

Vai trò của Data Warehouse trong Business Analytics và Business Intelligence

Data Warehouse (Kho dữ liệu) được dùng làm gì?

Cuộc cách mạng Big Data được kỳ vọng sẽ là mỏ vàng cho các doanh nghiệp trên toàn thế giới. Nhờ nó, chúng ta sẽ thu thập được vô vàn thông tin quan trọng, chi tiết về khách hàng, tổ chức và ngành nghề - hoặc bất cứ điều gì mang lại lợi thế cho kinh doanh. Từ đó, dữ liệu có thể được phân tích và biến thành những hiểu biết dẫn tới hành động, đưa doanh nghiệp tiến xa hơn.

LƯỢNG DỮ LIỆU NGÀY CÀNG TĂNG, VẤN ĐỀ NGÀY CÀNG RỒ: NÊN LƯU TRỮ Ở ĐÂU?

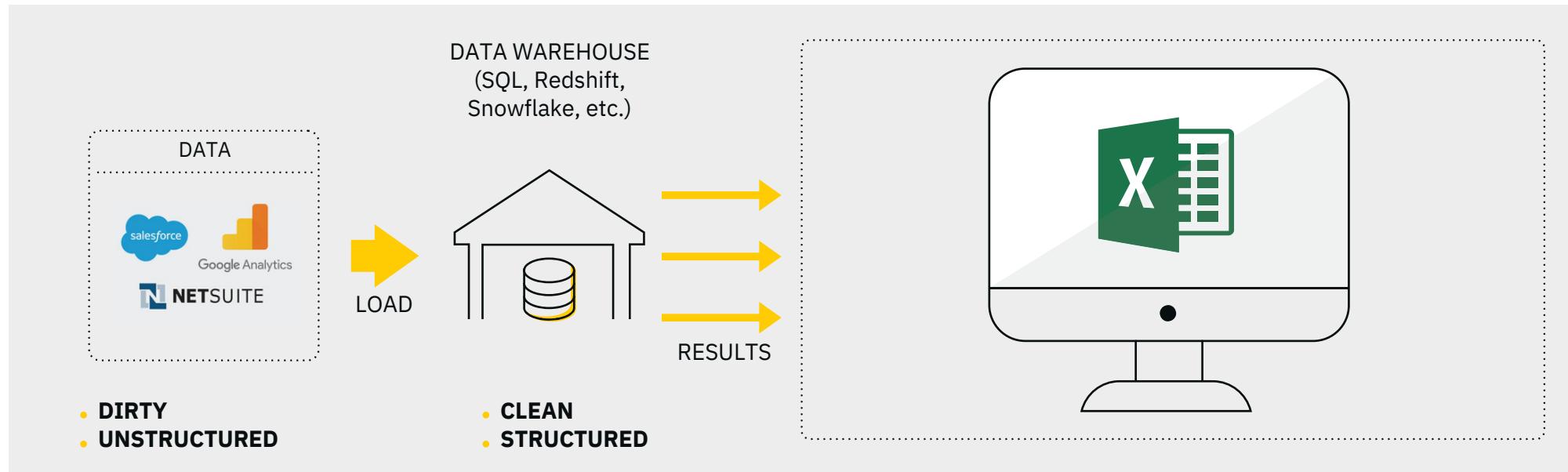
Đó chính là lúc ***data warehouses*** (**Kho dữ liệu**) ra đời. *Data warehouse* cho phép tập hợp tất cả dữ liệu từ nhiều nguồn khác nhau về một nơi, sắp xếp một cách rõ ràng và tổng quát. Cuối cùng, tạo ra một phiên bản đáng tin cậy duy nhất. Đã có rất nhiều khoản đầu tư lớn vào danh mục này nhằm đảm bảo dữ liệu được lưu trữ, truy xuất một nhanh chóng, thuận tiện dựa trên các phương pháp tổng hợp cần thiết có sẵn. Sau đó, doanh nghiệp có thể sử dụng nó như một nguồn dữ liệu tổng hợp duy nhất phục vụ cho 2 công việc BI và BA (*Business Intelligence* và *Business Analytics*).

... Và đó là cũng là lúc chúng ta cần thêm nhiều công cụ BI hơn. Kho dữ liệu có thể là hệ thống lưu trữ hoàn hảo, nhưng vẫn cần một công cụ để trích xuất dữ liệu từ đó, kết nối các nguồn khác nhau, và sau đó là một công cụ để phân tích dữ liệu - Nghĩa là bạn cần nhiều công cụ khác nhau, lập trình viên và báo cáo để làm việc trực tiếp với dữ liệu đã được lưu trữ. Một công cụ nào đó cho phép người dùng “đào sâu” vào dữ liệu, tìm những gì họ cần, phân tích và rút ra những thông tin chính để đưa ra quyết định kinh doanh tốt hơn.

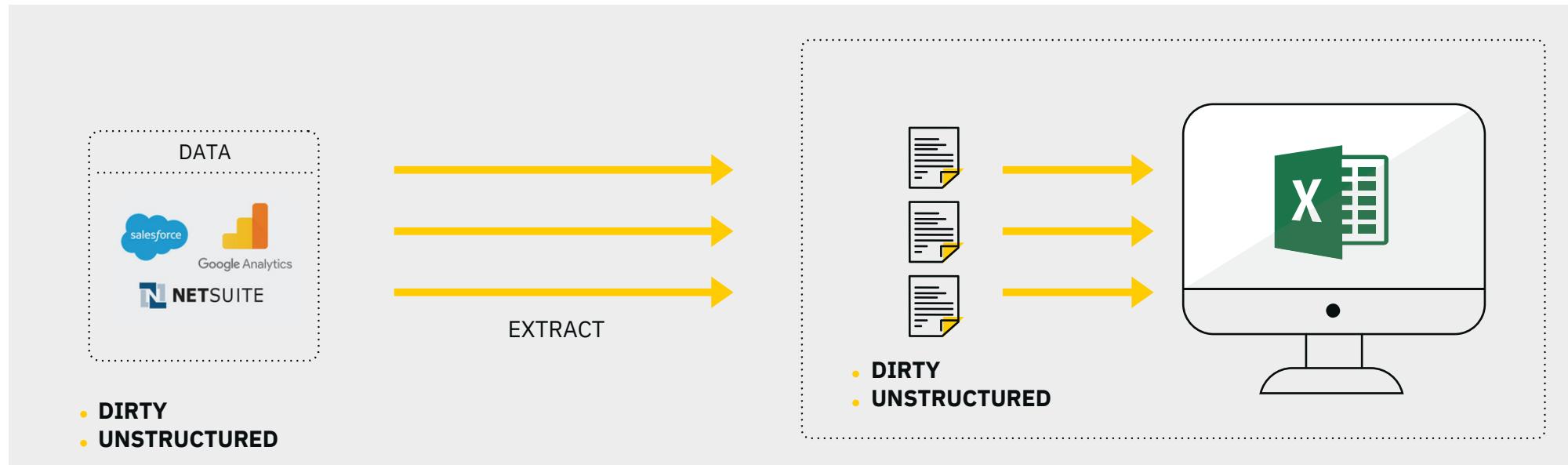
Ví dụ, có những giải pháp cho phép trích xuất dữ liệu bạn cần, thao tác và phân tích nó một cách thuận tiện, đồng thời cập nhật dữ liệu một cách nhanh chóng mỗi khi dữ liệu thay đổi. Với công nghệ này, thậm chí có thể truy cập trực tiếp vào nguồn dữ liệu gốc mà không cần phải tải mọi thứ lên kho dữ liệu trước. Điều này sẽ được trình bày ở các slide phía sau.

Data Warehouse Hoạt Động Như Thế Nào

SỬ DỤNG DATA WAREHOUSE



KHÔNG SỬ DỤNG DATA WAREHOUSE



Khi chưa có *data warehouse*, doanh nghiệp phải đổi mới với lượng lớn dữ liệu chi tiết nhưng lộn xộn và không có cấu trúc, khiến end-user rất khó trích xuất thông tin hữu ích. *Data warehouse* thu thập tất cả dữ liệu đó và sắp xếp gọn gàng trong một cơ sở dữ liệu quan hệ khổng lồ, qua quy trình Extract-Transform-Load (ETL). Điều này hỗ trợ giảm tải cho các hệ thống giao dịch, cải thiện chất lượng dữ liệu và chuẩn bị dữ liệu cho quá trình phân tích. Đồng thời, đảm bảo khi trích xuất hoặc thêm dữ liệu, các hệ thống vận hành không bị ảnh hưởng.

Khi thực sự dùng dữ liệu đó để tiến hành phân tích với BI, vấn đề mới thực sự bắt đầu. Không có công cụ phân tích nào đủ mạnh để chạy các truy vấn, sàng lọc từng mảnh dữ liệu có trong cơ sở dữ liệu. Để làm được điều đó, cần một hệ thống phần cứng cực kỳ mạnh mẽ, mà chi phí sẽ rất đắt đỏ!

Để giải quyết vấn đề này, các doanh nghiệp bắt đầu sử dụng “*data marts*”. Các *data marts* cho phép các nhóm người khác nhau quyền truy cập vào một phần nhỏ của toàn bộ kho dữ liệu. Tuy điều này giảm áp lực lên máy tính cá nhân khi chạy một truy vấn, nhưng cũng có nghĩa là không một ai có được cái nhìn tổng quan về dữ liệu mà công ty thu thập. Điều này không phải là một kết quả lý tưởng.

Tuy nhiên, BI truyền thống không có cách nào giải quyết vấn đề này. Nó không thể truy cập vào dữ liệu thô trong kho dữ liệu, hoặc chọn lọc những gì nó muốn trong cơ sở dữ liệu. Kết quả, các giải pháp BI cũ đã sử dụng các *data marts* hoặc các khối OLAP để kết nối với dữ liệu. Các khối OLAP rút ngắn thời gian truy vấn bằng cách tóm tắt trước một số yếu tố của dữ liệu trước khi bắt đầu tìm kiếm trong cơ sở dữ liệu.

Cả hai lựa chọn đều nhanh hơn so với làm việc cùng tập dữ liệu phức tạp và khổng lồ, nhưng mất đi sự tổng quát hoặc chi tiết trong dữ liệu, cũng như làm phân tán dữ liệu... Vì thế cũng mất đi mục đích ban đầu của việc có một kho dữ liệu hoàn chỉnh.

Hạn chế của *Data Warehouse* là gì?

DATA WAREHOUSE là phương án tuyệt vời để làm sạch và lưu trữ dữ liệu, hay sắp xếp dữ liệu gọn gàng để dễ dàng tìm kiếm. Đây cũng là mục đích chính của data warehouse.

Quan trọng là, cần thực sự hiểu mục đích của *data warehouse*. Chúng không có chức năng phân tích, và cần có một công cụ phân tích bổ sung. Điểm mạnh của *data warehouse* là sắp xếp dữ liệu, và đây cũng chính là điểm yếu của nó. Về bản chất, nó là các cơ sở dữ liệu quan hệ, khiến nó rất "kén chọn" trong loại dữ liệu để lưu trữ và phương thức lưu trữ.

Trong phân tích dữ liệu, cách sắp xếp và số lượng cột của từng danh mục đều được cố định. Vì vậy, nếu dữ liệu nhập vào không đúng cấu trúc sẽ gây ra khó khăn lớn - Ví dụ như nội dung ảnh hoặc video, phân tích ngôn ngữ, vv. Đồng thời, *data warehouse* cũng chỉ lưu trữ dữ liệu quá khứ, hay dữ liệu phân tích. Không thể lưu trữ dữ liệu thời gian thực (real-time). Ví dụ như dữ liệu về thông tin giao dịch.

Sự hạn chế này ảnh hưởng đến loại thông tin (insight) có thể trích xuất từ dữ liệu trong kho, ngay cả khi đã kết nối với một công cụ phân tích. Tùy vào nguồn dữ liệu, để có được bức tranh tổng thể và chính xác, cần phải lấy dữ liệu trực tiếp từ các ứng dụng hoặc các cơ sở dữ liệu khác nhau và đưa vào nền tảng Business Intelligence (BI) để phân tích.

Ngoài ra, quy trình này cần có sự hỗ trợ từ đội ngũ kỹ thuật (IT). Để theo kịp xu hướng và áp dụng cách tiếp cận mới trong tương tác với khách hàng, thu thập thông tin về doanh nghiệp và quy trình, cần thường xuyên thêm luồng dữ liệu và loại dữ liệu mới vào data warehouse. Nếu không, bạn sẽ phải làm việc với rất nhiều nền tảng, hệ thống cùng lúc.

Việc xây dựng *data warehouse* chắc chắn đã "ngốn" của doanh nghiệp rất nhiều ngân sách và thời gian.Thêm vào đó, để chạy "quy trình" một cách trơn tru cũng không hề đơn giản - Để duy trì, chắc chắn cần sự hỗ trợ từ bộ phận IT. Toàn bộ quá trình có thể dẫn tới trì hoãn truy cập và phân tích dữ liệu có sẵn. Lúc này, *data warehouse* không những không giúp đẩy nhanh tốc độ, mà lại khiến mọi thứ trì trệ hơn.

Để khắc phục, doanh nghiệp tiếp tục phải đổ thêm tiền mua thêm không gian *data warehouse* mới có thể tải thêm dữ liệu từ các nguồn mới (mà ban đầu khi xây kho dữ liệu chưa có).

Ví dụ: Một bên cung ứng dịch vụ sẽ lưu trữ thông tin thanh toán định kỳ. Khi có thêm một dịch vụ mới, dữ liệu mới phải chờ lập trình viên xử lý trước khi tải lên *data warehouse*. Cho đến lúc đó, không thể tiến hành phân tích, quản lý cũng không nắm được liệu dịch vụ mới có đem lại hiệu quả hay không.

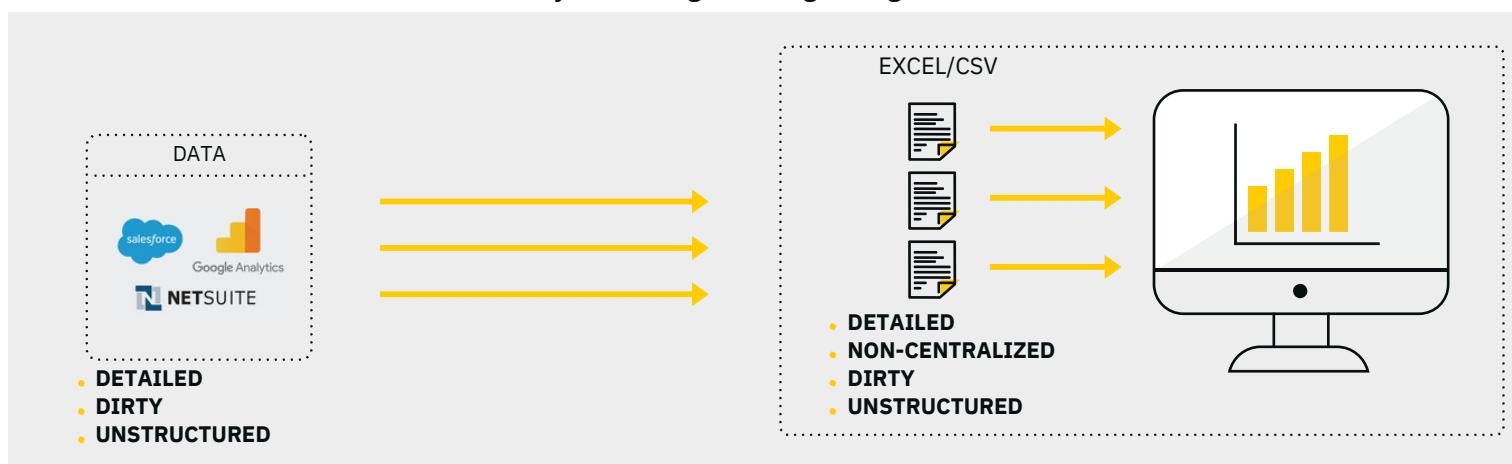
Data Warehouse có dành cho tất cả mọi người?

Với sự giới hạn không gian lưu trữ của *data warehouse*, bạn sẽ phải sử dụng thêm các luồng dữ liệu khác trong quá trình phân tích. Tuy nhiên, nếu có một hệ thống Business Intelligence (BI) đủ tinh vi để kết nối trực tiếp với các nguồn dữ liệu và đảm bảo truy cập vào dữ liệu sạch và hoàn chỉnh... thì có thể sẽ không cần tới *data warehouse* nữa...

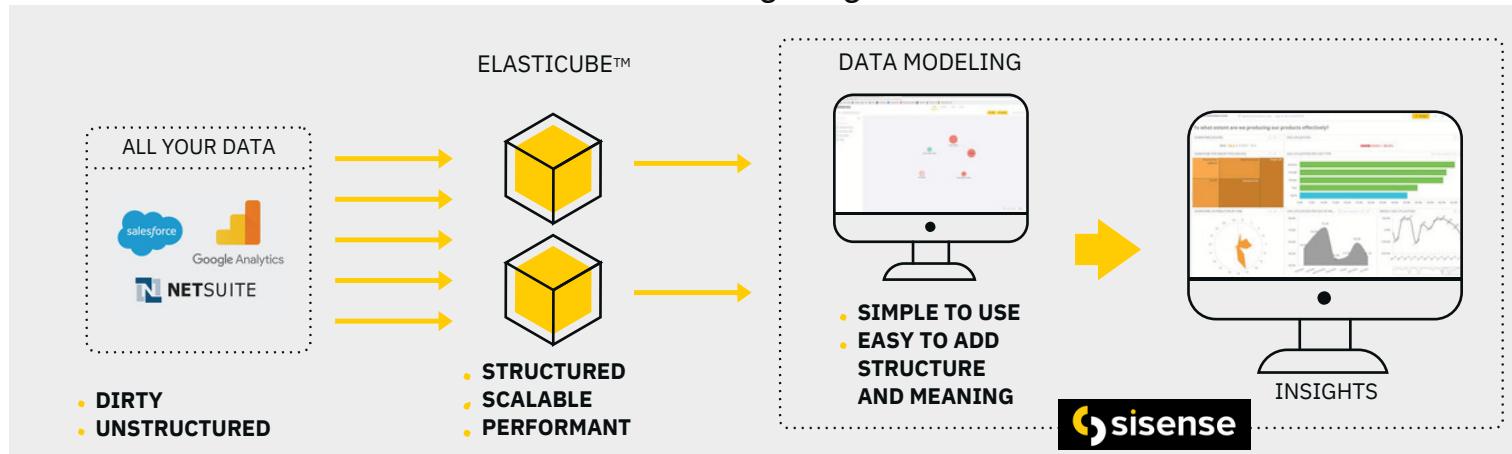
Ví dụ, nền tảng Sisense xây dựng ElastiCubes - Cơ sở dữ liệu phân tích độc đáo, hiệu suất cao với tốc độ lưu trữ siêu nhanh. ElastiCubes sử dụng phương pháp cơ sở dữ liệu cột trong bộ nhớ:

- a) Chỉ tải những phần dữ liệu cần thiết trong cơ sở dữ liệu vào nền tảng BI để phân tích, thay vì tải toàn bộ cùng lúc.
- b) Tối ưu hóa việc sử dụng ổ đĩa, RAM và CPU của máy tính. Một máy chủ thông dụng có thể xử lý cả terabytes dữ liệu, ngay cả khi có nhiều người dùng truy vấn cùng một lúc.

BI Truyền thống (Không dùng DW)



Sisense (Không dùng DW)

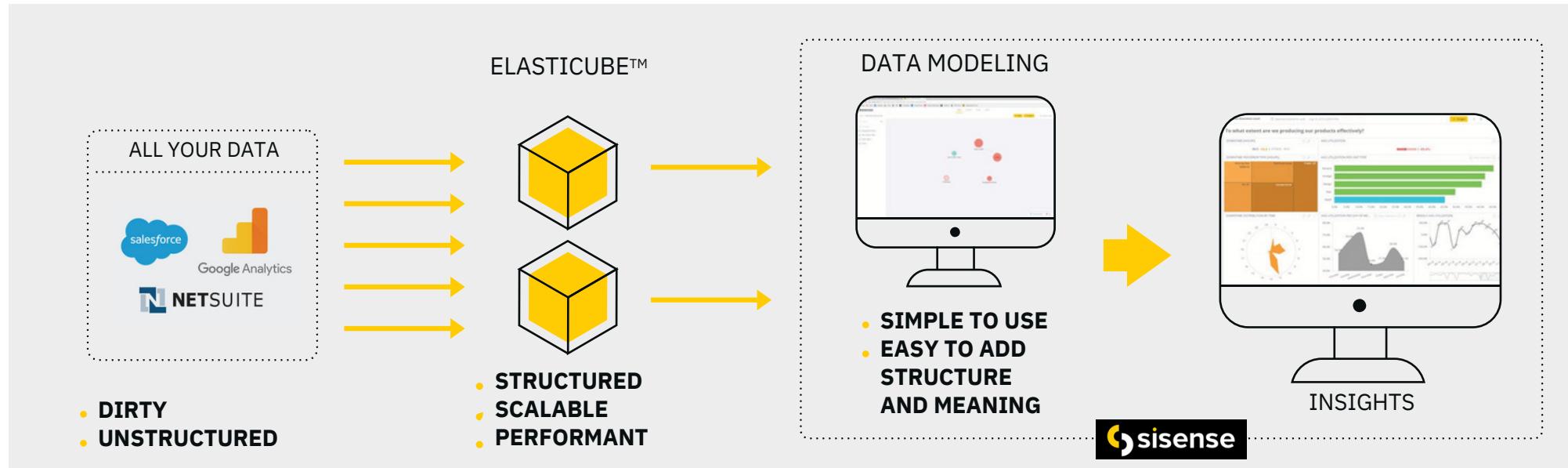


Sisense giúp làm sạch, tập trung và cấu trúc dữ liệu chi tiết, thực hiện các chức năng ETL trong máy chủ ElastiCube. Sau đó, hiển thị dữ liệu bằng bảng biểu trực quan, giúp người dùng BI tối ưu hóa việc sử dụng và khai thác dữ liệu. Bảng biểu cũng có thể dễ dàng chia sẻ trong nội bộ công ty.

Nhờ vậy, bạn có thể truy cập tất cả dữ liệu để phục vụ bất kỳ nhu cầu phân tích cụ thể nào, dữ liệu cũng được định dạng vô cùng gọn gàng và ngắn nắp. Không còn cần đầu tư nhiều vào phần cứng hoặc phụ thuộc vào bộ phận IT nữa. Có thể nói, với Sisense, *data warehouse* (kho dữ liệu) đã trở nên thừa thãi.

sisense

không dùng data warehouse



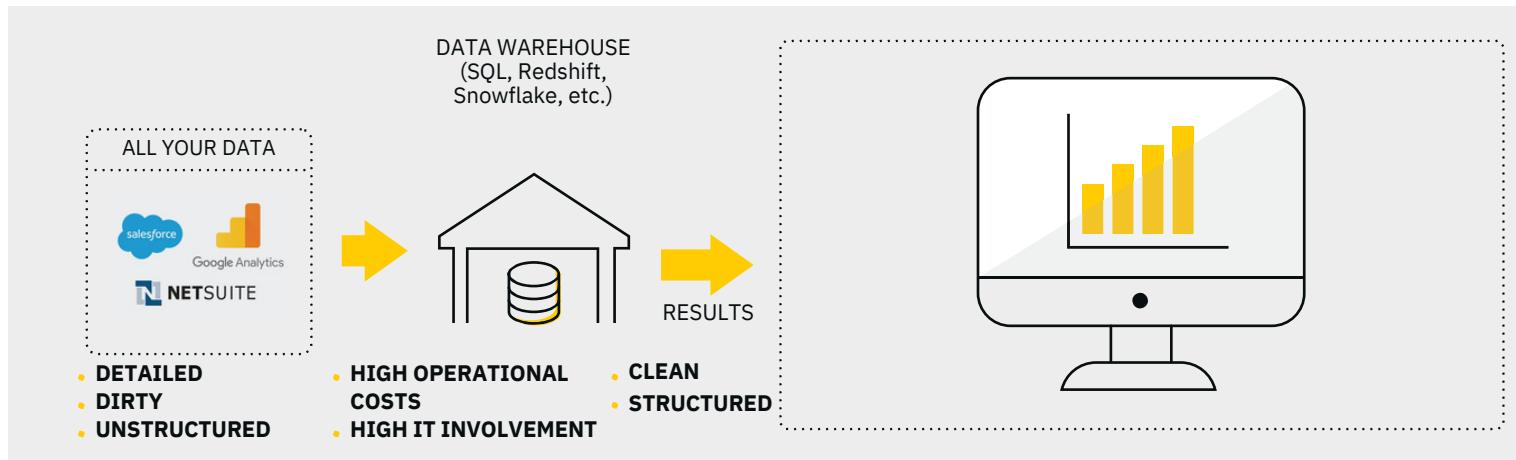
Khoan... nếu mọi dữ liệu đều có sẵn ở trong kho thì sao?

Nếu bạn đã có một kho dữ liệu, công sức bạn bỏ ra cũng không phải vô ích. Thực tế, nó cũng có thể là một tín hiệu tốt.

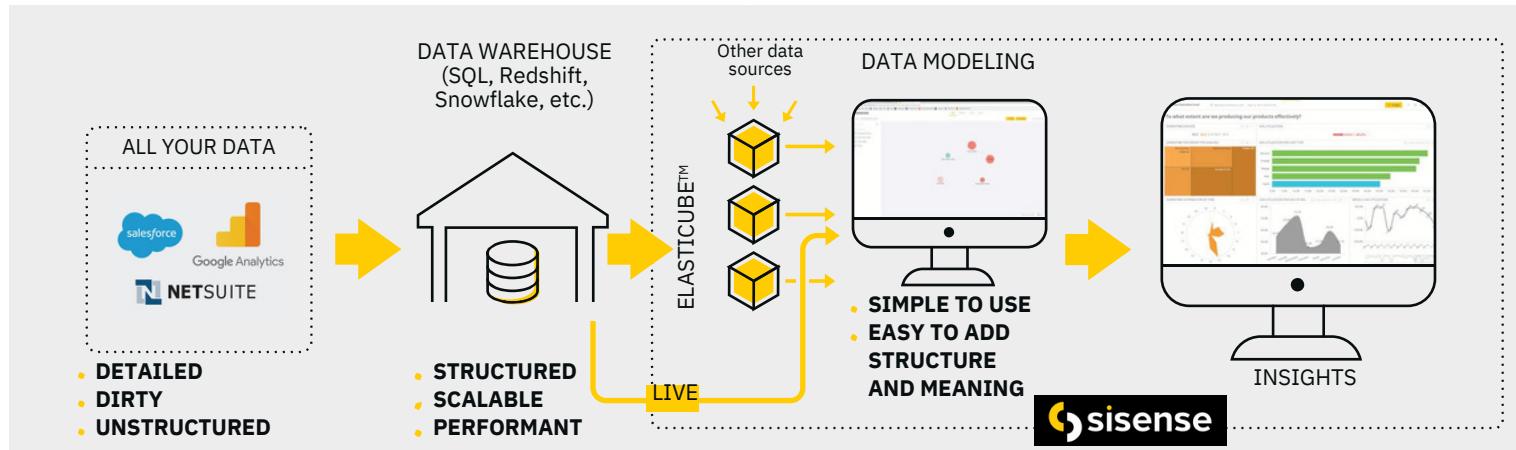
Nếu đang dùng một kho dữ liệu, không có lý gì để ngắt kết nối nó với hệ thống Business Intelligence (BI). Trên thực tế, giải pháp BI đúng có thể tận dụng tối đa kho dữ liệu hiện có, tối ưu hóa đầu tư của bạn.

Ví dụ, với Sisense, bạn có thể kết nối ElastiCubes với dữ liệu gốc trong kho dữ liệu, giữ lại từng chi tiết nhỏ của dữ liệu đã được làm sạch, tập trung và cấu trúc chuẩn. So sánh dữ liệu đó với các dữ liệu gốc sử dụng **data mart** hoặc **OLAP cubes** để kết nối, sau đó tóm tắt và phân phối. Công nghệ BI tiên tiến không đào thải cơ sở dữ liệu cũ (trừ khi bạn muốn thế). Ngược lại, chúng tận dụng tối đa hệ thống sẵn có.

BI SỬ DỤNG DATA WAREHOUSE



Sisense sử dụng data warehouse

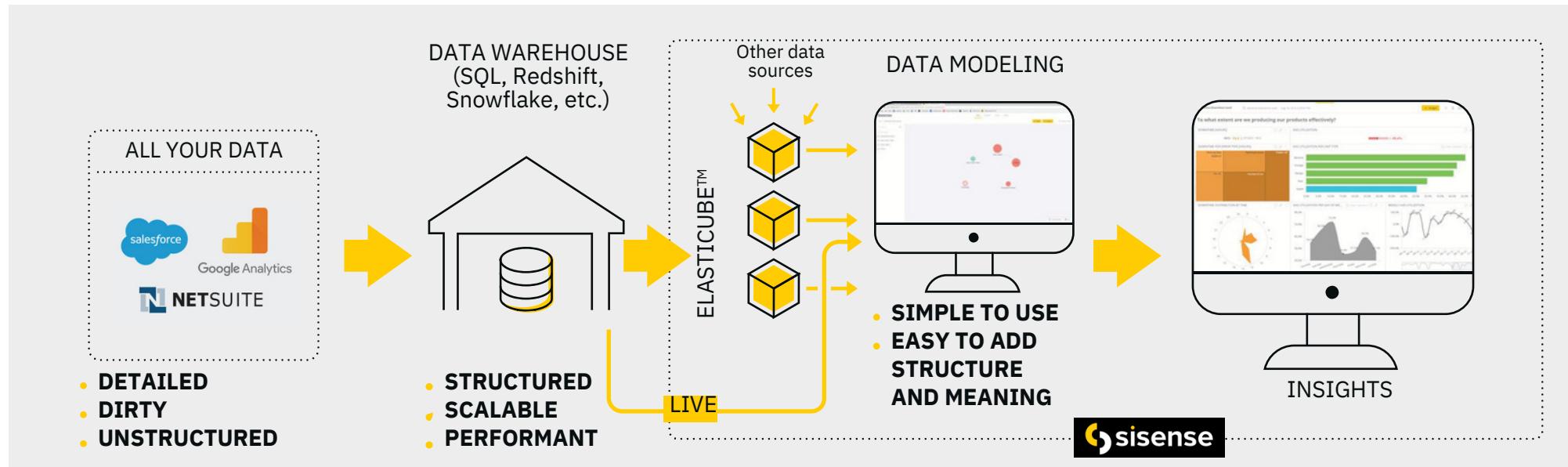


Sisense giải quyết mọi vấn đề về lưu trữ và truy vấn, đồng thời đảm bảo tính ổn định cho chiến lược dữ liệu trong tương lai. Không cần thêm luồng dữ liệu mới vào cơ sở dữ liệu vì bạn hoàn toàn có thể kết nối trực tiếp tới dữ liệu gốc.

Không còn phải lo lắng khi kho dữ liệu ngày càng lớn hơn, vì sẽ không có bất kỳ trở ngại nào trong việc kết nối dữ liệu từ kho và đưa vào hệ thống BI. Đồng thời, bạn có thể dễ dàng mở rộng quy mô dữ liệu của mình.

sisense

khi kết hợp với data warehouse



Kết Luận: Hãy Đầu Tư Thông Minh Hơn

Hiện nay, bước đầu trong triển khai BI cho bất kỳ tổ chức nào là xây dựng cơ sở hạ tầng để thu thập và lưu trữ dữ liệu. Thông thường quá trình này đòi hỏi đầu tư vào một kho dữ liệu, và ngân sách đầu tư sẽ tăng lên theo lượng dữ liệu cần lưu trữ. Dữ liệu trong kho sau đó sẽ được đưa vào nền tảng BI để phân tích. Có thể nói, khoản đầu tư vào kho dữ liệu và các công cụ để trích xuất và phân tích dữ liệu là rất lớn.

Trong thế giới BI hiện nay, vẫn có thể làm vậy - nhưng đó không phải là cách duy nhất. Bạn cũng có thể thiết lập kết nối trực tiếp, hoặc thậm chí kết nối trực tiếp với thời gian thực (real-time), đến nguồn dữ liệu gốc. Không nhất thiết **chỉ phụ thuộc** vào nguồn dữ liệu trong kho.

Có thể nói, *data warehouse* (kho dữ liệu) không còn là trung tâm trong quy trình BI, nhưng cũng không cần loại bỏ nó hoàn toàn. Tốt nhất là nên sử dụng một nền tảng BI vừa có thể kết nối dữ liệu từ kho dữ liệu, vừa hỗ trợ kết nối dữ liệu trực tiếp.

Sisense là một giải pháp BI giúp bạn giải quyết bài toán dữ liệu mà vẫn “vẹn cả đôi đường”. Không cần tái cấu trúc quy trình hiện tại, không gặp bất kỳ khó khăn nào trong quá trình, hoặc phải trả thêm tiền khi thêm nguồn dữ liệu mới. Có thể nói, nền tảng Sisense thực sự khiến dữ liệu trong kho trở nên dễ truy cập, hữu ích và có giá trị hơn bao giờ hết.

Trân
trọng
Cảm
Ơn



Liên hệ

(+84) 961 916 131

resolve.com.vn

resolve@resolve.com.vn

N04 Trang Tien, tầng 4, tòa M5, số 91
Nguyễn Chí Thanh, Phường Láng Hạ,
Quận Đống Đa, Thành phố Hà Nội.