1. What is data cleansing and some other names?

* *Data cleaning đề cập đến việc phát hiện và loại bỏ các lỗi và sự không nhất quán khỏi dữ liệu để cải thiện chất lượng của dữ liệu. Các vấn đề về chất lượng dữ liệu xuất hiện trong các bộ sưu tập dữ liệu đơn lẻ, chẳng hạn như tệp và cơ sở dữ liệu (lỗi chính tả trong quá trình nhập dữ liệu, thông tin bị thiếu hoặc dữ liệu không hợp lệ khác). Khi nhiều nguồn dữ liệu cần được tích hợp, ví dụ, trong kho dữ liệu, hệ thống cơ sở dữ liệu liên hợp hoặc hệ thống thông tin dựa trên web toàn cầu, nhu cầu làm sạch dữ liệu tăng lên đáng kể. Điều này là do các nguồn thường chứa dữ liệu dư thừa trong các biểu diễn khác nhau. Để cung cấp quyền truy cập vào dữ liệu chính xác và nhất quán, việc hợp nhất các biểu diễn dữ liệu khác nhau và loại bỏ thông tin trùng lặp trở nên cần thiết.*
* *Data cleaning (làm sạch dữ liệu) còn được gọi là data cleansing (tẩy rửa dữ liệu)* hay *scrubbing (quét dữ liệu).*