Big Data

Nghề "hot" của thế kỉ 21

■ CHÂM ANH

Ngày nay, dữ liệu chính là tiền bạc của doanh nghiệp. Chỉ cần doanh nghiệp biết khai thác hiệu quả, Big Data (dữ liệu lớn) là công cụ không chỉ giúp tăng lợi nhuận cho chính họ mà còn giúp tiết kiệm thời gian cho khách hàng trong mua sắm. Báo cáo mới đây của Công ty Nghiên cứu thị trường IDC cho thấy, doanh thu đến từ thị trường Big Data đã tăng lên 16,9 tỷ USD vào năm 2015, tiếp tục tăng trưởng kép với tốc độ 27% và đạt đến 32,4 tỷ USD trong năm 2017 và còn tiếp tục tăng mạnh trong những năm tiếp theo.

NGUỒN TÀI NGUYÊN QUÝ HƠN VÀNG

Big Data (dữ liệu lớn) là thuật ngữ dùng để chỉ một tập hợp dữ liệu rất lớn và phức tạp đến nỗi những công cụ, ứng dụng xử lý dữ liệu truyền thống không thể đảm đương được. Thống kê cho thấy, trong hai năm qua, khối lượng dữ liệu trên toàn cầu đã chiếm đến 90% lượng dữ liệu số được tạo ra kể từ khi công nghệ số hóa ra đời. Kích cỡ của Big Data tăng lên từng ngày, tính đến năm 2012 đã lên hàng exabyte (1 exabyte bằng 1 tỷ gigabyte). Dữ liệu lớn thường bao gồm tập hợp dữ liệu với kích thước vượt xa khả năng của các công cụ phần mềm thông thường để thu thập, hiển thị, quản lý và xử lý dữ liệu trong một thời gian có thể chấp nhận được. Kích thước dữ liệu lớn là một mục tiêu liên tục thay đổi. Như năm 2012 thì phạm vi một vài tá terabytes tới nhiều petabytes dữ liệu. Dữ liệu lớn yêu cầu một tập các kỹ thuật và công nghệ được tích hợp theo hình thức mới để khai phá từ tập dữ liệu đa dạng, phức tạp, và có quy mô lớn.

Dữ liệu lớn bao gồm các thách thức như phân tích, thu thập, giám sát dữ liệu, tìm kiếm, chia sẻ, lưu trữ, truyền nhận, trực quan, truy vấn và tính riêng tư. Thuật ngữ này thường chỉ đơn giản đề cập đến việc sử dụng các phân tích dự báo, phân tích hành vi người dùng, hoặc một số phương pháp phân tích dữ liệu tiên tiến khác trích xuất giá trị từ dữ liệu mà ít khi đề cập đến kích thước của bộ dữ liệu.

Nguồn dữ liệu lớn đã tồn tại dưới nhiều hình thức, thường được xây dựng bởi các công ty cho những nhu cầu đặc biệt. Bắt đầu từ những năm 1990, các nhà cung cấp thương mại tham gia cung cấp các hệ thống quản lý cơ sở dữ liệu song song cho các dữ liệu lớn. Trong nhiều năm, WinterCorp là công ty phát hành báo cáo lớn nhất về cơ sở dữ liệu.

Năm 1984, Tập đoàn Teradata đưa ra thị trường hệ thống xử lý dữ liệu song song DBC 1012. Các hệ thống của Teradata là những hệ thống đầu tiên lưu trữ và phân tích đến 1 terabyte dữ liệu vào năm 1992. Ổ đĩa cứng đã đạt đến mức dung lượng 2.5GB vào năm 1991 nên định nghĩa dữ liệu lớn liên tục phát triển theo quy luật Kryder. Teradata đã cài đặt hệ thống đầu tiên dựa trên RDBMS có thể phân tích hàng petabytes dữ liệu vào năm 2007. Đến năm 2017, có hàng chục các cơ sở dữ liệu dựa trên hệ thống của Teradata có dung lượng hàng petabyte, trong đó dữ liệu lớn nhất vượt quá 50 petabytes. Cho đến năm 2008, 100% hệ thống đều xử lý các dữ liệu quan hệ có cấu trúc. Do đó, Teradata đã thêm các kiểu dữ liệu phi cấu trúc bao gồm XML, JSON và Avro.

Năm 2000, Seisint Inc. (nay là Tập đoàn LexisNexis) đã phát triển một khung chia sẻ tệp dựa trên cấu cấu trúc C++ để lưu trữ và truy vấn dữ liệu. Hệ thống này lưu trữ và phân phối dữ liệu có cấu trúc, bán cấu trúc, và phi cấu trúc trên nhiều máy chủ.

Người dùng có thể truy vấn bằng một phương ngữ C ++ gọi là ECL. ECL sử dụng phương thức "áp dụng giản đồ khi truy cập dữ liệu" để suy luận cấu trúc dữ liệu được lưu trữ khi nó được truy vấn, thay vì khi nó được lưu trữ. Năm 2004, LexisNexis mua lại Seisint Inc. và trong năm 2008 đã mua lại ChoicePoint, Inc. cùng với nền tảng xử lý song song tốc độ cao của họ. Hai nền tảng đã được sáp nhập vào hệ thống HPCC (High-Performance Computing Cluster) và HPCC có mã nguồn mở dựa trên giấy phép Apache v2.0 vào năm 2011. Khoảng cùng thời điểm đó, hệ thống Quantcast File đã được phát hành.

Năm 2004, Google xuất bản một bài báo về một quá trình gọi là MapReduce sử dụng một kiến trúc tương tự. MapReduce cung cấp một mô hình xử lý song song, và phát hành những ứng dụng liên quan để xử lý lượng dữ liệu khổng lồ. Với MapReduce, các truy vấn được chia nhỏ và truyền đi qua các nút mạng song song và được xử lý song song (bước Map). Các kết quả sau đó được thu thập và phân phối (Bước Reduce). Khuôn mẫu này rất thành công nên những công ty khác cũng muốn sao chép các thuật toán của nó. Do đó, Google đã triển khai khuôn mẫu MapReduce thông qua dự án mã nguồn mở Apache Hadoop.

Các nghiên cứu vào năm 2012 cho thấy cấu trúc nhiều lớp là một lựa chọn để giải quyết các vấn đề của xử lý dữ liệu lớn. Một kiến trúc phân tán song song phân tán dữ liệu trên nhiều máy chủ; những môi trường thực hiện song song này có thể cải thiện đáng kể tốc độ xử lý dữ liệu. Kiểu cấu trúc này chèn dữ liệu vào một DBMS song song, thực hiện việc sử dụng các khung nền MapReduce và Hadoop. Loại khung nền này sẽ tăng sức mạnh xử lý thông suốt đến người dùng cuối bằng cách sử dụng một máy chủ ứng dụng đầu cuối.

Phân tích dữ liệu lớn ứng dụng vào việc sản xuất được giới thiệu như một cấu trúc 5C (connection - kết nối, conversion - chuyển đổi, cyber - không gian mạng, cognition - nhận thức và configuration - cấu hình).

Hồ dữ liệu cho phép một tổ chức thay đổi định hướng từ mô hình kiểm soát tập trung sang mô hình chia sẻ thông tin để năng động đáp ứng với sự thay đổi của việc quản lý thông tin. Điều này cho phép phân tách nhanh chóng dữ liệu vào hồ dữ liệu, do đó làm giảm thời gian xử lý thông tin.

NHỮNG LỢI ÍCH KHỔNG LỒ

Phân tích tập dữ hợp liệu có thể tìm ra tương quan mới tới "xu hướng kinh doanh hiện tại, phòng bệnh tật, chống tội phạm… Các nhà khoa học, điều hành doanh nghiệp, y bác

sĩ, quảng cáo và các chính phủ cũng thường xuyên gặp những khó khăn với các tập hợp dữ liệu lớn trong các lĩnh vực bao gồm tìm kiếm internet, thông tin tài chính doanh nghiệp. Các nhà khoa học gặp giới hạn trong công việc cần tính toán rất lớn, bao gồm khí tượng học, bộ gen, mạng thần kinh, các mô phỏng vật lý phức tạp, sinh vật học và nghiên cứu môi trường.

Dữ liệu lớn có thể kết hợp với công nghệ Mạng lưới vạn vật kết nối Internet. Dữ liệu được chiết xuất từ các thiết bị IoT cung cấp một bản đồ kết nối giữa các thiết bị. Những sự kết nối này đã được ngành công nghiệp truyền thông, các công ty và chính phủ sử dụng để nhắm mục tiêu chính xác hơn đối tượng của họ và tăng hiệu quả của phương tiện truyền thông. IoT cũng ngày càng được chấp nhận như một phương tiện thu thập dữ liệu cảm giác, và dữ liệu cảm giác này đã được sử dụng trong các ngành như y học và sản xuất.

Kevin Ashton, chuyên gia đổi mới kỹ thuật số người được cho là người tạo ra thuật ngữ định nghĩa Internet vạn vật đã phát biểu: "Nếu chúng ta có máy tính biết tất cả mọi thứ - nó sẽ sử dụng dữ liệu mà nó thu thập được mà không có sự trợ giúp từ chúng ta - chúng ta sẽ có thể theo dõi và kiểm soát mọi thứ, giảm đáng kể lượng chất thải, tổn thất và chi phí. Chúng ta sẽ biết khi nào cần thay thế, sửa chữa hoặc thu hồi lại, và liệu rằng thức ăn chúng ta đang ăn có tươi hay không."

Việc sử dụng các dữ liệu lớn trong các quy trình của chính phủ cho phép tăng hiệu quả về mặt chi phí, năng suất và sự đổi mới, nhưng không phải là không có sai sót. Phân tích dữ liệu thường yêu cầu nhiều bộ phận của chính phủ (trung ương và địa phương) hợp tác và tạo ra các quy trình mới và sáng tạo để mang lại kết quả mong muốn. Chính phủ các nước có thể ứng dụng Big Data để dự đoán tỷ lệ thất nghiệp, xu hướng nghề nghiệp của tương lai để đầu tư cho những hạng mục phù hợp hoặc cắt giảm chi tiêu, kích thích tăng trưởng kinh tế, thậm chí dự đoán sự phát triển của mầm bệnh và khoanh vùng sự lây lan của bệnh dịch. Nói cách khác, Big Data sẽ là công cụ thúc đẩy sự phát triển kinh tế - xã hội trong tương lai.

Việc sử dụng các dữ liệu lớn dưới dạng lịch sử các giao dịch tài chính được gọi là phân tích kỹ thuật. Sử dụng dữ liệu phi tài chính để dự đoán thị trường đôi khi được gọi là dữ liệu thay thế.

Dữ liệu lớn cung cấp cơ sở hạ tầng cho ngành công nghiệp sản xuất, đó là khả năng cải thiện năng suất và tính khả dụng. Việc lên kế hoạch sản xuất chính là một cách tiếp cận dữ liệu lớn cho phép giảm thời gian

chết về gần như bằng không và cụ thể hóa số lượng lớn dữ liệu và các công cụ dự đoán khác cho phép tạo ra một quá trình nhằm hệ thống hóa dữ liệu thành các thông tin hữu ích. Khái niệm về việc dự báo sản xuất bắt đầu bằng việc thu thập dữ liệu cảm quan khác nhau như âm thanh, chuyển động, áp suất, điện áp... Số lượng lớn các dữ liệu cảm quan cộng với dữ liệu lịch sử sản xuất tạo thành dữ liệu lớn trong sản xuất.

Nhờ giải pháp Big Data, năm 2013, Amazon đạt doanh thu tới 74 tỷ USD, IBM đạt hơn 16 tỷ USD. Big Data là nhu cầu tăng trưởng lớn đến nỗi từ năm 2010, Software AG, Oracle, IBM, Microsoft, SAP, EMC, HP và Dell đã chi hơn 15 tỷ USD cho các công ty chuyên về quản lý và phân tích dữ liệu.

Các dữ liệu lớn này như là đầu vào cho các công cụ dự báo và các chiến lược phòng ngừa tương tự như việc dự báo trong lĩnh vực Quản lý Y tế. Phân tích dữ liệu lớn đã giúp cải thiện việc chăm sóc sức khoẻ bằng cách cá nhân hóa các phương pháp trị liệu và chẩn đoán lâm sàng, làm giảm thiểu chi phí và thời gian khám bệnh, tự động báo cáo và lưu trữ thông tin sức khỏe và dữ liệu bệnh nhân trong nội bộ cũng như mở rộng ra bên ngoài, chuẩn hóa các thuật ngữ y học và chống phân mảnh trong lưu trữ dữ liệu và thông tin của bệnh. Một số lĩnh vực có sự cải tiến mang tính hướng dẫn hơn là thực hành. Lượng dữ liệu được tạo ra trong các hệ thống chăm sóc sức khoẻ là không nhỏ. Với sự bổ sung thêm của mHealth, eHealth và các thiết bị công nghệ theo dõi sức khỏe được thì khối lượng dữ liệu sẽ tiếp tục gia tăng. Điều này bao gồm dữ liệu ghi chép sức khoẻ điện tử, dữ liệu hình ảnh, dữ liệu được tạo ra của bệnh nhân, dữ liệu cảm biến và các dạng dữ liệu khó xử lý khác. Hiện nay, nhu cầu lớn hơn đối với các môi trường như vậy là chú ý nhiều hơn đến chất lượng dữ liệu và thông tin. "Dữ liệu lớn rất thường có nghĩa là dữ liệu chưa được xử lý và một phần số liệu không chính xác tăng lên khi có sự tăng trưởng khối lượng dữ liệu." Việc theo dõi bằng con người ở quy mô dữ liệu lớn là không thể và có một nhu cầu cấp thiết về các công cụ thông minh để kiểm soát chính xác và xử lý thông tin bị mất trong dịch vụ y tế. Mặc dù dữ liệu trong lĩnh vực chăm sóc sức khoẻ hiện nay thường được lưu trữ dưới dạng điện tử, nhưng nó nằm ngoài phạm vi của dữ liệu lớn vì hầu hết không có cấu trúc và khó sử dụng.

Một nghiên cứu của Viện nghiên cứu toàn cầu McKinsey cho thấy, ngành dữ liệu lớn đang thiếu hụt 1,5 triệu chuyên gia cũng như nhà quản lý dữ liệu, và một số trường đại học bao gồm Đại học Tennessee và UC Berkeley đã tạo ra các chương trình thạc sĩ để đáp ứng nhu cầu này. Các khóa huấn luyện tư nhân cũng phát triển các chương trình để đáp ứng nhu cầu đó, bao gồm các chương trình miễn phí như The Data Incubator hoặc chương trình trả tiền như General Assembly.

"CƠN KHÁT" NHÂN LỰC BIG DATA

Theo ước tính của Gartner, một công ty nghiên cứu và tư vấn về công nghệ hàng đầu của Mỹ, năm 2015 Big Data tạo thêm 4,4 triệu việc làm

trong ngành IT toàn cầu và trong 5 năm (2012-2017), thế giới đầu tư 232 tỷ USD cho Big Data.

Có thể nói chưa bao giờ các doanh nghiệp lại "khát" nhân lực về "Big Data" như hiện nay do ngày càng có nhiều công ty nhận ra được lợi ích to lớn từ việc khai thác và phân tích dữ liệu đối với hoạt động kinh doanh của họ. Khi ngày càng nhiều doanh nghiệp nhận ra được lợi ích từ việc phân tích xử lý dữ liệu, Big Data sẽ là một trong những lĩnh vực khát nhân lực nhất trong thời gian tới.

Big data là công nghệ thu thập thông tin quy mô lớn từ các website. Các doanh nghiệp thường vận dụng công cụ này nhằm phục vụ công việc dự đoán xu hướng thị trường, nâng cao chất lượng sản phẩm hoặc dịch vụ hiện có, tạo ra sản phẩm mới hoặc tìm hiểu về hành vi khách hàng

Phân tích dữ liệu cũng có thể giúp các doanh nghiệp thích nghi, tạo ra nội dung website thu hút nhiều khách hàng hơn, có được cái nhìn sâu sắc vào hành vi mua hàng. Dữ liệu càng nhiều thì càng tốt cho công ty. Để làm được như vậy, doanh nghiệp nên cung cấp nội dung trên nhiều nền tảng social media, nhằm thu thập được nhiều thông tin từ những điểm tiếp xúc với khách hàng.

Bằng cách tìm hiểu qua hệ thống cơ sở dữ liệu, công ty có thể tạo ra nội dung có liên quan hơn với người đọc. Chính ý tưởng này đã giúp Craig Rayner – Giám đốc tuyển dụng hãng SEO.io thu hút nhân tài. Nhờ vào việc phân tích và tổng hợp những dữ liệu nội bộ phòng nhân sự, ông đã tạo ra những quảng cáo tuyển dụng hấp dẫn đối với người tìm việc.

Đảo qua thị trường việc làm, sẽ không khó để bạn nhìn ra những mức lương hậu hĩnh cùng hàng tá những phụ cấp hấp dẫn khác cho công việc như "data scientist" (tạm dịch: chuyên gia dữ liệu) hay "data analyst" (phân tích dữ liệu).

Big Data là từ khoá được tìm kiếm nhiều nhất trên mạng xã hội LinkedIn, và tất nhiên những ai đang đi đầu trong lĩnh vực này hẳn sẽ được các headhunter (chuyên gia săn đầu người) săn đón thường xuyên. Do nhu cầu tăng vọt là vậy, nên nếu bạn là người có đầu óc phân tích và khả năng xử lý dữ liệu, việc bước chân vào ngành này sớm bao nhiêu thì cơ hội thăng tiến của bạn. Tin tốt là trong vài năm qua xuất hiện nhiều chương trình đào tạo (dưới nhiều hình thức khác nhau) đáp ứng nhu cầu chuyên gia về dữ liệu lớn. Tuy nhiên, các đơn vị đào tạo hầu hết là các trường đại học ở nước ngoài, ví dụ như các chương trình khoa học dữ liệu chuyên sâu hay phân tích nâng cao tại Viện Nghiên cứu Khoa học và Kỹ thuật Dữ liệu Columbia, Đại học Berkeley, Đại học Carnegie Mellon, Viện Công nghệ Illinois, Đại học Imperial, Đại học North Carolina, Đại học Syracuse và Đại học Tennessee…