ĐẠI HỌC QUỐC GIA TPHCM – TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN – BỘ MÔN HỆ THỐNG THÔNG TIN

A picture containing logo

Description automatically generated

**BÀI TẬP NHÓM 2**

**Môn học: Hệ thống thông tin phục vụ trí tuệ kinh doanh**

*GVHD: Hồ Thị Hoàng Vy*

|  |  |  |
| --- | --- | --- |
| **Mã nhóm** | **MSSV** | **Họ và tên** |
| **TTKD-24** | 18120227 | Phạm Văn Minh Phương |
| 18120299 | Trương Công Quốc Cường |
| 18120035 | Đoàn Nguyễn Tấn Hưng |
| 18120534 | Hoàng Công Sơn |

**BẢNG PHÂN CÔNG**

|  |  |  |  |
| --- | --- | --- | --- |
| MSSV | Họ và tên | Phân công | Đánh giá % |
| 18120227 | Phạm Văn Minh Phương | Câu hỏi 2 | 100% |
| 18120299 | Trương Công Quốc Cường | Câu hỏi 3 | 100% |
| 18120035 | Đoàn Nguyễn Tấn Hưng | Câu hỏi 1 | 100% |
| 18120534 | Hoàng Công Sơn | Tổng hợp, làm báo cáo | 100% |

Mục lục

[Câu 1: What is data cleansing and some other names? 2](#_Toc82873662)

[Câu 2: Let describe some data quality problems for: Single source; Multiple sources 2](#_Toc82873663)

[I. One source 2](#_Toc82873664)

[II. Multipile source 4](#_Toc82873665)

[Câu 3: List some tools to support data transformation and data cleaning 5](#_Toc82873678)

# **Câu 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).*

# **Câu 2:** **Let describe some data quality problems for: Single source; Multiple sources**

**1. Một nguồn**

*a. 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

*b. 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

**2. 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)

**Câu 3:** **List some tools to support data transformation and data cleaning**

1. Hevo Data: Hỗ trợ tích hợp dữ liệu được xây dựng trước từ hơn 100 nguồn. Là nền tảng tự động hoàn toàn, không yêu cầu việc quản lý và bảo trì. Kiên trúc fault-tolorent đảm bảo dữ liệu được nhất quán và bảo mật.
2. IBM InfoSphere DataStage: Được xây dựng trên công nghệ microservice và container nên dễ dàng phân tích dữ liệu thời gian thực. Các chuyển đổi từ công cụ cung cấp thông tin phong phú và phù hợp tới tổ chức sử dụng. Các tính năng như tìm kiếm tích hợp sẵn, truyền siêu dữ liệu tự đông và đánh dấu đống thời các lỗi biên dịch.
3. Pentaho Data Integration: Là công cụ ETL mã nguồn mở, hỗ trợ nhiều kiến trúc đám mây. Tập các nguồn dữ liệu lớn.
4. Xplenty: Giao diện thân thiện người dùng, không yêu cầu viết mã. Là công cụ dựa trên nền tảng đám mây, giảm tải việc quản lý và bảo trì định kỳ.
5. Tibco Clarity: Giao diện làm sạch dữ liệu trực quan. Có thể trực quan hóa dữ liệu đang được xử lý, cung cấp cái nhìn rõ hơn về tập dữ liệu. Cấu hình việc xác thực dựa trên quy tắc cung cấp thêm 1 lớp quản lý chất lượng dữ liệu.