Tích hợp dữ liệu và nhiệm vụ xây dựng hệ thống cơ sở dữ liệu Quốc gia về biến đổi khí hậu

KS. NGUYỄN HỮU CHÍNH
Cục Công nghệ Thông tin - Bộ TN&MT

Nhằm từng bước triển khai tốt các văn bản QPPL đã ban hành, và Chương trình mục tiêu quốc gia ứng phó với BĐKH, Bộ TN&MT đã triển khai đề tài "Nghiên cứu cơ sở khoa học, công nghệ xây dựng hệ thống CSDL Quốc gia về BĐKH và tác động của BĐKH phục vụ ứng phó với BĐKH", trong đó có nội dung "Nghiên cứu khung kiến trúc hệ thống CSDL Quốc gia về BĐKH và tác động của BĐKH phục vụ ứng phó với BĐKH". Trong khuôn khổ của một bài báo, tác giả sẽ đưa ra giải pháp tích hợp dữ liệu trong khung kiến trúc hệ thống CSDL Quốc gia đã đề xuất.

Tổng quan về tích hợp dữ liệu

Vấn đề tích hợp dữ liệu

Khi nói đến tích hợp dữ liệu, ta thường hình dung ngay đến việc tổng hợp thông tin từ nhiều nguồn dữ liệu sẵn có của các hệ thống khác nhau, hay trên cùng một hệ thống thành phần một nguồn dữ liệu mới có thể dùng cho một hệ thống mới nào đó hay chỉ để lưu trữ.

Vậy, các công việc trong việc tích hợp dữ liệu là gì? chúng được thực hiện như thế nào? trình tự ra sao? Ta sẽ lần lượt đi vào từng vấn đề, bắt đầu từ khái niệm về tích hợp dữ liệu.

Khái niệm tích hợp dữ liệu

Tích hợp dữ liệu là một khái niệm khá trừu tượng, thậm chí là hơi mơ hồ khiến nhiều người không thể định nghĩa chính xác và cụ thể. Thông thường, tích hợp dữ liệu được hiểu là quá trình kết hợp dữ liệu từ nhiều nguồn thông tin khác nhau nhằm cung cấp cho người dùng một cái nhìn tổng quan và duy nhất về các dữ liệu này.

Đặc điểm của hệ thống tích hợp dữ liệu

Các nguồn dữ liệu là phân tán. Các nguồn dữ liệu này có thể là cơ sở dữ liệu (CSDL) trong các hệ thống khác nhau, cũng có thể là các trang web ở các địa chỉ khác nhau, hoặc cũng có thể là những con người với các quan điểm khác nhau về một vấn đề nào đó.

Các nguồn dữ liệu là không đồng nhất. Sự không đồng nhất này thể hiện ở các ngôn ngữ biểu diễn và từ vựng biểu diễn dữ liệu. Các nguồn dữ liệu có thể có ngôn ngữ biểu diễn khác nhau. Ví dụ, CSDL của một nguồn được biểu diễn theo dạng XML, nhưng một nguồn dữ liệu khác lại được biểu diễn theo CSDL quan hệ. Các nguồn dữ liệu cũng có thể sử dụng các từ vựng khác nhau để cùng biểu diễn một dữ liệu.

Một hệ tích hợp dữ liệu thường không cần toàn bộ thông tin dữ liệu trong các nguồn cần tích hợp. Với mỗi nhiệm vụ cụ thể, hệ thống chỉ cần những dữ liệu liên quan đến việc thực hiện nhiệm vụ đó. Như vậy, nếu tập hợp toàn bộ các nguồn dữ liệu và hệ thống trước khi tích hợp thì sẽ rất lãng phí và nhiều khi không thể thực hiện được.

Với những đặc điểm nêu trên, việc xây dựng hệ thống tích hợp dữ liệu yêu cầu kiến thức về nhiều lĩnh vực khác nhau như lý thuyết về CSDL, các phương pháp ước lượng, lý thuyết về ngôn ngữ và biểu diễn thông tin... Do vậy, khi xây dựng hệ thống tích hợp dữ liệu cho lĩnh vực nào, cần phải tìm hiểu sâu về lĩnh vực đó, và cần thiết phải có sự tham gia của các chuyên gia trong lĩnh vực vào dự án.

Các mức độ tích hợp dữ liệu

Theo Khaled Bashir Shaban, tích hợp dữ liệu được chia thành ba mức dựa vào đặc điểm đầu vào và đầu ra của quá trình tích hợp.

Mức I: Tích hợp dữ liệu. Đây là mức thấp nhất. Trong mức này, đầu vào là các bản ghi dữ liệu, đầu ra cũng có dạng các bản ghi dữ liệu hoặc một dạng cao hơn nhưng vẫn đóng vai trò là dữ liệu cung cấp cho một ứng dụng nào đó.

Mức II: Tích hợp thông tin. Trong mức này, cả

đầu vào và đầu ra của quá trình tích hợp đều là thông tin, tức là một cấu trúc đầy đủ, tập hợp từ các bản ghi dữ liệu. Mức này xảy ra với các hệ thống nhiều nguồn dữ liệu mà cấu trúc của các nguồn dữ liệu này là khác nhau và mỗi nguồn thông tin không thể tách ra từ một nguồn khác.

Mức III: Tích hợp quyết định. Đây là mức tích hợp thông tin dữ liệu cao nhất. Đầu vào của hệ thống này có thể là thông tin, dữ liệu, hoặc các quyết định từ các hệ thống khác. Nhiệm vụ của hệ thống tích hợp dữ liệu ở mức này là phải đưa ra tập quyết định phục vụ yêu cầu đặt ra của hệ thống. Có thể nói, tích hợp quyết định phục vụ yêu cầu đặt ra của hệ thống, tích hợp quyết định ở mức trừu tượng cao hơn hai mức trước, do đó nó bao hàm cả hai mức trên. Một điểm khác nữa, nếu như ở mức 1 và mức 2 vẫn có những trường hợp quá trình tích hợp thông tin dữ liệu không thực hiện được (do không thoả mãn các điều kiện nào đó) thì ở mức 3 sẽ luôn được thực hiện vì nó không phụ thuộc vào bản chất và đặc điểm của các nguồn dữ liệu. Tuy chia ra làm ba mức như trên nhưng thực tế một hệ thống tích hợp dữ liệu thường có đủ ba mức. Các mức thấp sẽ làm cơ sở cho các mức cao hơn.

Phương pháp tích hợp dữ liệu

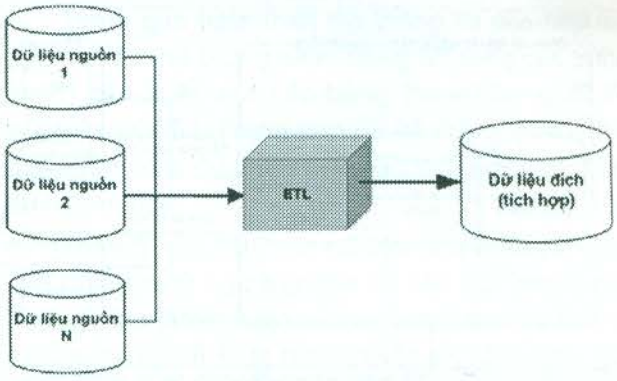
Có nhiều phương pháp tích hợp dữ liệu khác nhau, mỗi phương pháp sẽ phù hợp với một dạng hệ thống và các nguồn dữ liệu cụ thể.

Có thể kể ra một số phương pháp tích hợp dữ liệu như: Tích hợp dữ liệu dựa trên ước lượng không chắc chắn, Tích hợp dữ liệu dựa trên các ràng buộc dữ liệu, Tích hợp dữ liệu tự động dựa vào ontology,... Tuy nhiên, trong khuôn khổ của một bài báo, tác giả chỉ trình bày phương pháp Tích hợp dữ liệu dựa trên các ràng buộc dữ liệu, và đây cũng là phương pháp ứng dụng vào bài toán tích hợp dữ liệu trong nhiệm vụ Xây dựng CSDL Quốc gia về BĐKH.

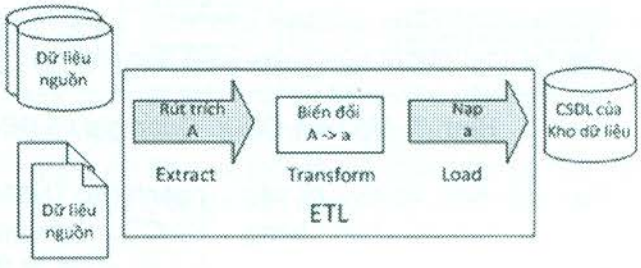
Tích hợp dữ liệu dựa trên các ràng buộc dữ liệu:

Phương pháp này dựa trên các ràng buộc dữ liệu. Các phương pháp thuộc về dạng này được áp dụng cho hệ thống bao gồm các nguồn dữ liệu biểu diễn dưới dạng CSDL và cấu trúc, ràng buộc trong CSDL này là có thể biết trước. Mục đích của các hệ thống này là trả lời các truy vấn của người dùng về thông tin dữ liệu trong nhiều nguồn khác nhau mà không cần truy nhập trực tiếp vào tất cả các nguồn thông tin này. Tiêu biểu cho phương pháp tích hợp dữ liệu thuộc loại này là phương pháp dùng cho hệ thống IBIS.

Phương pháp tích hợp dữ liệu được đưa ra dựa trên bộ ba lược đồ (G,S,M) được xây dựng từ các nguồn thông tin dữ liệu cần tích hợp.



Hình 1. *Mô hình tích hợp dữ liệu cơ bản.*



Hình 2. *Mô tả quá trình tích hợp dữ liệu trong ETL*

Lược đồ toàn cục G: giống như lược đồ quan hệ trong lý thuyết về CSDL, mô tả các rằng buộc nhất quán, các rằng buộc khoá và các yêu cầu về tính độc lập giữa các nguồn thông tin dữ liệu.

Lược đồ nguồn S: Mô tả cấu trúc của tập các nguồn dữ liệu cần tích hợp.

Các ánh xạ M: Bao gồm các ánh xạ được thiết lập giữa lược đồ toàn cục và các lược đồ nguồn dữ liệu.

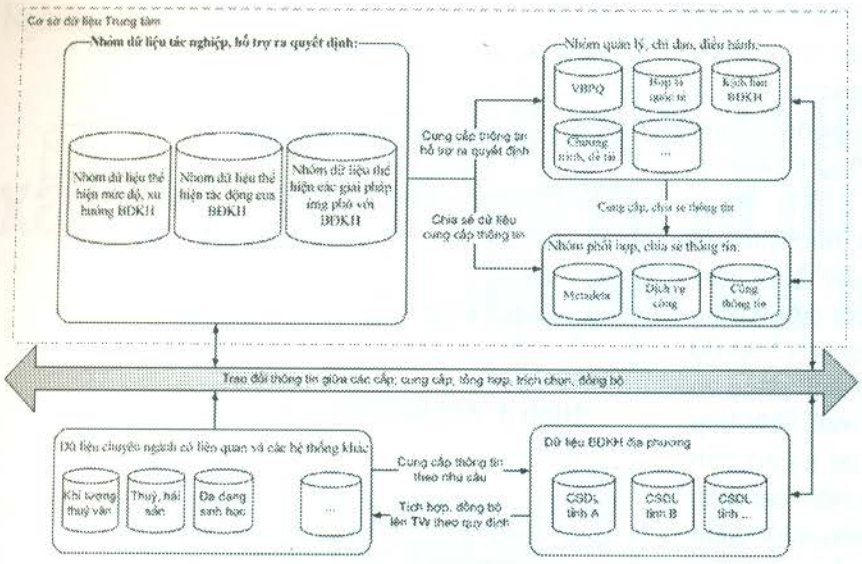
Trên cơ sở xem xét các rằng buộc được định nghĩa trong G và cấu trúc biểu diễn trong S, người thiết kế hệ thống sẽ xác định các ánh xạ tương ứng giữa các thực thể dữ liệu trong các nguồn dữ liệu (ở đây là các CSDL).

Phương pháp này có ưu điểm là biểu diễn được các ngữ nghĩa thông tin dữ liệu thông qua bộ ba (G,S,M). Nhưng nhược điểm là cần biết cấu trúc và rằng buộc của các CSDL trong hệ thống.

Có nhiều cách biến đổi từ dữ liệu nguồn để đạt được mục tiêu về thông tin dữ liệu sau tích hợp cần thiết (dữ liệu đích).

Trích xuất (E): Là quá trình theo dõi các thay đổi trên dữ liệu nguồn, cập nhật dữ liệu tương ứng vào hệ thống đích. Việc này có tác dụng: Giảm ảnh hưởng lên hệ thống nguồn; giảm thời gian xử lý; các phương pháp: 4 phương pháp.

Biến đổi (T): Kiểm tra tính hợp lệ; làm sạch dữ liệu; giải nghĩa và ánh xạ dữ liệu; tạo và quản lý khóa; Tổng hợp dữ liệu.



Hình 3. *Mô hình CSDL Quốc gia về BĐKH.*

Nạp (L): *Nạp và duy trì các chiều:* Chiều thay đổi chậm; bảng cầu nối; chiều thời gian.

Nạp dữ kiện: Dữ liệu lớn; tham chiếu đến các bảng chiếu.

Tích hợp dữ liệu trong cơ sở dữ liệu Quốc gia về BĐKH

Cơ sở dữ liệu Quốc gia về BĐKH

Theo nghiên cứu và đề xuất của đề tài, mô hình CSDL Quốc gia về BĐKH được thể hiện như *Hình 3*. Trong đó, có CSDL Trung tâm, CSDL chuyên ngành có liên quan, và CSDL BĐKH địa phương.

Cơ sở dữ liệu Trung tâm: Được gom thành 03 nhóm như sau: Nhóm dữ liệu phục vụ quản lý, chỉ đạo, điều hành; Nhóm dữ liệu tác nghiệp, hỗ trợ ra quyết định; Nhóm dữ liệu phục vụ phối hợp, chia sẻ và công bố thông tin.

Cơ sở dữ liệu Trung tâm được tổ chức lưu trữ và quản lý tại Cục KTTV & BĐKH, được cập nhật trực tiếp hoặc được đồng bộ, tích hợp một lần (hoặc định kỳ) từ các CSDL thành phần (CSDL chuyên ngành có liên quan), từ các CSDL BĐKH địa phương.

Cơ sở dữ liệu Chuyên ngành có liên quan: được tổ chức lưu trữ, quản lý và cập nhật tại các bộ ngành (Bộ TN&MT, Bộ TN&PTNT, Bộ CT,...), chúng được tổ chức lưu trữ và quản lý thành các dữ liệu với phạm vi chuyên ngành nhỏ hơn như: Khí tượng thuỷ văn; kiểm kê khí nhà kính, thuỷ, hải sản; đa dạng sinh học.

Cơ sở dữ liệu BĐKH địa phương: được tổ chức lưu trữ, quản lý và cập nhật tại các tỉnh/thành phố.

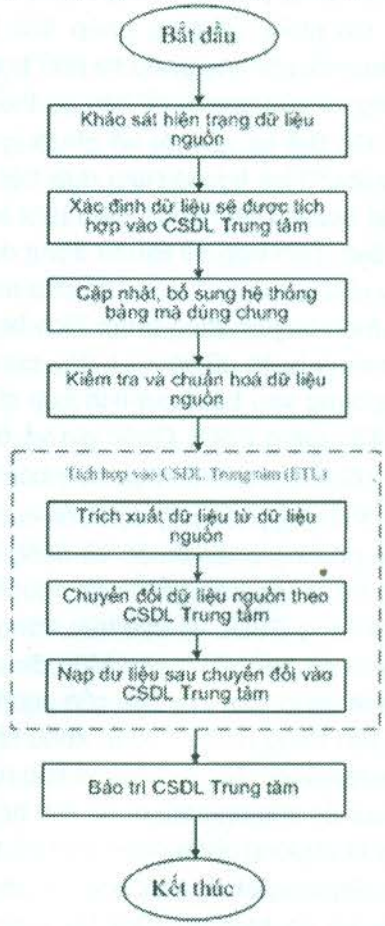
Quy trình tích hợp CSDL Quốc gia về BĐKH.

Trên cơ sở tổng quan về vấn đề tích hợp dữ liệu, hiện trạng dữ liệu BĐKH và mô hình tổ chức lưu trữ, quản lý và cập nhật CSDL Quốc gia về BĐKH, Quy trình tích hợp CSDL trong CSDL Quốc gia về BĐKH được đưa ra như Hình 4. Trong quy trình này, 04 bước đầu hướng đến việc chuẩn bị tích hợp dữ liệu, 03 bước tiếp theo (đặt trong khung có đường nét đứt) thực hiện tích hợp dữ liệu nguồn vào CSDL Trung tâm, và bước cuối cùng là thực hiện bảo trì CSDL Trung tâm để đảm bảo việc tích hợp thêm dữ liệu vào CSDL Trung tâm không làm ảnh hưởng (tác động xấu) đến CSDL Trung tâm, và hệ thống phần mềm đã triển khai tại CSDL Trung tâm.

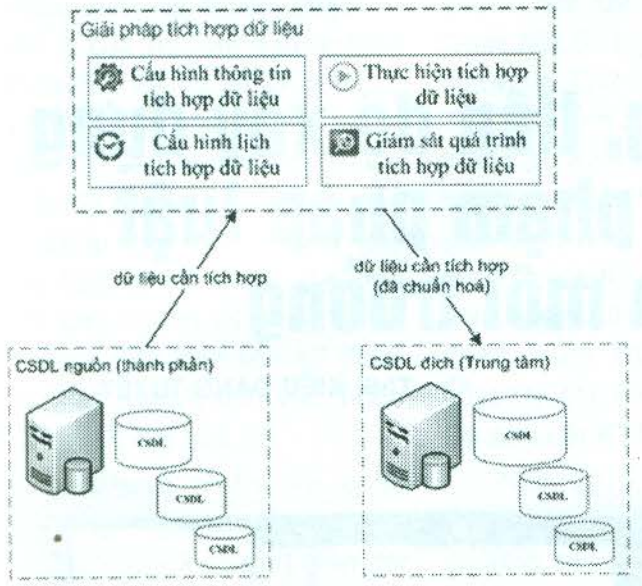
Giải pháp tích hợp CSDL Quốc gia về BĐKH.

Về cơ bản, việc tích hợp dữ liệu thường chia làm hai trường hợp. Trường hợp thứ nhất, dữ liệu chỉ cần tích hợp một lần, trường hợp này áp dụng cho các trường hợp dữ liệu nguồn sẽ không có sự thay đổi trong tương lai (tức là dữ liệu nguồn không được cập nhật sau khi tích hợp vào CSDL đích, ở đây là CSDL Trung tâm). Khi đó, giải pháp đưa ra là viết công cụ phần mềm (tool) thực hiện chuyển đổi dữ liệu nguồn theo khuôn dạng (cấu trúc) của CSDL Trung tâm là xong, trường hợp này có thể gọi là chuyển đổi (convert) dữ liệu.

Trường hợp thứ hai, dữ liệu nguồn cần tích hợp một cách định kỳ vào CSDL Trung tâm (do dữ liệu nguồn được cập nhật thường



Hình 4. *Quy trình tích hợp CSDL*



Hình 5. *Giải pháp tích hợp dữ liệu*

xuyên). Khi đó, cần phải có giải pháp tổng thể để giải quyết trường hợp này. Dưới đây là mô hình giải pháp tích hợp dữ liệu định kỳ ở mức cơ bản (*Hình 5*).

Yêu cầu chung của giải pháp tích hợp dữ liệu định kỳ, tự động:

- Có thể cấu hình, quản lý nguồn dữ liệu cần tích hợp về CSDL Trung tâm. Với mỗi nguồn dữ liệu, cho phép lựa chọn các đối tượng dữ liệu sẽ tích hợp về CSDL Trung tâm.

Việc tích hợp dữ liệu nguồn về CSDL Trung tâm có thể thực hiện theo yêu cầu, hoặc theo lịch được thiết lập tự động.

Có thể theo dõi, giám sát, kiểm soát được quá trình tích hợp dữ liệu, ghi lại nhật ký tích hợp dữ liệu.

Có thể tuy chọn thực hiện tích hợp trực tiếp giữa dữ liệu nguồn và CSDL Trung tâm hoặc thông qua file dữ liệu trung gian.

Dữ liệu tích hợp có thể là dữ liệu không gian hoặc dữ liệu phi không gian.

Hỗ trợ tích hợp dữ liệu thông qua mạng internet.

Trên cơ sở các yêu cầu này, *Hình 5* được mô tả sơ bộ như sau:

Môđun Cấu hình thông tin tích hợp dữ liệu

Thực hiện cấu hình và quản lý thông tin cấu hình tích hợp dữ liệu, cung cấp thông tin cấu hình tích hợp dữ liệu cho môđun Thực hiện tích hợp dữ liệu. Thông tin cấu hình tích hợp bao gồm: Thông tin Nguồn dữ liệu; Các đối tượng dữ liệu của Nguồn dữ liệu xác định sẽ tích hợp với CSDL Trung tâm; Thực hiện tích hợp trực tiếp hay thông qua file dữ liệu trung gian?...

Trong quá trình thiết lập thông tin cấu hình tích hợp dữ liệu, hệ thống sẽ tự động bổ sung các trường (field) cần thiết vào các bảng (table) trong dữ liệu nguồn và CSDL Trung tâm để hệ thống có thể theo dõi, kiểm soát được dữ liệu nào cần tích hợp ở phía dữ liệu nguồn.

Trường hợp tích hợp dữ liệu qua mạng internet, cần thống nhất cấu trúc file dữ liệu tích hợp ở dạng Thông điệp (Message) và xây dựng dịch vụ tích hợp (webservice) để thực hiện truyền tải Message từ dữ liệu nguồn về CSDL Trung tâm.

Môđun Cấu hình lịch tích hợp dữ liệu

Thực hiện cấu hình và quản lý thông tin Cấu hình lịch tích hợp, cung cấp thông tin này cho dịch vụ kích hoạt tích hợp dữ liệu. Việc cấu hình lịch tích hợp được thực hiện cho từng nguồn dữ liệu, mỗi nguồn dữ liệu có thể có lịch tích hợp dữ liệu khác nhau và tuỳ thuộc vào tần suất thay đổi/cập nhật của dữ liệu nguồn.

Môđun Thực hiện tích hợp dữ liệu

Được kích hoạt từ giao diện người dùng hoặc từ dịch vụ kích hoạt tích hợp dữ liệu. Thực hiện đọc thông tin cấu hình tích hợp để trích xuất dữ liệu từ dữ liệu nguồn, biến đổi dữ liệu nguồn vừa đọc được phù hợp với CSDL Trung tâm, sau đó nạp vào CSDL Trung tâm. Trong quá trình tích hợp, hệ thống phải lưu lại nhật ký, sao lưu cần thiết để có thể khôi phục lại CSDL Trung tâm nếu có sự cố xảy ra.

Môđun Giám sát quá trình tích hợp dữ liệu

Giám sát quá trình tích hợp được thực hiện qua việc đọc liên tục quá trình tích hợp dữ liệu đang diễn ra, sau đó thể hiện thông tin này dưới dạng biểu đồ cho người quản trị theo dõi. Ngoài ra, việc giám sát còn thể hiện qua việc tra cứu lại nhật ký tích hợp dữ liệu

Kết luận

Với kết quả nghiên cứu đạt được của đề tài, có thể nói, việc đề xuất mô hình tổ chức lưu trữ, quản lý và cập nhật CSDL Quốc gia về BĐKH, kèm theo mô hình, giải pháp tích hợp dữ liệu từ các CSDL chuyên ngành có liên quan (CSDL Thành phần) và CSDL BĐKH địa phương vào CSDL Trung tâm là phù hợp với mô hình phân cấp quản lý, phù hợp với Chương trình mục tiêu quốc gia ứng phó với biến đổi khí hậu đã được phê duyệt.

Trên cơ sở kết quả nghiên cứu, thử nghiệm của đề tài, việc thực hiện nhiệm vụ Xây dựng CSDL Quốc gia về BĐKH sẽ trở lên hiện thực hơn, và dễ thành công hơn.■