

KHÁI QUÁT VỀ BIGDATA TRONG THỐNG KÊ NHÀ NƯỚC

Cao Văn Hoạch

Vụ Phương pháp chế độ thống kê và Công nghệ thông tin, Tổng cục Thống kê

Bigdata là những nguồn dữ liệu có dung lượng lớn, có cấu trúc hoặc phi cấu trúc, đa dạng và có tốc độ tăng lên nhanh chóng, đòi hỏi có phương thức đặc biệt để khai thác. Bigdata được hình thành từ nhiều nguồn dữ liệu khác nhau, nhưng chủ yếu từ nguồn dữ liệu hành chính; dữ liệu giao dịch; dữ liệu cảm biến, như ảnh vệ tinh, cảm biến giao thông, cảm biến khí hậu...; dữ liệu lưu vết thiết bị điện tử, như lưu vết điện thoại di động, định vị toàn cầu (GPS)...; dữ liệu hành vi, như tìm kiếm trực tuyến... Bigdata ngày càng có vai trò quan trọng trong thống kê Nhà nước nhằm đáp ứng tính kịp thời của số liệu thống kê, Bigdata có thể được sử dụng để sản xuất số liệu thống kê Nhà nước theo những cách khác nhau: (i) thay thế hoàn toàn các nguồn đầu vào; (ii) thay thế một phần các nguồn đầu vào; và (iii) cung cấp các số liệu thống kê hoàn toàn mới mà có thể bổ sung cho các thông tin thống kê có sẵn...

Ủy ban Kinh tế châu Âu của Liên hợp quốc (UNECE) đã thành lập Ban cấp cao thực hiện việc hiện đại hóa công tác thống kê. Ban này đã thiết lập nhóm công tác để nghiên cứu, ứng dụng Bigdata trong thống kê Nhà nước gồm các chuyên gia từ các quốc gia và các tổ chức quốc tế do Ban Thư ký UNECE làm đầu mối. Nhóm công tác đã nghiên cứu và đưa ra dự thảo phân loại Bigdata trong thống kê Nhà nước như:

(1) *Mạng xã hội (nguồn thông tin về con người)*: thông tin này là các ghi chép cá nhân của mỗi người về kinh nghiệm, về công việc, tâm sự, hình ảnh,..., trước đây được lưu giữ trong sổ, sách.... hiện nay phần lớn được số hóa và lưu trữ ở khắp mọi nơi từ các máy tính cá nhân đến lưu trữ đám mây với các mạng xã hội. Dữ liệu loại này có cấu trúc lỏng lẻo và thường không được quản trị.

1100. Mạng xã hội: Facebook, Twitter, Tumblr, etc.

1200. Bài viết và các lời bình luận

1300. Các tài liệu cá nhân

1400. Hình ảnh: Instagram, Flickr, Picasa,

1500. Videos: Youtube,....

1600. Tra cứu internet

1700. Dữ liệu điện thoại: Tin nhắn, ...

1800. Bản đồ do người dùng tự thiết lập

1900. Thư điện tử

(2) *Hệ thống kinh doanh truyền thống (dữ liệu xử lý trung gian)*: các ghi chép và theo dõi các sự kiện kinh doanh như đăng ký khách hàng, sản phẩm sản xuất, đơn

đặt hàng... Dữ liệu kinh doanh truyền thống phần lớn được quản lý và xử lý bằng tin học, thường có cấu trúc và được lưu trữ trong các hệ thống cơ sở dữ liệu quan hệ (một số có thể là "dữ liệu hành chính").

21. Dữ liệu của các tổ chức công

2110. Bệnh án

22. Dữ liệu của các cơ sở kinh doanh

2210. Giao dịch thương mại

2220. Hồ sơ ngân hàng, chứng khoán

2230. Thương mại điện tử

2240. Thẻ tín dụng

(3) *Dữ liệu thiết bị kết nối Internet (dữ liệu máy móc tạo ra)*: Bắt nguồn từ sự bùng nổ của các thiết bị cảm biến, các máy quan sát trái đất, các máy móc thiết bị kết nối internet khác, dữ liệu do chúng tạo ra (từ các ghi chép cảm biến đơn giản đến các bản ghi máy tính phức tạp), dữ liệu máy móc tạo ra. Khối lượng dữ liệu này đang tăng lên nhanh chóng, trở thành thành phần quan trọng của việc lưu trữ và xử lý. Cấu trúc của dữ liệu này phù hợp cho việc xử lý máy tính, nhưng kích thước và tốc độ của nó vượt ra ngoài các phương pháp xử lý truyền thống.

31. Dữ liệu từ các thiết bị cảm biến

311. Dữ liệu từ thiết bị cảm biến cố định

3111. Dữ liệu từ thiết bị tự động hóa trong gia đình

3112. Dữ liệu từ Cảm biến thời tiết, môi trường

3113. Dữ liệu từ Cảm biến quan sát giao thông

3114. Dữ liệu từ Cảm biến nghiên cứu khoa học

3115. Dữ liệu từ hình ảnh, video giám sát an ninh

312. Dữ liệu cảm biến điện thoại di động (theo dõi lộ trình-tracking)

3121. Vị trí điện thoại di động

3122. Ô tô

3123. Ảnh vệ tinh

32. Dữ liệu từ hệ thống máy tính

3210. Nhật ký máy tính

3220. Nhật ký các trang mạng internet

Trên thế giới hiện nay có nhiều dự án cũng như nhiều nước ứng dụng Bigdata trong sản xuất số liệu thống kê, tuy nhiên vẫn gặp phải một số khó khăn như: chưa có căn cứ pháp lý cho việc sử dụng Bigdata (nhất là đối với vấn đề thông tin riêng tư, cá nhân); chưa có căn cứ khẳng định mức độ chính xác, trung thực của nhiều thông tin khai thác từ Bigdata. Đồng thời chưa có phương pháp tính mức độ đại diện cũng như sai số chọn mẫu của mẫu chọn từ nguồn Bigdata; chưa có phương pháp chuẩn mực thống nhất của việc khai thác Bigdata, đồng thời chưa có mô hình hệ thống CNTT

phục vụ xử lý khối dữ liệu đặc biệt này. Sử dụng Bigdata sẽ đòi hỏi phương pháp lưu trữ và xử lý phải thoát ra khỏi những phương pháp truyền thống hiện có. Bên cạnh đó, do tính rất không đồng nhất về định dạng, cấu trúc, nội dung,... của nguồn Bigdata, nên phương pháp sản xuất thông tin thống kê cần được áp dụng phù hợp cho từng trường hợp riêng biệt. Do đó mô hình Quy trình sản xuất thông tin thống kê (GSBPM) có thể không áp dụng được. Về lâu dài cần nghiên cứu GSBPM trong điều kiện khai thác Bigdata đòi hỏi cán bộ thực hiện phải có cả kỹ năng thống kê và kỹ năng về CNTT.

Để ứng dụng Bigdata ở Việt Nam trước mắt cần làm thí điểm ở phạm vi hẹp để lấy kinh nghiệm, tập trung vào đo lường sự thay đổi, thay vì khai thác giá trị tuyệt đối. Điều này có thể đánh giá được độ tin cậy của số liệu khai thác từ Bigdata; và bước đầu thử nghiệm với các chỉ tiêu đơn giản, ngắn hạn; trong khi đó vẫn thực hiện song song với phương pháp truyền thống để đối chứng. Chúng ta cần khảo sát kỹ các đặc thù của nguồn Bigdata sẽ sử dụng để có biện pháp xử lý, khai thác phù hợp với điều kiện cho phép. Hiện nay, chúng ta có thể sử dụng dữ liệu tờ khai nộp thuế, tờ khai hàng hóa dịch vụ XNK để tổng hợp một số chỉ tiêu liên quan; sử dụng dữ liệu mua bán trực tuyến để thu thập giá của một số nhóm mặt hàng trong rõ hàng hóa tính chỉ số giá tiêu dùng; sử dụng dữ liệu ảnh vệ tinh để thống kê diện tích tự nhiên, diện tích gieo trồng và sản phẩm trồng trọt; sử dụng ảnh vệ tinh để thống kê các công trình xây dựng; sử dụng dữ liệu mạng xã hội để nghiên cứu quan điểm, tư tưởng của các thế hệ nhân dân; sử dụng dữ liệu điện thoại di động để thống kê về di cư; và sử dụng dữ liệu truy cập internet để thống kê tình hình sử dụng internet, hiện trạng thương mại điện tử, giao dịch trực tuyến,... để thí nghiệm.

Để có thể khai thác và sử dụng Bigdata, chúng ta phải xây dựng phương pháp luận thống kê mới trong điều kiện sử dụng các nguồn Bigdata và phương pháp tiếp cận và phân tích dữ liệu có khối lượng khổng lồ; xây dựng cơ sở pháp lý cho việc sử dụng Bigdata trong thống kê Nhà nước; cập nhật và xây mới các bảng Phân loại thống kê phù hợp với công tác thống kê sử dụng Bigdata; xây dựng hạ tầng CNTT phù hợp và phát triển các phần mềm chuyên biệt phục vụ khai thác Bigdata; xây dựng quan hệ hợp tác chặt chẽ giữa Cơ quan thống kê và các cơ quan, đơn vị quản lý nguồn Bigdata; và phải đào tạo kỹ năng thiết yếu để khai thác Bigdata về CNTT cũng như về thống kê.

Tóm lại, để khai thác, sử dụng thông tin từ nguồn dữ liệu khổng lồ Bigdata cho thống kê Nhà nước, cần đầu tư nghiên cứu để xác định: phương pháp tiếp cận, mô hình hệ thống xử lý, các lý thuyết về thống kê liên quan và đào tạo đội ngũ cán bộ thực hiện đáp ứng yêu cầu. Trước đó, quan điểm về thống kê phải thoát ra khỏi những lý thuyết của thống kê truyền thống.