MỘT KHÍA CẠNH CẦN LƯU Ý VỀ PHÁT TRIỂN ỨNG DỤNG TIN HỌC TRONG NGÀNH THỐNG KÊ XÂY DỰNG CÁC KHO DỮ LIỆU THỐNG KÊ

Kỹ sư Hoàng Minh Thiện

Hiện nay hầu hết các Cục Thống kê tỉnh, thành phố đã và đang áp dụng các công nghệ tin học vào công tác nghiệp vụ. Điều đó đặt ra cho các Cục một đòi hỏi cấp bách là xây dựng các kho dữ liệu.

Ở đây chúng tôi chỉ tạm gọi là kho dữ liệu chứ không gọi như tên kỹ thuật là cơ sở dữ liệu nhằm hàm ý việc hình thành và phát triển kho dữ liệu hiện nay là tự phát và không bài bản lắm; hay có thể nói nó chưa đủ điều kiện để gọi là các cơ sở dữ liệu. Song đây lại là thực tế hiện nay, khi các cán bộ nghiệp vụ sau một thời gian ứng dụng rời rạc các sản phẩm mềm ứng dụng đã tự nhiên hình thành các kho dữ liệu thống kê về lĩnh vực nghiệp vụ của riêng mình. Ban đầu các cán bộ nghiệp vụ ứng dụng các công cụ mềm cho một việc cụ thể, họ lưu lại số liệu cũng như các tính toán dẫn xuất từ đó; trong lần ứng dụng sau đó họ cũng làm như vậy và có thể họ đã sử dụng lại số liệu đã nhập hay tính toán trước đó,... cứ như vậy thì tới một lúc họ sẽ có hàng loạt tệp số liệu được lưu trữ và cả các dữ liệu khác như danh mục, hoặc lời văn. Ít lâu sau việc sử dụng lại số liệu hay dữ liệu đã có một số vấn đề nảy sinh, thông thường cán bộ nghiệp vụ thường gặp các vấn đề sau:

- Làm sao tìm thấy dữ liệu cần tìm nhanh chóng.

- Làm sao lấy một phần số liệu cũ hoà vào số liệu mới mà không ảnh hưởng tới tệp đã lưu.

- Làm sao loại bỏ những số liệu không còn cần tới mà không ảnh hưởng tới những gì cần giữ lại.

- Làm sao cất trữ dữ liệu đã có gọn gàng và lâu dài hơn vì có thể chúng đã làm cho chật cả ổ đĩa cứng.

- Làm sao dữ liệu đã lưu được an toàn trong khi làm việc và đặc biệt khi có nhiều người cùng làm việc trên một máy tính.

- ....

Những điểm cơ bản trên là những vấn đề thường xuyên các cán bộ nghiệp vụ gặp phải trước kho dữ liệu tự hình thành của họ trong quá trình ứng dụng máy tính. Thực chất của những công việc này là sử dụng số liệu, tìm kiếm, lưu trữ, hiệu chỉnh, loại bỏ và an toàn dữ liệu... đây cũng là những chức năng chính phải có của một cơ sở dữ liệu.

Do phần lớn các kho dữ liệu hình thành tự phát như đã nói ở trên và do những người tạo ra là không chuyên cho nên khi dữ liệu nhiều lên làm cho người sử dụng mất rất nhiều thời gian trong tìm kiếm, hiệu chỉnh lại dữ liệu. Một số cách làm đơn giản mà người sử dụng không chuyên tin học hiện nay thường dùng để đối phó trước thực tế đó là:

- Tạo ra các thư mục khác nhau để việc lưu trữ trở nên có cấu trúc và trên cấu trúc đó có thể dễ dàng tìm thấy dữ liệu thông qua con đường dẫn đến dữ liệu. Cách thức này làm cho việc lưu trữ cần

phải thận trọng và tuân thủ theo cấu trúc thông tin trên đĩa cứng do hệ điều hành quản lý. Song nó cũng không giúp giải quyết mọi vấn đề nảy sinh.

- Cách thứ hai là người sử dụng lập một quy tắc riêng cho mình trong đặt tên cho các tệp dữ liệu, cùng với một sổ ghi chép các tên tệp theo hướng gợi nhớ tới nội dung dữ liệu đã có. Nên khi tìm có thể dễ dàng tìm thấy. Cách này cũng gây ra lích kích cho người ứng dụng và cũng chỉ gói trong việc kiểm soát tệp mà thôi.

- Cách thứ ba cũng thường được sử dụng là lưu trữ ngoài trên các đĩa mềm có ghi nhãn rõ ràng để nhận biết dữ liệu. Khi sử dụng thì đọc dữ liệu từ đĩa mềm. Cách này vừa để an toàn dữ liệu vừa để bảo mật. Tuy nhiên dung lượng nhớ của đĩa mềm nhỏ, độ tin cậy của thông tin lưu trữ không lâu và không cao nên phải thường xuyên cập nhật, cũng có khó khăn.

Mấy cách trên chắc chắn nhiều cán bộ nghiệp vụ đã và đang sử dụng tuy nhiên các khó khăn không cải thiện được bao nhiêu, đặc biệt khi kho dữ liệu đã khá lớn và người sử dụng không có tính kiên nhẫn; thậm chí có khi "Tìm cho ra dữ liệu còn lâu hơn cả tính toán!" hay "có khi nhập mới lại nhanh hơn!". Vì như vậy và vì không thể bỏ đi các dữ liệu đã có sau bao nhiêu năm làm việc chúng ta tất nhiên là phải tiến tới một bước quan trọng là ***tổ chức lại kho dữ liệu*** đã có. Đây là việc mà các cán bộ nghiệp vụ không chuyên tin học ít có kinh nghiệm cho nên lời khuyên của chúng tôi là hãy chọn lúc thật rảnh việc, thận trọng tiến hành và có thể sẽ thất bại trong những thử nghiệm đầu tiên, cho nên cần lưu trữ an toàn trước và hãy làm thử trên các bản sao.

Tổ chức lại kho dữ liệu. Thực tế việc tổ chức lại kho dữ liệu là làm cho kho dữ liệu hoàn thiện dần tới chuẩn mực như một cơ sở dữ liệu để đáp ứng tốt hơn yêu cầu khai thác của người ứng dụng trong thực tế; vì thế một số thao tác họ có thể làm trên kho dữ liệu của mình theo hướng đó là:

- Gộp lại. Trong thực tế các tệp dữ liệu của một cán bộ nghiệp vụ thống kê thường có cùng dạng vì thế chúng ta có thể ghép nhiều tệp lại thành một tệp như thế sẽ thuận lợi hơn khi tìm kiếm và sử dụng chúng. Tuy nhiên cũng có khi không thể gộp đơn thuần vì các cài đặt tính toán (hoặc các tập lệnh) đã có sẵn (thường trên các dạng bảng tính như EXEL, LOTUS...) và như thế có thể cần phải tách các tệp gốc đó ra thành thuần nhất trước khi gộp chúng lại. Để làm như thế trên rất nhiều tệp thì cần thận trọng suy tính những gì giữ lại những gì bỏ đi trong các tệp gộp tránh sự lặp dữ liệu hay thừa ra những dữ liệu không cần thiết. Việc này cũng có khi dẫn tới những tệp quá lớn vượt quá cho phép của bộ nhớ như thế là chúng ta sẽ thất bại.

- Bổ sung. Vấn đề thường gặp hơn lại là việc phân biệt số liệu giữa các tệp ghép lại khi sử dụng; vì thế đôi khi cần bổ sung thêm các chỉ tiêu trước khi gộp. Ví dụ chúng ta có một loạt tệp theo thời gian, mà trước nay ta vẫn lấy thời hạn của dữ liệu đặt làm tên tệp; nên khi ghép các tệp lại sẽ cần có thêm chỉ tiêu là thời hạn của số liệu để phân biệt giữa các thời kỳ. Như ta bổ sung năm, quý hay tháng... để khi tính toán sau này có thể dựa vào đó để chọn lọc dữ liệu cho các tính toán khác nhau.

- Xác định dữ liệu cở sở. Một việc nữa mà cán bộ nghiệp cần lưu ý là xác định rõ những dữ liệu căn bản nhất không thể thiếu và các dữ liệu khác có thể đều tính được từ đó; đó là những dữ liệu bắt buộc phải giữ lại và không cần giữ lại những dữ liệu do tính toán mà có. Khi quyết định

như vậy cũng đồng nghĩa người sử dụng phải từ bỏ các cấu trúc bảng tính đã hình thành, bỏ các tính toán, các đoạn lệnh...đã có và sẵn sàng cho việc làm mới các công cụ tiện ích thay cho công cụ đã dùng bấy lâu nay. Nếu tiến tới bước này thì kho dữ liệu đã rất gần với tên gọi cơ sở dữ liệu.

- Xây dựng quy trình tổng hợp mới. Sau khi tiến hành các việc loại bỏ, bổ sung, tách, ghép hay gộp các tệp số liệu nhằm thu gọn kho dữ liệu thì người sử dụng sẽ có ít hơn các tệp cần quản lý; nhưng vấn đề quan trọng hơn là người sử dụng sẽ phải xây dựng một quy trình tính toán tổng hợp mới để đảm bảo nhiệm vụ được giao. Trong nhiều trường hợp người cán bộ nghiệp vụ nên tham khảo các chuyên gia tin học để việc tổ chức lại kho dữ liệu an toàn và thành công.

Những việc làm trên có thể phần nào giúp cho người sử dụng tránh được một số khó khăn, song giải pháp toàn diện hơn vẫn phải có sự đầu tư của ngành cho những khoá bồi dưỡng chi tiết hay trợ giúp cụ thể đối với từng lĩnh vực mới hy vọng có được cả hệ thống cơ sở dữ liệu ngành hoàn chỉnh. Trong phạm vi ngành cùng với việc phát triển ứng dụng mạng làm tăng lên khả năng trao đổi dữ liệu trong nội bộ ngành, chúng tôi cho rằng cần xem xét kỹ vấn đề này để các nhà quản lý có thể vạch ra các cấu trúc hợp lý cho hệ thống các kho dữ liệu toàn ngành, tránh trùng lắp dữ liệu cũng như sự chia rẽ giữa các kho dữ liệu chuyên ngành.