu nhập hay không thể khám bệnh vì thiếu chi phí. Ta thấy nhóm người không thể khám chữa bệnh thường tập trung ở nhóm người có tổng thu nhập gia đình nhỏ hơn 20.000$, ứng với tỉ lệ mắc bệnh ở nhóm này luôn ở mức cao. Tỉ lệ mắc bệnh ở nhóm người có tổng thu nhập gia đình qua các năm dưới 15.000$ là lớn nhất. Từ đó trả lời cho câu hỏi trước, thì người Mĩ gốc Phi đa số có tổng thu nhập gia đình thấp do vấn đề phân biệt chủng tộc, cuộc sống vật chất, hay chất lượng cuộc sống cũng ảnh hưởng rất nhiều đến sức khỏe.

- Một yếu tố quan trọng khi đánh giá tỉ lệ mắc bệnh đó là tình trạng béo phì. Biểu đồ chỉ số BMI, chỉ số đánh giá tình trạng dinh dưỡng cơ thể theo chiều cao và cân nặng với 4 mức: BMI<18:gầy; BMI 18-25 : bình thường; từ 20-30 là thừa cân, >30 béo phì.
- Khi khảo sát những người mắc bệnh tiểu đường ở người dân Alabama, đa số người mắc bệnh đều nằm trong khoảng thừa cân hay béo phì, thừa cân và béo phì chiếm gần 60%.
  Chẳng hạn: với những người bị béo phì tỉ lệ mắc bệnh lên đến 28.25% cao hơn rất nhiều so với trung bình thực tế.

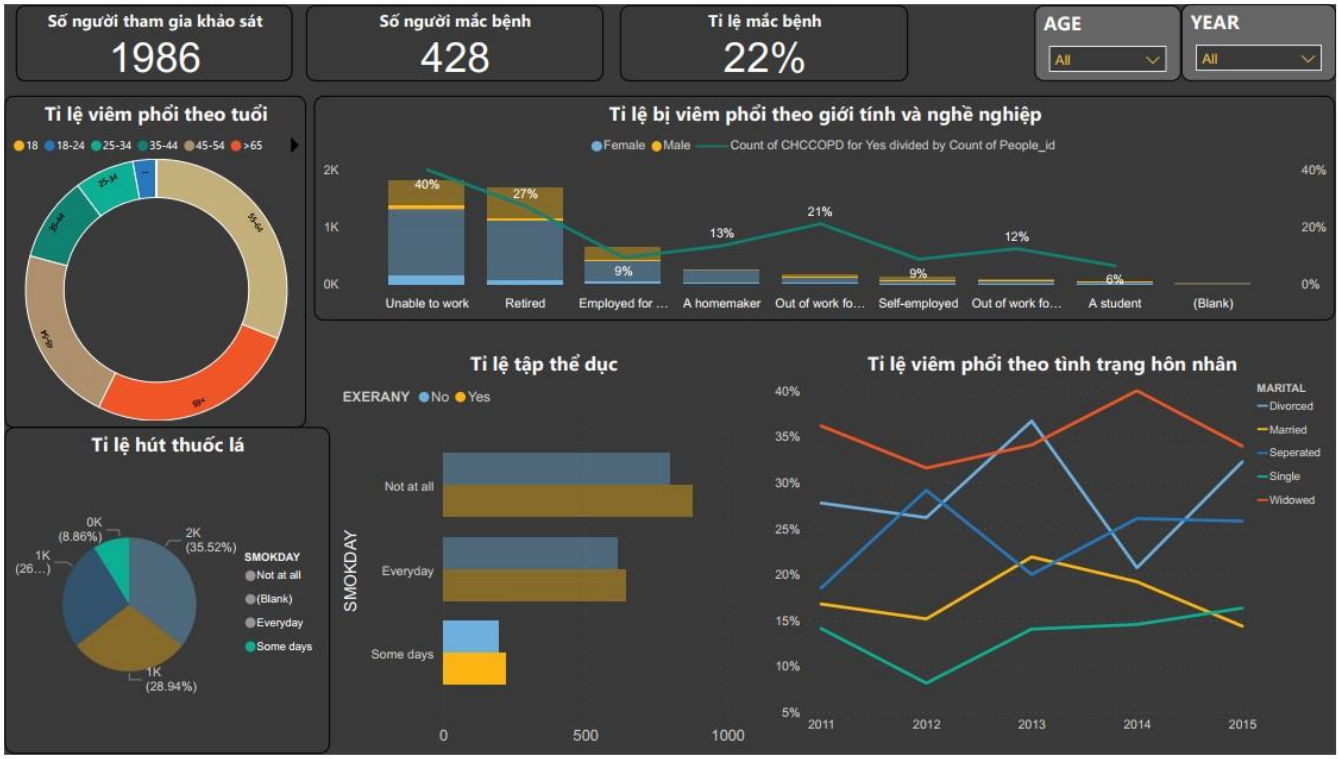
❾ Tóm lại ta có thể rút ra kết luận, bệnh tiểu đường ở bang Alabama trong 5 năm từ 2011-2015 là rất cao, tập trung chủ yếu ở người lớn, những người thừa cân hay béo phì. Đặc biệt là những người béo phì có tỉ lệ mắc bệnh rất cao, ở mức báo động, ngoài ra lối sống sinh hoạt cũng ảnh hưởng rất nhiều đến việc mắc bệnh.

✞ Dashboard đánh giá tỉ lệ mắc bệnh viêm phổi của người dân ở Alabama



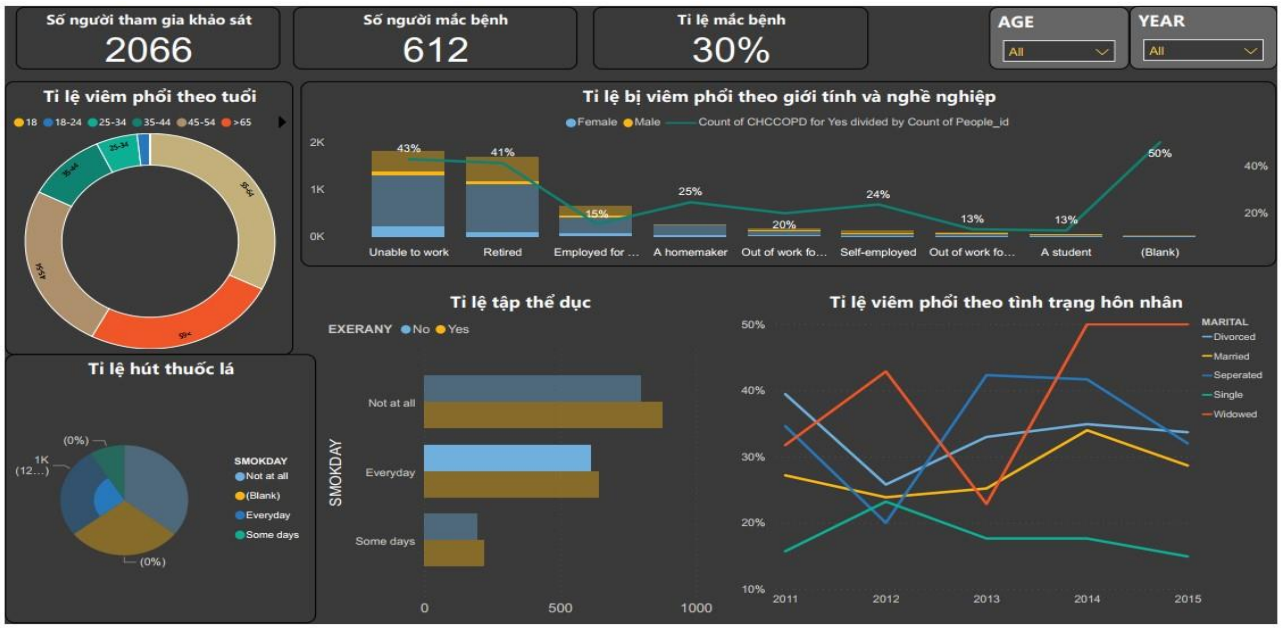
- Đánh giá tỉ lệ bị viêm phổi của người dân Alabama qua các khía cạnh: sử dụng thuốc lá, tập thể dục, giới tính, nghề nghiệp.
- Với hai lát cắt, ta có thể đánh giá tỉ lệ qua các năm hoặc qua những độ tuổi khác nhau. Trước hết với mẫu dữ liệu ban đầu, ta thấy tỉ lệ mắc bệnh viêm phổi trung bình qua các năm là 12% và tập trung chủ yếu ở lứa tuổi trên 35 tuổi.
- Tỉ lệ mắc bệnh ở những người không thể làm việc là cao nhất, lên đến gần 30%.
- Sang cột tình trạng mắc bệnh theo tình trạng hôn nhân ta thấy tỉ lệ mắc bệnh ở nhóm người có tình trạng ly hôn là cao nhất, cao gầm gấp đôi trung bình, và qua các năm đều có xu hướng tăng lên. Ngoài ra tỉ lệ mắc bệnh ở nhóm người ly hôn, li thân, góa luôn nằm ở mức cao. Trong khi đó tỉ lệ mắc bệnh ở nhóm độc thân, hôn nhân luôn ở mức thấp. Lí giải cho việc này thì có lẽ một cuộc sống hạnh phúc sẽ sản sinh enzym hạnh phúc, tăng cường hệ miễn dịch chống lại bệnh tật, kéo dài tuổi thọ.
- Trong những người bị viêm phổi thì tỉ lệ họ hút thuốc mỗi ngày, hay hút thuốc một vài lần trong ngày là cao nhất, ta có thể xem xét tỉ lệ hút thuốc lá mắc bệnh.

Tỉ lệ mắc bệnh viêm phổi ở những người hút thuốc là mỗi ngày: 25%



Tỉ lệ mắc bệnh viêm phổi ở những người hút lá một vài lần trong ngày: 22%

- Ta thấy tỉ lệ mắc bệnh ở những người hút thuốc mỗi ngày hay hút một vài lần trong ngày là cao, điều này dễ hiểu thì hút thuốc là nguyên nhân chính gây bệnh viêm phổi. Ngoài hút thuốc thì thói quen sinh hoạt cũng ảnh hưởng đến tỉ lệ mắc bệnh. Chẳng hạn những người hút thuốc mỗi ngày và không tập thể dục



Tỉ lệ mắc bệnh rất cao, lên đến 30%, tương đương với việc cứ 3 người là có một người mắc bệnh, đây là một con số báo động cao. Bên cạnh đó những người hút thuốc lá mỗi ngày nhưng có tập thể dục thì tỉ lệ mắc bệnh thấp hơn chỉ còn 22%.

❾ Từ đó ta thấy thuốc lá là nguyên nhân chính gây viêm phổi, vì thế nhằm chăm sóc sức khỏe cho bản thân, hạn chế thuốc lá, đồng thời tăng cường tập luyện thể thao để phòng chống giảm nguy cơ mắc bệnh.

Như vậy: *Qua việc phân tích tỉ lệ mắc bệnh tiểu đường và viêm phổi ở người dân Alabama ta thấy tỉ lệ mắc bệnh ở đây rất cao, là những con số báo động, từ đó đòi hỏi chính quyền, bộ y tế phải có những chính sách biện pháp nhằm nâng cao nhận thức của người dân về việc sử dụng thuốc lá, tình trạng cân nặng, và các vấn đề chăm sóc bản thân như việc tập thể dục, sống lành mạnh. Hơn nữa, tập trung giải quyết các vấn đề cộng đồng, tạo công ăn việc làm ổn định cho toàn dân, nâng cao chất lượng cuộc sống giảm tỉ lệ mắc bệnh.*

LỜI KẾT

Sau báo cáo này, nhóm chúng em đã hiểu hơn về các tìm hiểu các cấu trúc dữ liệu, quá trình xử lí và làm sạch dữ liệu, từ đó tổng hợp và xây dựng nên Data Warehouse. Việc xử lí và làm sạch dữ liệu đã được thực hiện trong Excel và SQL Server. Sau khi xử lí và xây dựng xong thì nhóm có làm một số truy vấn dữ liệu trong OLAP để kiểm tra tính đúng đắn khách quan của dữ liệu được đưa vào. Tiếp sau đó nhóm có sự dụng công cụ Power BI để hỗ trợ trong quá trình trích xuất dữ liệu đầu ra, chia cắt kho dữ liệu theo từng lát để thuận tiện cho việc tạo Dashboard.

Tuy nhiên trong quá trình thực hiện, việc lựa chọn dữ liệu còn nhiều sai sót dẫn đến việc phân tích đưa ra quyết định còn non nớt, chưa đi vào mục đích chi tiết.

Tài liệu tham khảo

1. https://labs.flinters.vn/data-science/gioi-thieu-tong-quan-ve-data-warehouse/
2. https://www.hyperlogy.com/vi/tong-quan-ve-business-intelligence
3. https://a1digihub.com/data-dashboard-la-gi/
4. https://www.cdc.gov/brfss/annual_data/2014/pdf/CODEBOOK14_LLCP.pdf
5. https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system