1. Một số vấn đề về chất lượng dữ liệu
2. Một nguồn
3. Mức lược đồ (Schema Level)

* Nguyên nhân:
* Thiếu các ràng buộc toàn vẹn: Chất lượng dữ liệu của một nguồn chủ yếu phụ thuộc vào mức độ mà nó bị chi phối bởi lược đồ và các ràng buộc toàn vẹn. Đối với các nguồn không có lược đồ, chẳng hạn như tệp, có rất ít giới hạn về những dữ liệu được nhập và lưu trữ, dẫn đến khả năng cao xảy ra lỗi và không nhất quán.
* Thiết kế lược đồ tệ: Các lược đồ và mô hình dữ liệu tệ có rất ít các ràng buộc toàn vẹn để ngăn sự thiếu nhất quán.
* Hệ quả:
* Giá trị không hợp lệ: Ngoài vùng, sai format
* Vi phạm ràng buộc giá trị
* Vi phạm tính độc nhất
* Vi phạm tham chiếu

1. Mức thể hiện (Instance level)

* Hệ quả:
* Thiếu giá trị
* Sai chính tả
* Giá trị được mã hóa, viết tắt
* Nhiều thông tin được nhúng vào (sai trường)
* Giá trị bị đặt vào sai trường
* Vi phạm ràng buộc thuộc tính
* Chuyển vị từ
* Bản ghi bị lặp lại
* Bản ghi bị mâu thuẫn
* Tham chiếu sai

1. Multipile source

* Các vấn đề hiện tại trong các nguồn đơn lẻ càng trầm trọng hơn khi nhiều nguồn cần được tích hợp. Mỗi nguồn có thể chứa dữ liệu xấu và dữ liệu trong các nguồn có thể được biểu diễn khác nhau, chồng chéo hoặc mâu thuẫn với nhau. Điều này là do các nguồn thường được phát triển, triển khai và duy trì một cách độc lập để phục vụ các nhu cầu cụ thể. Điều này dẫn đến một mức độ không đồng nhất lớn trong hệ thống quản lý dữ liệu, mô hình dữ liệu, thiết kế lược đồ và dữ liệu thực tế
* Ở mức lược đồ, sự khác biệt về mô hình dữ liệu và thiết kế lược đồ phải được giải quyết tương ứng bằng các bước dịch lược đồ và tích hợp lược đồ. Các vấn đề chính trong thiết kế lược đồ là đặt tên và xung đột cấu trúc. Xung đột đặt tên nảy sinh khi cùng một tên được sử dụng cho các đối tượng khác nhau hoặc các tên khác nhau được sử dụng cho cùng một đối tượng. Xung đột cấu trúc xảy ra trong nhiều biến thể và đề cập đến các biểu diễn khác nhau của cùng một đối tượng trong các nguồn khác nhau, ví dụ: biểu diễn thuộc tính so với biểu diễn bảng, cấu trúc thành phần khác nhau, các kiểu dữ liệu khác nhau, các ràng buộc toàn vẹn khác nhau, v.v.
* Ngoài xung đột mức lược đồ, nhiều xung đột chỉ xuất hiện ở mức thể hiện (xung đột dữ liệu). Tất cả các vấn đề từ trường hợp nguồn đơn có thể xảy ra với các đại diện khác nhau trong các nguồn khác nhau (ví dụ: bản ghi trùng lặp, bản ghi mâu thuẫn,…). Hơn nữa, ngay cả khi có cùng tên thuộc tính và kiểu dữ liệu, có thể có các biểu diễn giá trị khác nhau (ví dụ: cho tình trạng hôn nhân) hoặc cách giải thích khác nhau về các giá trị (ví dụ: đơn vị đo lường Dollar so với Euro) trên các nguồn. Hơn nữa, thông tin trong các nguồn có thể được cung cấp ở các cấp độ tổng hợp khác nhau (ví dụ: doanh số trên mỗi sản phẩm so với doanh số trên mỗi nhóm sản phẩm) hoặc đề cập đến các điểm khác nhau trong thời gian (ví dụ: doanh số bán hàng hiện tại tính đến ngày hôm qua cho nguồn 1 so với tuần trước cho nguồn 2).

1. Tham chiếu:

* Erhard Rahm; Hong Hai Do: Data Cleaning: Problems and Current Approaches