Data Cleaning

Là quá trình thay đổi hoặc loại bỏ dữ liệu không chính xác, trùng lặp, bị hỏng không đầy đủ bên trong cơ sở dữ liệu (database). Nếu dữ liệu không chính xác, các thuật toán và kết quả cho ra không đáng tin cậy (dù cho nó có vẻ đúng). Data Cleaning là quá trình tìm ra phương pháp tối đa hóa tính xác thực của tập dữ liệu mà không cần phải xóa thông tin, là quá trình chuyển từ một tập dữ liệu từ không clean sang một tập dữ liệu clean.

Để cleaning một tập dữ liệu thì cần phải thu thập dữ liệu, loại bỏ giá trị trùng lặp, giải quyết các giá trị trùng lặp, quy trình làm sạch tiêu chuẩn,…

Data Nomalization

Là quá trình chuẩn hóa cơ sở dữ liệu, là quá trình phân tách một cơ sở dữ liệu có cấu trúc phức tạp thành những bảng có cấu trúc đơn giản theo những quy luật đảm bảo không làm mất thông tin dữ liệu. Giúp làm giảm bớt sự dư thừa và loại bỏ những sự cố mâu thuẫn về dữ liệu, tiết kiệm không gian lưu trữ. Trong CSDL có các chuẩn phổ biến như 1NF, 2NF, 3NF, 4NF, BCNF (Boyce-Codd).

Data Transformation

Là quá trình chuyển đổi dữ liệu, là quá trình sửa đổi, tính toán, phân tách và kết hợp dữ liệu thô thành các mô hình dữ liệu sẵn sàng phân tích.