DATA QUALITY

Định nghĩa

Data quality:

- là 1 chức năng của Data Management
- Mục đích: cải thiển chất lượng dữ liệu, đảm bảo tiêu chuẩn, đảm bảo phù hợp với mục tiêu nghiên cứu
- Triển khai song song trong suốt quá trình (từ tổ chức, thu thập, lưu trữ … đến phân tích)
- Nguyên nhân: chất lượng dữ liệu bị ảnh hưởng khi ta thao tác trong các quá trình đó

Data quality management (DQM)

- Lập kế hoạch, thực hiện và kiểm soát => việc áp dụng kĩ thuật quản lí chất lượng => vào dữ liệu
- Mục đích: hiểu biết sâu sắc về dữ liệu, đáp ứng nhu cầu người dùng, đảm bảo, duy trì, cải thiện chất lượng để phân tích => đạt được kết quả tốt hơn
- Triển khai trong suốt quá trình tương tự Data Quality phía trên
- Đo lường chất lượng dữ liệu dựa trên phạm vi, tiêu chuẩn, tình huống cụ thể (không thể chỉ nói tốt hay xấu)

Tại sao cần quản lí chất lượng dữ liệu?

- Dữ liệu quan trọng => Vì sao quan trọng?

- Cung cấp thông tin, kiến thức, hiểu biết về đối tượng dữ liệu đó mô tả
- Hỗ trợ việc quản lí, phân tích, dự báo, báo cáo

- Chất lượng dữ liệu tốt => tăng khả năng đề xuất những ý tưởng, giải pháp, quyết định hiệu quả và kịp thời, hạn chế rủi ro.

- DQM đảm bảo thiết lập 1 tiêu chuẩn chung giữa các phòng ban trong việc chia sẻ, gửi nhận

- DQM cắt giảm chi phí không cần thiết (chi phí do dữ liệu kém)

- DQM hỗ trợ Data Governance and Data Management được vận hành trơn tru

- Là tiền đề để những thuật toán Machine Learning, Data Mining, AI, Deep Learning phát huy tối đa hiệu quả. “Garbage in, garbage out”

Tiêu chí đánh giá Data Quality

1. Tiêu chuẩn **Strong -Wang Framework (1996):** tập trung vào mong muốn, ý kiến, nhận thức của người sử dụng dữ liệu (Dựa vào tài liệu **The Knowledge of Data Management**)
    - Intrinsic Data Quality
        - Accuracy
        - Objectivity (tính khách quan)
        - Believability (mức độ tin cậy)
        - Reputation (ý kiến cá nhân về bộ dữ liệu)
    - Contextual Data Quality
        - Value-added (giá trị có được từ dữ liệu)
        - Relevancy (mức độ liên quan, phù hợp)
        - Timeliness (tính kịp thời)
        - Completeness (tính đầy đủ)
        - Appropriate amount of data (mức độ phù hợp của khối lượng dữ liệu)
    - Representational DQ (trình bày, thể hiện dữ liệu)
        - Interpretability (khả năng giải thích, diễn giải)
        - Ease of understanding (độ dễ hiểu)
        - Representational consistency (tính nhất quán khi trình bày)
        - Concise representation (súc tích, ngắn gọn khi trình bày)
    - Accessibility DQ (khả năng tiếp cận dữ liệu)
        - Accessibility (khả năng tiếp cận)
        - Access security (khả năng truy cập bảo mật)
2. Dựa theo tài liệu **Data Quality for the Information Age (1996).** Tiêu chuẩn dựa trên **cấu trúc dữ liệu**, data model
    - Content
        - Relevance of data
        - The ability to obtain the values (khả năng có được giá trị)
        - Clarity of definitions (mức độ rõ ràng về định nghĩa, ý nghĩa loại dữ liệu)
    - Level of detail
        - Attribute granularity (độ chi tiết của thuộc tính)
        - Precision of attribute domains (độ chính xác của các miền thuộc tính)
    - Consistency
        - Tính nhất quán về ngữ nghĩa của các thành phần mô hình
        - Cấu trúc thống nhất của các thuộc tính trên các loại dữ liệu
    - Reaction to change (khả năng thích ứng sự thay đổi)
    - Data Values
        - Accuracy
        - Completeness
        - Currency (khả năng được công nhận)
        - Consistency

- Representation (trình bày dữ liệu)
  - Appropriateness
  - Interpretability
  - Portability (khả năng truyền đạt, chia sẻ)
  - Format precision (định dạng chính xác)
  - Format flexibility (mức độ linh hoạt của định dạng)

3. Trong tài lệu **Improving Data Warehouse and Business Information Quality (1999)**
   - Inherent (gôm các thuộc tính dữ liệu phải có độc lập với mục đích sử dụng)
   - Pragmatic (các thuộc tính dữ liệu liên quan đến mục đích sử dụng dữ liệu)

➔ Kết hợp các phương pháp trên để phù hợp với thời đại công nghệ mới, ta sử dụng phương pháp sau:

- Accuracy
- Completeness
- Consistency
- Integrity (tính toàn vẹn, gắn kết)
- Relevance
- Timeliness
- Validity
- Uniqueness
- Auditability (khả năng dễ dàng thống kê)

Tác hại của dữ liệu kém chất lượng

- Tác động đến hầu hết các bộ phận chức năng, các hoạt động kinh doanh => gây tổn hại về mặt chi phí và hiệu suất công việc

- Phá hủy, bỏ lỡ những giá trị, lợi nhuận mà một tổ chức, cá nhân có thể nhận được => mất lợi thế cạnh tranh, cơ hội phát triển và mở rộng

- Khách hàng, quản lí, người trực tiếp hay gián tiếp sử dụng hoặc tiếp nhận thông tin từ dữ liệu mất niềm tin vào những kết quả, hành động từ việc phân tích dữ liệu có được => các tổ chức, doanh nghiệp, cá nhân đánh mất sự uy tín

Thế nào là dữ liệu kém chất lượng?

- Dữ liệu cung cấp thông tin không đáng tin cậy

- Dữ liệu không đầy đủ, thiếu giá trị tại các ô quan sát

- Dữ liệu bị trùng lặp

- Dữ liệu mơ hồ, khó khăn diễn giải ý nghĩa, không có thông tin mô tả về dữ liệu cụ thể

- Dữ liệu cung cấp thông tin lỗi thời

- Dữ liệu cập nhật trễ

- Dữ liệu có định dạng phức tạp, không nhất quán

…

Nguyên nhân và thách thức đối với Data quality management

- Đến từ cơ cấu tổ chức, và cách thức quản lí của các cấp lãnh đạo

  - Lỗ hổng trong việc vận hành, quản lí các quy trình kĩ thuật và kinh doanh kém, thiếu sự thống nhất và quản lí chặt chẽ hiệu quả
  - Thiếu hệ thống luân chuyển chia sẻ dữ liệu hiệu quả và nghiêm ngặt (sự khác biệt về cấu trúc, định dạng trước và sau khi chia sẻ, dữ liệu bị thất lạc, không đầy đủ, …)
  - Việc quản lí hệ thống cơ sở dữ liệu không hỗ trợ tốt cho việc chia sẻ thông tin

- Đến từ quá trình nhập dữ liệu

  - Vấn đề về giao diện nhập dữ liệu
  - Vấn đề từ việc sử dụng giao diện và các trường dữ liệu trong thời gian dài mà không cơ cấu lại phục vụ dự án cụ thể
  - Ở những bài toán ML việc gán nhãn dữ liệu cần phải có kiến thức, kinh nghiệm về lĩnh vực => những người không được đào tạo hay chuyên môn về lĩnh vực có thể gán nhãn sai
  - Vấn đề từ việc vận hành các quy trình kinh doanh không nhất quán

- Đến từ các chức năng, quy trình xử lí dữ liệu bị tác động bởi tác động của việc thay đổi mục tiêu, hướng phát triển, thay đổi hệ thống dữ liệu (tiếp cận nguồn dữ liệu bị sai)

- Đến từ cách thức sửa chữa, xử lí các vấn đề dữ liệu và từ hệ thống bảo mật thông tin dữ liệu

- Khối lượng dữ liệu ngày càng lớn, sự đa dạng về nguồn dữ liệu và về cấu trúc định dạng dữ liệu phức tạp

- Dữ liệu hiện nay thay đổi nhanh và nhanh chóng trở nên lỗi thời, và dễ dàng gặp phải dữ liệu không chính xác, không đáng tin cậy

Giải pháp cải thiện chất lượng dữ liệu

Data profiling

Data profiling là 1 hình thức phân tích dữ liệu => dùng để kiểm tra dữ liệu và đánh giá chất lượng dữ liệu bằng các kĩ thuật thống kê để khám phá cấu trúc, nội dung và chất lượng => rồi so sánh đối chiếu, lập báo cáo

- Xác định số lượng null values
- Xác định min, max, outliers or unusual values
- Xác định phân phối tương ứng các biến
- Xác định format, loại dữ liệu

Data quality processing

- Data cleansing
- Data enhancement
- Data parsing
- Data formatting
- Data transformation and standardization
- Data reporting
- Data repair