BIG DATA VÀ ỨNG DỤNG TRONG BẢO MẬT THÔNG TIN

ThS Lò Thị Phương Nhung, ThS Nguyễn Mai Phương
Viện Thông tin Khoa học, Học viện Chính trị quốc gia Hồ Chí Minh

Big data đã và đang là một trong những vấn đề trung tâm, nhận được nhiều sự quan tâm trong cuộc Cách mạng công nghiệp (CMCN) 4.0. Big data chính là cốt lõi để sử dụng, phát triển internet vạn vật (IoT) và trí tuệ nhân tạo (AI). Theo dự báo, CMCN 4.0 sẽ tạo ra một lượng lớn dữ liệu (đến năm 2020, lượng dữ liệu sẽ tăng gấp 50 lần hiện nay) [1]. Thông qua thu thập, phân tích và xử lý lượng dữ liệu lớn này sẽ tạo ra những tri thức mới, hỗ trợ tích cực trong quản lý, sản xuất kinh doanh và nhiều lĩnh vực của đời sống xã hội, trong đó có bảo mật thông tin.

Tổng quan về big data

Hiện nay, có nhiều quan điểm khác nhau về khái niệm big data - "dữ liệu lớn". Theo Viện Nghiên cứu toàn cầu McKinsey (Mỹ), big data được hiểu là 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. Nhìn từ góc độ giá trị của dữ liệu, có ý kiến cho rằng, big data không chỉ có nghĩa là dung lượng lớn mà còn có nghĩa thông qua việc tích hợp, phân tích và xử lý đối với những dữ liệu này, con người có thể phát hiện được tri thức mới và thu được giá trị mới, từ đó mang đến cho con người tri thức, lợi nhuận và sự phát triển lớn hơn. Để có thể hiểu thêm khái niệm big data, chúng ta cần thấy được các thuộc tính kỹ thuật và thuộc tính xã hội của nó.

Về thuộc tính kỹ thuật

Dung lượng lớn. Trong xã hội thông tin hiện nay, mỗi người đều là chủ thể tạo ra dữ liệu. Qua các công cụ khác nhau như tin nhắn, mạng xã hội, mạng mua sắm điện tử, truyền hình... những hành vi thường ngày trong công việc và cuộc sống của mỗi cá nhân đều có thể trở thành nguồn dữ liệu.

Trong xã hội thông tin hiện nay, mỗi người đều là chủ thể tạo ra dữ liệu.

Thiết bị di động ngày càng rẻ và nhiều, anten, nhật ký phần mềm, các thiết bị thu hình, thu thanh, đầu đọc RFID, mạng cảm biến không dây… đều góp phần đắc lực cho quá trình tạo ra dữ liệu.

Tính đa dạng. Tính đa dạng của dữ liệu lớn thể hiện ở các phương diện: đa dạng về loại (dữ liệu kết cấu và dữ liệu phi kết cấu); đa dạng về nguồn gốc (tổ chức và cá nhân trong xã hội đều là chủ thể tạo ra nguồn dữ liệu); nội dung dữ liệu (tất cả các lĩnh vực, các khía cạnh của đời sống xã hội).

Tốc độ nhanh. Một đặc trưng nổi bật của xã hội thông tin là tính phức tạp và tính không xác định ở mức độ cao. Tốc độ nhanh của big data không chỉ thể hiện ở việc dữ liệu được tạo ra một cách nhanh chóng mà còn thể hiện ở tốc độ xử lý thông tin nhanh. Thời đại big data đòi hỏi phương thức vận hành của dữ liệu cần chuyển từ trạng thái dữ liệu động và tĩnh sang trạng thái dữ liệu đang sử dụng nhằm đạt được mục đích xử lý thông tin nhanh chóng.

Sự tồn tại đan xen giữa dữ liệu có giá trị cao và dữ liệu có giá trị thấp. Chỉ những dữ liệu đã được phân tích, xử lý và chọn lọc thì mới là những dữ liệu có giá trị thật sự.

Trong thời đại big data, bên cạnh dữ liệu có giá trị thì cũng tồn tại những dữ liệu ít có ý nghĩa đối với chúng ta.

Về thuộc tính xã hội

Thứ nhất, big data là một loại năng lực và kỹ thuật. Ưu thế của thời đại big data chính là ở chỗ con người có thể tiến hành phân tích, lưu trữ và sử dụng nguồn dữ liệu khổng lồ mà kỹ thuật truyền thông không thể thực hiện được. Thông qua việc phân tích đối với nguồn dữ liệu khổng lồ, con người không chỉ tận dụng được giá trị tiềm năng của dữ liệu mà còn sử dụng nó vào việc đổi mới và sáng tạo.

Thứ hai, big data là kết cấu hạ tầng. Trong xã hội nông nghiệp, đất đai và thủy lợi là những hạ tầng chủ yếu; trong xã hội công nghiệp, năng lượng, đường bộ, đường sắt, hàng không... là những kết cấu hạ tầng chủ yếu. Trong bối cảnh của CMCN 4.0, thông tin trở thành nguồn lực chiến lược quan trọng; điện toán đám mây (cloud computing), trung tâm dữ liệu, mạng di động tốc độ cao... sẽ trở thành kết cấu hạ tầng quan trọng. Việc xây dựng kết cấu hạ tầng này vừa cần vai trò quy hoạch và đầu tư của nhà nước, vừa cần sự tham gia và đầu tư của doanh nghiệp.

Thứ ba, big data là nguồn lực cốt lõi. Các loại nguồn lực vật chất truyền thống như đất đai, năng lượng... đều là những nguồn lực khan hiếm, người này sử dụng sẽ ảnh hưởng đến việc sử dụng của người khác. Nhưng đặc tính của nguồn thông tin là ở chỗ, việc người này sử dụng và tiêu dùng không ảnh hưởng đến việc sử dụng của người khác, không hề làm giảm đi mà còn làm tăng thêm giá trị của nó. Quan trọng hơn là, quá trình tiêu dùng thông tin cũng đồng thời là quá trình tạo ra thông tin mới, người sử dụng và tiêu dùng thông tin càng nhiều, lượng thông tin được tạo ra sẽ càng lớn. Có thể nói, trong thời đại big data, kỹ thuật và công nghệ liên quan đến big data trở thành nguồn lực cốt lõi quan trọng nhất của một quốc gia [2].

Thứ tư, big data là một phương thức tư duy. Big data không chỉ là trạng thái dữ liệu lớn, một loạt kỹ thuật thông tin tiên tiến mà còn là một quan niệm và phương pháp liên ngành trong nhận thức và cải tạo thế giới. Nó tạo điều kiện để thực hiện một xã hội mở với mức độ cao; nhấn mạnh việc chia sẻ và tương tác về mặt thông tin. Chính điều này góp phần vào việc đổi mới quan niệm, phương pháp nhận thức của con người đối với thế giới. Big data làm cho tư duy của con người trở nên biện chứng hơn, giúp nhìn nhận vấn đề và sự việc một cách đa chiều hơn.

Thứ năm, big data là "một thời đại". Thời đại big data lấy dữ liệu làm nguyên tố cơ bản, làm nguồn lực chiến lược, chỉ cần nắm được dữ liệu thì sẽ có được năng lực cạnh tranh cốt lõi. Trong thời đại big data, mỗi một cá thể đều là "nguồn" của dữ liệu, thông qua những phương thức khác nhau mỗi cá thể đều có thể thể hiện tiếng nói của mình. Thời đại dữ liệu lớn cũng có nghĩa là thời đại xã hội mở, một thời đại mà quyền lực trở nên phân tán hơn, đời sống xã hội trở nên tự do và dân chủ hơn [3].

Ứng dụng trong bảo mật thông tin

Các nhà cung cấp các giải pháp an toàn thông tin cho các doanh nghiệp nhỏ và vừa đều xem big data là yếu tố không thể tách rời với các kết quả phân tích nguy cơ và rủi ro của hệ thống, đặc biệt các hệ thống thông tin thương mại điện tử với nguồn dữ liệu khổng lồ từ các phương tiện truyền thông xã hội. Đây là nhóm dữ liệu thường được dùng để phân tích, xác định hoặc dự báo về quan điểm, mối quan tâm, tình cảm của khách hàng về các sản phẩm và dịch vụ mà họ đã sử dụng. Ví dụ như hệ thống dữ liệu thu thập từ nhật ký máy chủ (Log Server), hệ thống dữ liệu từ các sự kiện nhấp chuột trên các website, hệ thống dữ liệu thu thập từ các máy cảm biến (Sensors), hệ thống dữ liệu thu thập từ hệ thống thông tin địa lý (GIS)… Việc bảo đảm an toàn cho hệ thống big data được các doanh nghiệp nhỏ và vừa rất quan tâm, bởi dữ liệu càng lớn càng là mục tiêu tấn công của các tội phạm công nghệ cao.

Sự xuất hiện của big data với những công nghệ ứng dụng mới cũng giúp mở rộng quy mô của các hệ thống dữ liệu để sử dụng một tập hợp các nguồn tài nguyên phân tán với các bộ vi xử lý nhanh hơn và lưu trữ nhiều dữ liệu hơn, giúp tận dụng được tất cả các nguồn dữ liệu sẵn có, để cung cấp các phân tích tốt hơn và nhanh hơn đối với việc phát hiện tấn công và phản ứng các sự cố. Big data sẽ chuyển đổi phân tích an toàn thông tin bằng cách thu thập dữ liệu ở một quy mô lớn từ nhiều nguồn (các bản ghi nhật ký hệ thống đến các cơ sở dữ liệu về lỗ hổng bảo mật, dữ liệu về tấn công mạng, dữ liệu mã độc…), sau đó được sử dụng với các ứng dụng chính như:

Một là, theo dõi và phát hiện Botnet. Botnet hiện đang là một trong những mối đe dọa lớn và là một thách thức đối với các chuyên gia an toàn thông tin. Việc phát hiện Botnet đòi hỏi phải thu thập một lượng lớn dữ liệu mạng để phân tích. Với việc ứng dụng big data, dự án nghiên cứu Botcloud do nhóm của Jer'ome Fraçois và đồng nghiệp tại Đại học Luxembourg thực hiện đã sử dụng mô hình MapReduce để phân tích một lượng lớn các dữ liệu Netflow để xác định các máy tính

Phân tích một lượng lớn các dữ liệu Netflow để xác định các máy tính bị lây nhiễm đang tham gia trong một mạng Botnet.

bị lây nhiễm đang tham gia trong một mạng Botnet. Dự án này đã mở ra nhiều hướng mới trong việc xây dựng các hệ thống thông minh để phát hiện Botnet. MapReduce được sử dụng cho dự án này, vì một lượng lớn các dữ liệu Netflow được thu thập cần phải phân tích. 720 triệu bản ghi Netflow (77 GB) được thu thập chỉ trong 23 giờ đồng hồ. BotCloud được xây dựng dựa trên kiến trúc BotTrack. Kiến trúc này được thiết kế để theo dõi và phát hiện Botnet bằng việc sử dụng Netflow và thuật toán PageRan, thực hiện việc theo dõi các kênh C&C (command - and - control) trong Botnet [4].

Hai là, ứng dụng big data trong phát hiện tấn công APT*. Tấn công APT thường do những đối tượng có trình độ chuyên môn cao thực hiện, được hậu thuẫn bởi những tổ chức có tiềm lực. Một thách thức trong việc dò tìm các cuộc tấn công APT là việc lọc toàn bộ số lượng dữ liệu nhằm phát hiện những bất thường đang xảy ra. Vì thế phân tích big data là một tiếp cận phù hợp trong việc dò tìm các cuộc tấn công APT. Tại Phòng thí nghiệm RSA (chuyên nghiên cứu để giải quyết các vấn đề về an ninh mạng cấp bách trên thế giới), một hệ thống dò tìm tấn công APT được nghiên cứu có tên là Beehive. Các kết quả nghiên cứu ban đầu cho thấy, Beehive cung cấp khả năng để xử lý khoảng 1 tỷ các thông điệp bản ghi sự kiện trong một giờ và nhận diện các hành động vi phạm chính sách, cũng như sự lây nhiễm phần mềm độc hại.

Ba là, ứng dụng big data trong phát hiện tấn công Zero-day. Tại Symantec, bằng việc ứng dụng big data, các kỹ sư an ninh mạng đã đưa ra một nền tảng WINE (Worldwide Intelligence Network Enviornment) cho việc tiến hành phân tích dữ liệu, trên phạm vi rộng, sử dụng các dữ liệu thu thập được. Nền tảng WINE đã từng được sử dụng để đo thời gian của 18 cuộc tấn công Zero-day bằng cách kết hợp các hệ nhị phân tin cậy và tập dữ liệu về các dấu hiệu ngăn chặn mã độc từ xa, tiến hành phân tích các trường dữ liệu được thu thập trên 11 triệu máy chủ (host) trên khắp thế giới, các tấn công này kéo dài từ 19 ngày đến 30 tháng. Hơn nữa, 60% các lỗ hổng bảo mật được xác định trong nghiên cứu này đã bị khai thác trong các tấn công Zero-day mà trước đó chưa tìm ra [5]…

Ngoài ra, Chương trình PRISM của cơ quan tình báo Mỹ đã ứng dụng công nghệ big data để thu thập lượng thông tin khổng lồ trên khắp thế giới (dữ liệu điện thoại, email, hình ảnh, video, trạng thái trên các trang mạng xã hội…), từ đó phân tích và đưa ra các cảnh báo về các dấu hiệu khủng bố có thể xảy ra. Với sự trợ giúp của những gã khổng lồ công nghệ thông tin như Microsoft, Yahoo, Google, Facebook, PalTalk, AOL, Skype, YouTube, và Apple cùng với việc hợp tác với FBI để thu thập dữ liệu điện thoại (cuộc gọi, tin nhắn, danh bạ...) thì cơ quan này đang giám sát dữ liệu thông qua PRISM [6]

*APT là tên viết tắt của Advanced Persistent Threat - thuật ngữ rộng dùng để mô tả một chiến dịch tấn công, thường do một nhóm sử dụng những kỹ thuật tấn công nâng cao để có thể hiện diện và tồn tại lâu dài trên mạng Internet nhằm khai thác dữ liệu có độ nhạy cảm cao. Mục tiêu chính của những vụ tấn công này thường được lựa chọn và nghiên cứu cẩn thận. Chúng thường bao gồm các doanh nghiệp lớn, các cơ quan chính phủ. Thực hiện tấn công APT đòi hỏi nhiều tài nguyên hơn tấn công ứng dụng web bình thường. Những kẻ phạm tội thường là những nhóm tội phạm mạng có kinh nghiệm và có hỗ trợ tài chính rất lớn. Một số cuộc tấn công APT còn được chính phủ tài trợ và được sử dụng làm vũ khí chiến tranh mạng, phục vụ công tác tình báo.

TÀI LIỆU THAM KHẢO

[1] Marek Obitko, *Industry 4.0 and big data*, http://www.stech.cz/Portals/0/Konference/2015/03%20Industry-/PDF/03_obitko.pdf.

[2] Li Shuqing, Jiao Fusen, Zhang Yong, Xu Xia (2019), "Problems and changes in digital libraries in the age of big data from the perspective of user services", *Journal of Academic Librarianship*, **45(1)**, pp.22-30.

[3] Astrid Mager (2019), "The politics of big data. Big data, big brother?", *Information, Communication & Society*, **22(10)**, pp.1523-1525.

[4] Alguliyev Rasim, Imamverdiyev Yadigar (2014), "Big Data: Big Promises for Information Security", *Conference Proceedings*, Publisher: IEEE.

[5] N. Miloslavskaya, A. Makhmudova (2016), "Survey of Big Data Information Security", *Conference Proceedings*, Publisher: IEEE.

[6] United States National Security Agency (2013), *PRISM Collection Manager*.