



XỬ LÝ VÀ LÀM SẠCH DỮ LIỆU

Nhóm biên soạn:

1. Lê Ngọc Thành
2. Nguyễn Ngọc Thảo
3. Phạm Trọng Nghĩa
4. Nguyễn Thái Vũ
5. Trương Tấn Khoa

Năm 2022



1 GIỚI THIỆU VỀ PHÂN TÍCH VÀ XỬ LÝ DỮ LIỆU



NỘI DUNG



Giới thiệu về phân tích dữ liệu



Vai trò của xử lý dữ liệu



Các bước xử lý dữ liệu



Tổng kết

NỘI DUNG



Giới thiệu về phân tích dữ liệu



Vai trò của xử lý dữ liệu



Các bước xử lý dữ liệu



Tổng kết

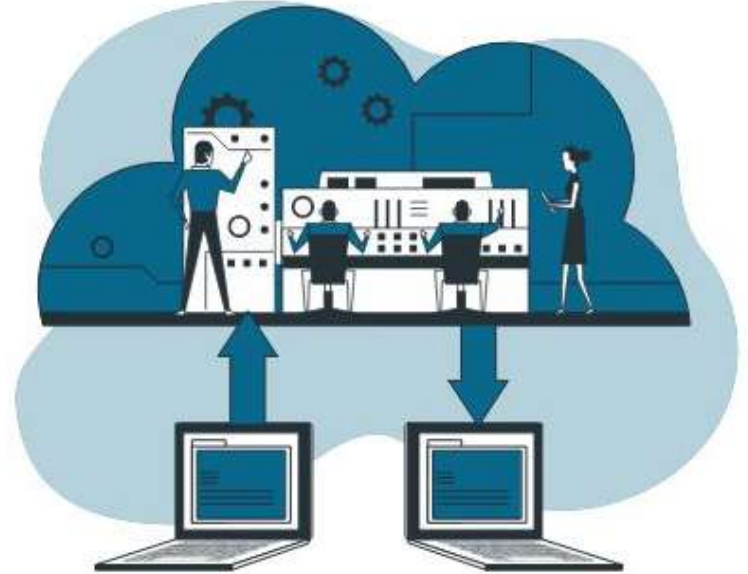
PHÂN TÍCH DỮ LIỆU

Phân tích dữ liệu (data analysis):

- Phát hiện
- Giải thích
- Truyền đạt

thông tin có ý nghĩa trong dữ liệu.

➤ Đưa ra những quyết định có lợi cho kinh doanh.



PHÂN TÍCH DỮ LIỆU

Dữ liệu đã sẵn sàng cho phân tích?

- Dữ liệu thiếu?
 - Dữ liệu sai?
 - Dư thừa?
 - ...
- Quá trình phân tích sai.



PHÂN TÍCH DỮ LIỆU

Nhắc lại quá trình phân tích dữ liệu gồm



Đưa ra câu hỏi
(ask)



Chuẩn bị
(prepare)



Xử lý
(process)



Phân tích
(analyze)



Chia sẻ
(share)



Triển khai
(act)

NỘI DUNG



Giới thiệu về phân tích dữ liệu



Vai trò của xử lý dữ liệu



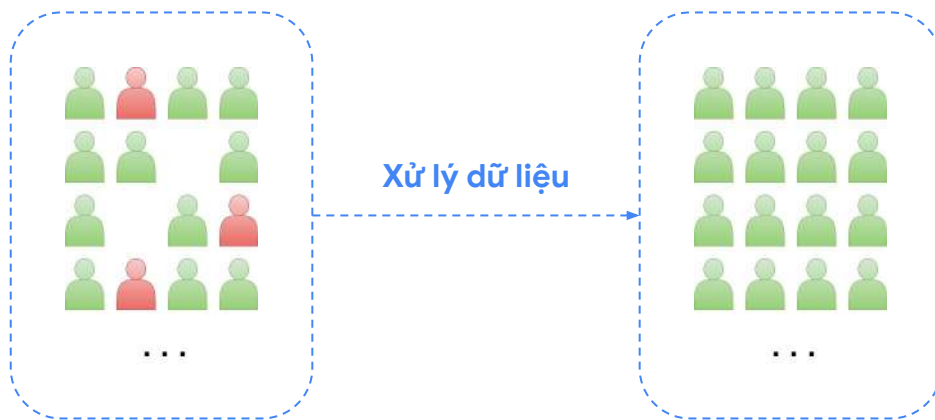
Các bước xử lý dữ liệu



Tổng kết

VAI TRÒ CỦA XỬ LÝ DỮ LIỆU

Xử lý (process) và làm sạch (clean) dữ liệu đảm bảo dữ liệu cho quá trình phân tích.



VAI TRÒ CỦA XỬ LÝ DỮ LIỆU

Ví dụ: xóa dữ liệu trùng nhau.

ID	Tên
123	Nguyễn Văn A
124	Trương Thị B
125	Võ Văn C
123	Nguyễn Văn A

Xử lý dữ liệu

ID	Tên
123	Nguyễn Văn A
124	Trương Thị B
125	Võ Văn C

NỘI DUNG



Giới thiệu về phân tích dữ liệu



Vai trò của xử lý dữ liệu



Các bước xử lý dữ liệu



Tổng kết

CÁC BƯỚC XỬ LÝ DỮ LIỆU



NỘI DUNG



Giới thiệu về phân tích dữ liệu



Vai trò của xử lý dữ liệu



Các bước xử lý dữ liệu



Tổng kết

TỔNG KẾT

Các bài học tiếp theo:

- Tìm hiểu chi tiết các bước xử lý dữ liệu
- Các kỹ thuật làm sạch dữ liệu
- Tầm quan trọng của xử lý dữ liệu trong quá trình phân tích





2 TOÀN VỆN DỮ LIỆU



NỘI DUNG



Khái niệm toàn vẹn dữ liệu



Nguy cơ của sự không toàn vẹn



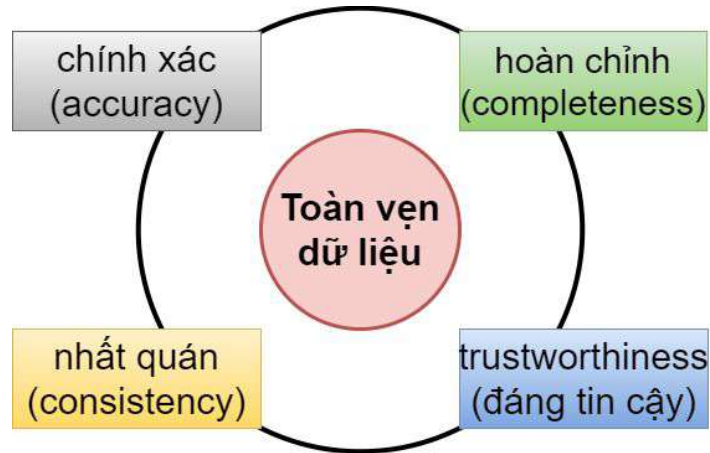
Nguyên nhân gây ra sự không toàn vẹn



Một số giải pháp

KHÁI NIỆM TOÀN VỆN DỮ LIỆU

Toàn vẹn dữ liệu ảnh hưởng đến kết quả phân tích.



1. KHÁI NIỆM TOÀN VỆN DỮ LIỆU

Dữ liệu được lưu trữ chính xác và không có lỗi.

Ví dụ dữ liệu không chính xác: Trường Đại học Khoa học Tự nhiên TP HCM, có địa chỉ ở Hà Nội.



1. KHÁI NIỆM TOÀN VỆN DỮ LIỆU

hoàn chỉnh
(completeness)

Đo lường mức độ đầy đủ của dữ liệu.

Ví dụ, dữ liệu gồm ba thuộc tính (bắt buộc): id, họ tên, số điện thoại.

Nếu một trong ba giá trị bị thiếu, thì dữ liệu mất đi tính hoàn chỉnh.



1. KHÁI NIỆM TOÀN VỆN DỮ LIỆU

nhất quán
(consistency)

Dữ liệu giống nhau được lưu trữ ở nơi khác nhau phải khớp với nhau.

Ví dụ: trong một công ty đa quốc gia, dữ liệu về tiền tệ đều dùng đơn vị \$



1. KHÁI NIỆM TOÀN VỆN DỮ LIỆU

trustworthiness
(đáng tin cậy)

Dữ liệu có đáng tin không có mâu thuẫn với các nguồn khác không?

Ví dụ thông tin khách hàng: cùng một khách hàng nhưng lại có ngày sinh khác nhau.

NỘI DUNG



Khái niệm toàn vẹn dữ liệu



Nguy cơ của sự không toàn vẹn



Nguyên nhân gây ra sự không toàn vẹn





Một số giải pháp

NGUY CƠ CỦA SỰ KHÔNG TOÀN VỆN

- Toàn vẹn dữ liệu quan trọng đối với các hệ thống lưu trữ, xử lý, truy xuất dữ liệu.
- Dữ liệu bị tổn hại (compromised data).
 - Không toàn vẹn dữ liệu.
 - Phân tích sai. Gây ra hậu quả nghiêm trọng.



NGUY CƠ CỦA SỰ KHÔNG TOÀN VỆN

	Ảnh đầu vào	Dự đoán
Toàn vẹn dữ liệu		Chó
Không toàn vẹn dữ liệu		???

NỘI DUNG



Khái niệm toàn vẹn dữ liệu



Nguy cơ của sự không toàn vẹn



Nguyên nhân gây ra sự không toàn vẹn



Một số giải pháp

NGUYÊN NHÂN GÂY RA SỰ KHÔNG TOÀN VỆN

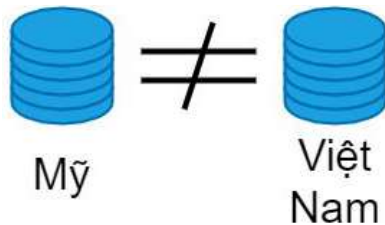
Dữ liệu không còn toàn vẹn, do:

- Nhân bản (Replicated) dữ liệu.
- Chuyển đổi (Transferred) dữ liệu.
- Thao tác (Manipulated) dữ liệu.



NGUYÊN NHÂN GÂY RA SỰ KHÔNG TOÀN VỆN

Nhân bản dữ liệu: lưu trữ dữ liệu ở nhiều nơi.

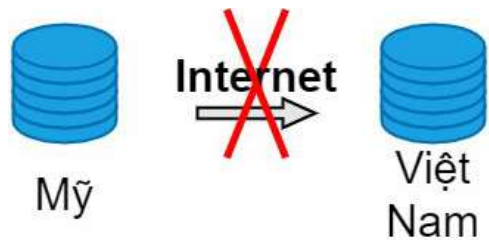


Nguyên nhân: địa lý, múi giờ,...

- Dữ liệu không đồng nhất.
- **Dữ liệu không toàn vẹn (cụ thể là không nhất quán).**

NGUYÊN NHÂN GÂY RA SỰ KHÔNG TOÀN VỆN

Chuyển đổi dữ liệu: truyền tải dữ liệu từ nơi này sang nơi khác.

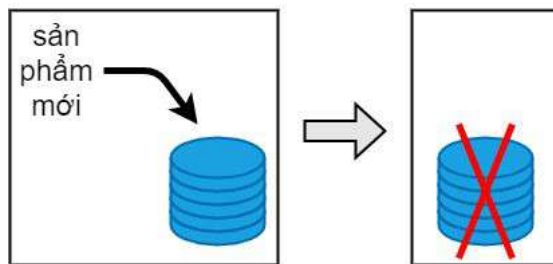


Nguyên nhân: đường truyền bị ngắt

- Dữ liệu hoàn chỉnh.
- **Dữ liệu không toàn vẹn (cụ thể là không nhất quán).**

NGUYÊN NHÂN GÂY RA SỰ KHÔNG TOÀN VỆN

Thao tác dữ liệu: thay đổi, sửa đổi dữ liệu cho dễ đọc và có tổ chức hơn.



Nguyên nhân: quá trình thao tác bị lỗi.

- **Dữ liệu không toàn vẹn (cụ thể là không nhất quán).**

NGUYÊN NHÂN GÂY RA SỰ KHÔNG TOÀN VỆ



NỘI DUNG



Khái niệm toàn vẹn dữ liệu



Nguy cơ của sự không toàn vẹn



Nguyên nhân gây ra sự không toàn vẹn



Một số giải pháp

MỘT SỐ GIẢI PHÁP

- Kết hợp với Kỹ sư dữ liệu (Data Engineer) hoặc Kỹ sư kho dữ liệu (Data warehouse engineer).
- Tiếp theo, sẽ tìm hiểu các cách kiểm tra tính toàn vẹn của dữ liệu.





3 MỐI QUAN HỆ GIỮA DỮ LIỆU VÀ MỤC TIÊU KINH DOANH



NỘI DUNG



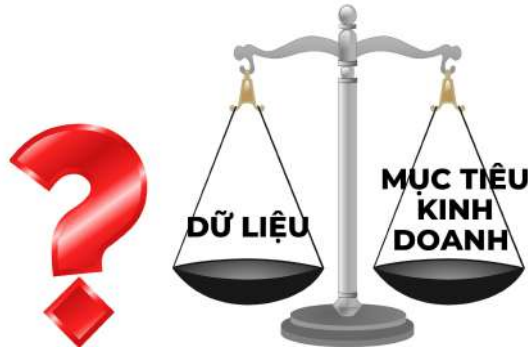
Giới thiệu mối quan hệ giữa dữ liệu và mục tiêu kinh doanh.



Vấn đề: không có dữ liệu để phân tích

MỐI QUAN HỆ GIỮA DỮ LIỆU VÀ MỤC TIÊU KINH DOANH

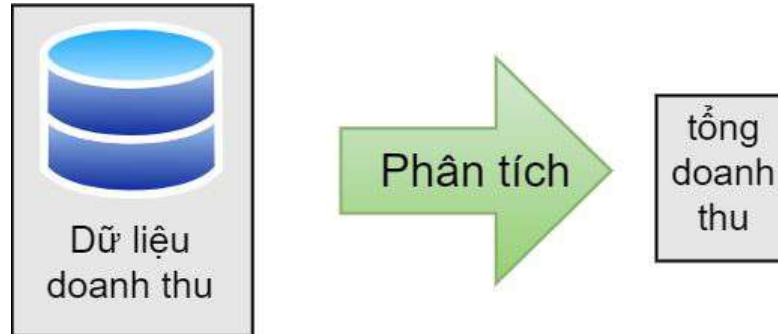
- Kiểm tra: dữ liệu phân tích có liên quan đến quyết định kinh doanh không?
- Sau khi phân tích dữ liệu, có trả lời được câu hỏi đặt ra?



MỐI QUAN HỆ GIỮA DỮ LIỆU VÀ MỤC TIÊU KINH DOANH

Ví dụ: giải quyết câu hỏi về doanh thu

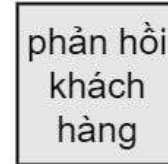
➤ “Dữ liệu doanh thu”



MỐI QUAN HỆ GIỮA DỮ LIỆU VÀ MỤC TIÊU KINH DOANH

Ví dụ: giải quyết câu hỏi về phản hồi của khách hàng

➤ “Dữ liệu về khách hàng”



MỐI QUAN HỆ GIỮA DỮ LIỆU VÀ MỤC TIÊU KINH DOANH

- Duy trì tính toàn vẹn của dữ liệu giúp đảm bảo sự liên kết giữa dữ liệu và hoạt động kinh doanh.
- Toàn vẹn dữ liệu: chính xác, đầy đủ, nhất quán và đáng tin cậy.



MỐI QUAN HỆ GIỮA DỮ LIỆU VÀ MỤC TIÊU KINH DOANH

Nếu nhà phân tích cho rằng: mục tiêu kinh doanh cần được điều chỉnh.

- Thảo luận với các bên liên quan.



NỘI DUNG



Giới thiệu mối quan hệ giữa dữ liệu và mục tiêu kinh doanh



Không có dữ liệu để phân tích

KHÔNG CÓ DỮ LIỆU ĐỂ PHÂN TÍCH

Câu hỏi kinh doanh đã được đặt ra

➤ *Không có dữ liệu phù hợp để phân tích?*

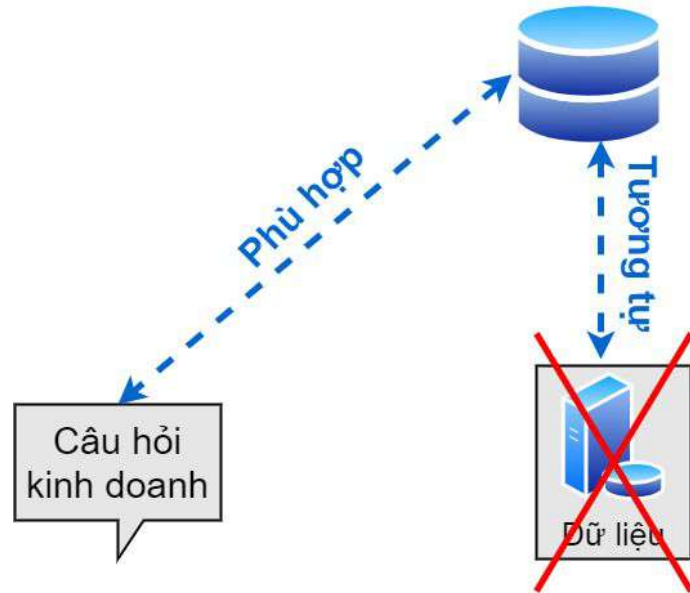
Câu hỏi
kinh doanh

PHÙ HỢP?



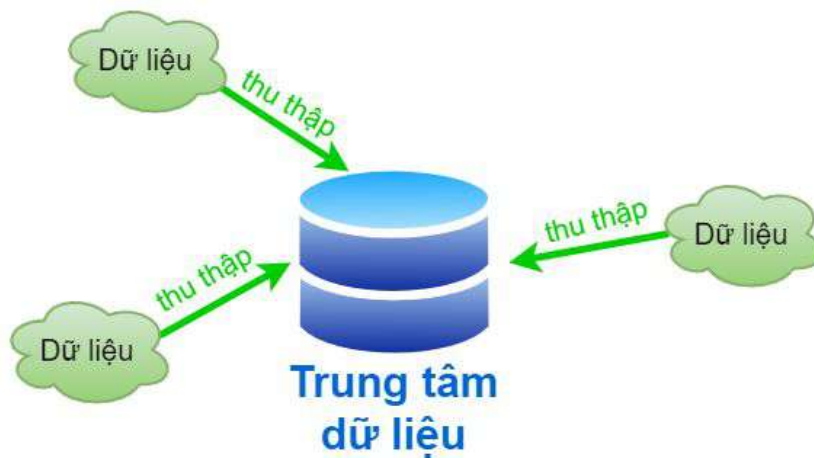
KHÔNG CÓ DỮ LIỆU ĐỂ PHÂN TÍCH

- Giải pháp 1: tìm nguồn dữ liệu tương tự và vẫn phù hợp với mục tiêu kinh doanh.



KHÔNG CÓ DỮ LIỆU ĐỂ PHÂN TÍCH

➤ Giải pháp 2: thu thập thêm dữ liệu





4 VẤN ĐỀ VỀ DỮ LIỆU KHÔNG ĐẦY ĐỦ



NỘI DUNG



Vấn đề dữ liệu không đầy đủ

Giới thiệu vấn đề dữ liệu không đầy đủ

Một số trở ngại thường gặp



Giới thiệu lấy mẫu ngẫu nhiên

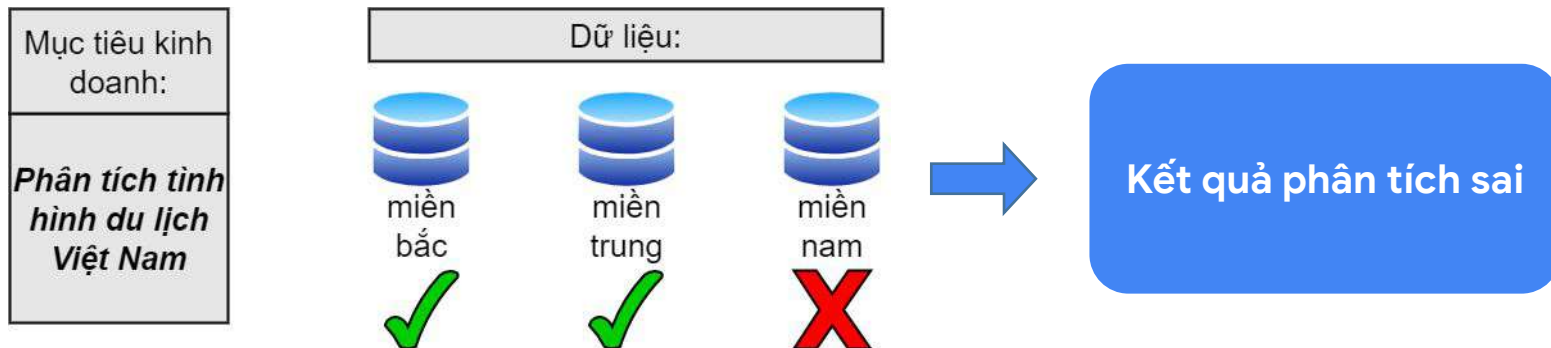
Giới thiệu lấy mẫu tổng thể

Giới thiệu lấy mẫu ngẫu nhiên

VẤN ĐỀ DỮ LIỆU KHÔNG ĐẦY ĐỦ

Tình trạng: đã xác định được mục tiêu kinh doanh, nhưng không đủ (hoặc không có) dữ liệu để phân tích.

Ví dụ



NỘI DUNG



Vấn đề dữ liệu không đầy đủ

Giới thiệu vấn đề dữ liệu không đầy đủ

Một số trở ngại thường gặp



Giới thiệu lấy mẫu ngẫu nhiên

Giới thiệu lấy mẫu tổng thể

Giới thiệu lấy mẫu ngẫu nhiên

VẤN ĐỀ DỮ LIỆU KHÔNG ĐẦY ĐỦ

Một số trở ngại thường gặp:

- Dữ liệu bị giới hạn từ một nguồn



VẤN ĐỀ DỮ LIỆU KHÔNG ĐẦY ĐỦ

Dữ liệu bị giới hạn từ một nguồn:

- *Dữ liệu có chính xác?*
- *Các nguồn khác có xu hướng khác?*
- **Giải pháp: Thu thập dữ liệu từ các nguồn khác.**



VẤN ĐỀ DỮ LIỆU KHÔNG ĐẦY ĐỦ

Một số trở ngại thường gặp:

- Dữ liệu bị giới hạn từ một nguồn
- Dữ liệu vẫn được cập nhật



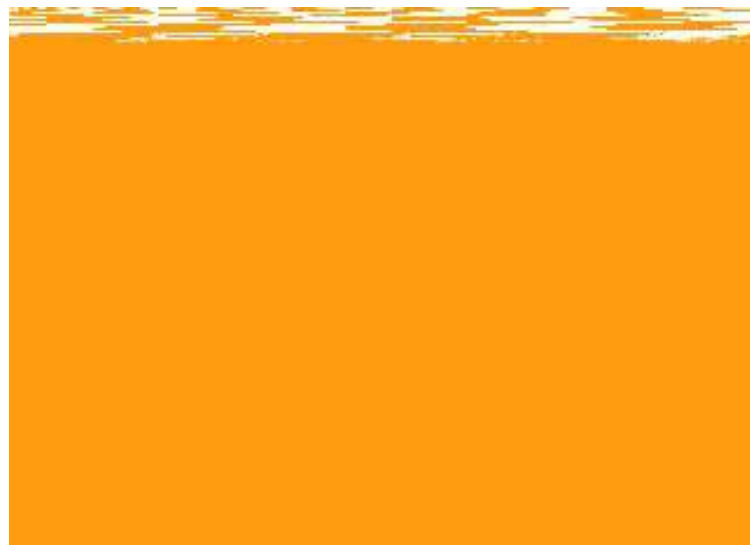
VẤN ĐỀ DỮ LIỆU KHÔNG ĐẦY ĐỦ

Dữ liệu vẫn được cập nhật. Ví dụ:

- Mục tiêu: phân tích tình hình du lịch tháng 6.
- Tình trạng: đang ở giữa tháng 6 và lượng khách vẫn đang tăng.

➤ **Giải pháp:**

- Chờ đợi dữ liệu hoàn thiện.
- Phân tích dựa vào dữ liệu trong quá khứ.



VẤN ĐỀ DỮ LIỆU KHÔNG ĐẦY ĐỦ

Một số trở ngại thường gặp:

- Dữ liệu bị giới hạn từ một nguồn
- Dữ liệu vẫn được cập nhật
- Dữ liệu bị lỗi thời



VẤN ĐỀ DỮ LIỆU KHÔNG ĐẦY ĐỦ

Dữ liệu lỗi thời. Ví dụ:

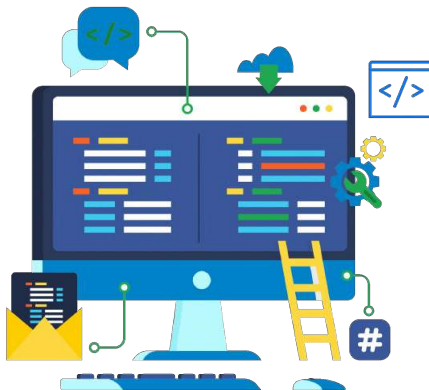
- Mục tiêu: phân tích địa điểm du lịch nổi tiếng gần đây.
- Dữ liệu: *được thu thập 10 năm về trước.*
- **Giải pháp: Thu thập dữ liệu cần thiết.**



VẤN ĐỀ DỮ LIỆU KHÔNG ĐẦY ĐỦ

Một số trở ngại thường gặp:

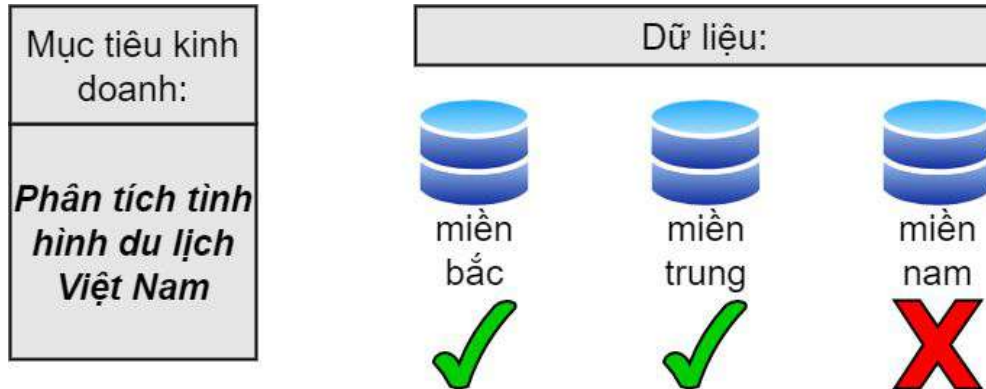
- Dữ liệu bị giới hạn từ một nguồn
- Dữ liệu vẫn được cập nhật
- Dữ liệu bị lỗi thời
- Dữ liệu bị giới hạn địa lý



VẤN ĐỀ DỮ LIỆU KHÔNG ĐẦY ĐỦ

Dữ liệu giới hạn địa lý.

- Ví dụ: không có toàn bộ dữ liệu ở Việt Nam



NỘI DUNG



Vấn đề dữ liệu không đầy đủ

Giới thiệu vấn đề dữ liệu không đầy đủ

Một số trở ngại thường gặp



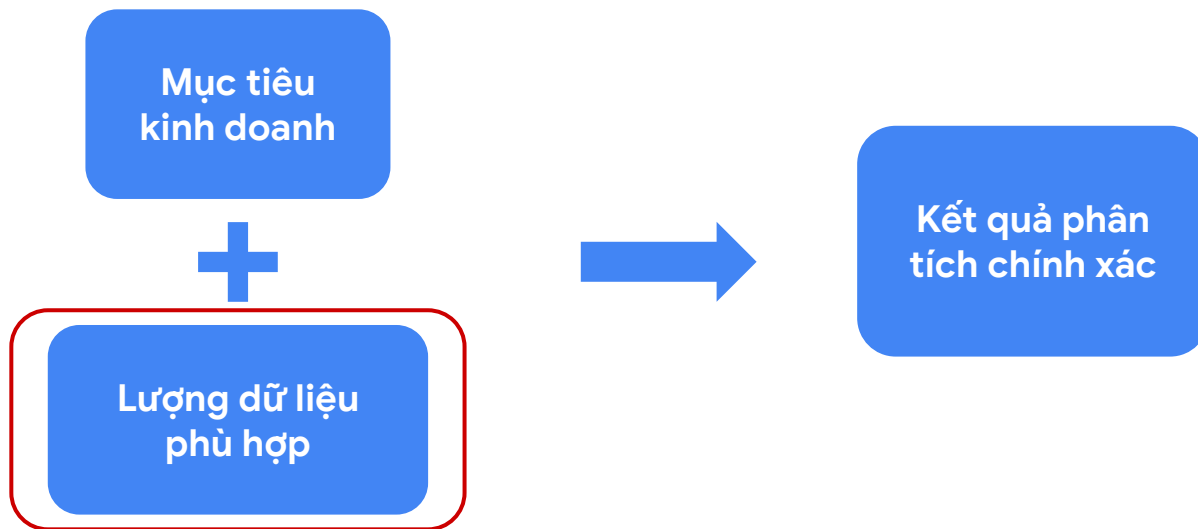
Giới thiệu lấy mẫu ngẫu nhiên

Giới thiệu lấy mẫu tổng thể

Giới thiệu lấy mẫu ngẫu nhiên

GIỚI THIỆU LẤY MẪU NGẪU NHIÊN

Nhắc lại



NỘI DUNG



Vấn đề dữ liệu không đầy đủ

Giới thiệu vấn đề dữ liệu không đầy đủ

Một số trở ngại thường gặp



Giới thiệu lấy mẫu ngẫu nhiên

Giới thiệu lấy mẫu tổng thể

Giới thiệu lấy mẫu ngẫu nhiên

GIỚI THIỆU LẤY MẪU TỔNG THỂ

Lấy mẫu tổng thể (population): sử dụng 100% lượng dữ liệu để phân tích

Mục tiêu:

*Phân tích tình
hình du lịch
Việt Nam*

Dữ liệu:



toàn bộ 100 triệu
người dân Việt Nam

Khả thi không?

NỘI DUNG



Vấn đề dữ liệu không đầy đủ

Giới thiệu vấn đề dữ liệu không đầy đủ

Một số trở ngại thường gặp



Giới thiệu lấy mẫu ngẫu nhiên

Giới thiệu lấy mẫu tổng thể

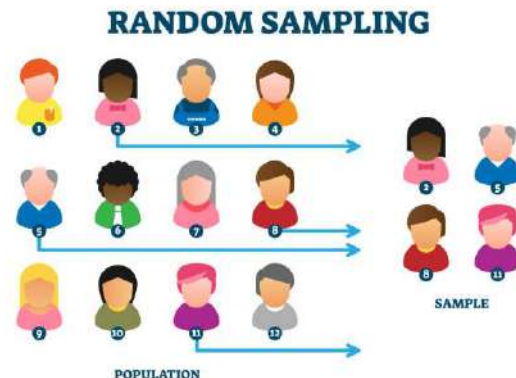
Giới thiệu lấy mẫu ngẫu nhiên

GIỚI THIỆU LẤY MẪU NGẪU NHIÊN

Lấy mẫu tổng thể không khả thi:

- Tổn kém chi phí
- Thời gian
- ...

➤ **Giải pháp**



**Lấy mẫu ngẫu nhiên
(random sampling)**

GIỚI THIỆU LẤY MẪU NGẪU NHIÊN

Lấy mẫu ngẫu nhiên: thu thập đủ thông tin từ một nhóm nhỏ để đưa ra phân tích về tổng thể.

Ví dụ:



Khả thi để thực hiện!

GIỚI THIỆU LẤY MẪU NGẪU NHIÊN

Trong bài học tiếp theo:

- Nói thêm về lấy mẫu.
- Các ý nghĩa thống kê của quá trình lấy mẫu.





5 KIỂM TRA TÍNH ĐÚNG CỦA DỮ LIỆU



NỘI DUNG



Sức mạnh thống kê



Tính toán kích thước mẫu

Một số khái niệm

Ví dụ về kích thước mẫu

SỨC MẠNH THỐNG KÊ

- **Sức mạnh thống kê (statistical power):** xác suất nhận được kết quả có ý nghĩa từ một thử nghiệm hoặc một bài kiểm tra.
- Bài kiểm tra được gọi là có ý nghĩa thống kê (**statistically significant**), tức là kết quả bài kiểm tra không xảy ra ngẫu nhiên mà do nguyên nhân cụ thể.



SỨC MẠNH THỐNG KÊ

Ví dụ

CHIẾN DỊCH
Quảng bá du lịch
phía Nam

KHẢO SÁT
Người dân có
thích chiến dịch
này không?

Sức mạnh
thống kê: 0.6
(60%)

60% khả năng kết
quả cuộc khảo sát có
ý nghĩa thống kê

60% kết quả của cuộc khảo sát
là đáng tin cậy và 40% là khảo
sát sai

SỨC MẠNH THỐNG KÊ

Thông thường, sức mạnh thống kê tối thiểu **0,8 (80%)** được gọi là có ý nghĩa thống kê

Kích thước mẫu càng lớn
➤ Cơ hội có ý nghĩa
thống kê càng lớn

Tiếp theo: tính toán kích
thước mẫu



NỘI DUNG



Sức mạnh thống kê



Tính toán kích thước mẫu

Một số khái niệm

Ví dụ về kích thước mẫu

KÍCH THƯỚC MẪU – KHÁI NIỆM

- **Lấy mẫu ngẫu nhiên:** thu thập thông tin từ một nhóm nhỏ để đưa ra phân tích về tổng thể.
- Kích thước lấy mẫu bao nhiêu là đủ?
- Nói cách khác: lấy mẫu ngẫu nhiên thế nào, để thống kê vẫn có ý nghĩa?



KÍCH THƯỚC MẪU – KHÁI NIỆM

Trong các phần tiếp theo, chúng tôi xin phép trình bày khái niệm và ý nghĩa, mà tạm thời bỏ qua phần toán học và xem như đó là phần mở rộng cho người học.

Kích thước mẫu phụ thuộc:

- Độ tin cậy
- Kích thước tổng thể
- Giới hạn sai số



KÍCH THƯỚC MẪU – KHÁI NIỆM

- **Kích thước tổng thể:** tổng số mẫu trong tập dữ liệu.
- **Độ tin cậy (confidence level):** xác suất mẫu dữ liệu phản ánh chính xác tổng thể.
- **Giới hạn sai số (margin of error):** mức độ tối đa mà kết quả phân tích mẫu sẽ khác với kết quả của tổng thể.



NỘI DUNG



Sức mạnh thống kê



Tính toán kích thước mẫu

Một số khái niệm

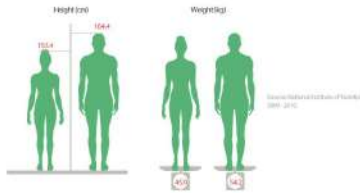
Ví dụ về kích thước mẫu

KÍCH THƯỚC MẪU – VÍ DỤ

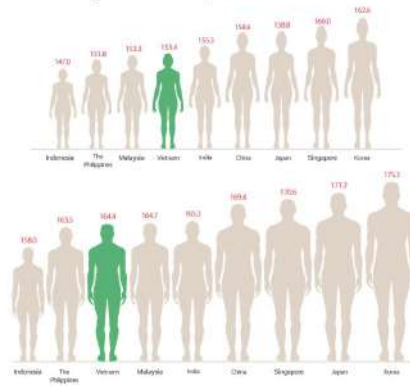
Khảo sát: chiều cao trung bình người việt là 160cm, với:

- Độ tin cậy là 95%
- Giới hạn sai số: ϵ
 - Nghĩa là: 95% chiều cao của người Việt nằm trong khoảng $[160 - \epsilon, 160 + \epsilon]$.

Height and weight of Vietnamese, age 20-24



Height of Vietnamese compared to other Asian countries



KÍCH THƯỚC MẪU – VÍ DỤ

Khảo sát về chiều cao của sinh viên trong trường đại học

Kích thước tổng thể: 500

Giới hạn sai số: 5%

Độ tin cậy: 95%

Kích thước mẫu: 218

- Khảo sát tối thiểu 218 sinh viên. Thì kết quả cũng đúng với 500 sinh viên (với độ tin cậy 95% và sai số 5%).

KÍCH THƯỚC MẪU – VÍ DỤ

Nếu thay đổi “giới hạn sai số” xuống còn 3%.

Kích thước tổng thể: 500

Giới hạn sai số: 3%

Độ tin cậy: 95%

Kích thước mẫu: 341

➤ Cần nhiều điểm dữ liệu hơn: 341 điểm (so với 218 điểm)

KÍCH THƯỚC MẪU – VÍ DỤ

Lưu ý: “giới hạn sai số” và “độ tin cậy” không cần phải có tổng là 100.

Người học có thể tính kích thước mẫu qua các công cụ:

- <http://www.raosoft.com/samplesize.html>
- <https://www.surveymonkey.com/mp/sample-size-calculator/>





6 ĐÁNH GIÁ SAI SỐ CỦA DỮ LIỆU



NỘI DUNG



Giới thiệu về giới hạn sai số



Tính toán giới hạn sai số



Tổng kết

NỘI DUNG



Giới thiệu về giới hạn sai số



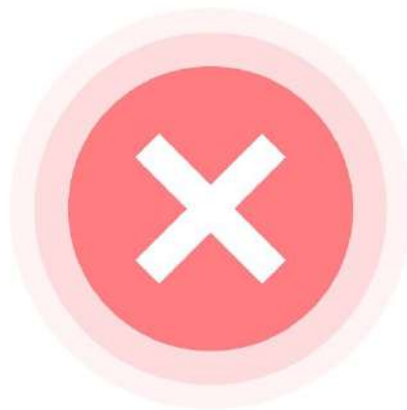
Tính toán giới hạn sai số



Tổng kết

GIỚI THIỆU VỀ GIỚI HẠN SAI SỐ

Giới hạn sai số (margin of error): mức độ tối đa mà kết quả phân tích mẫu sẽ khác với kết quả của tổng thể.



GIỚI THIỆU VỀ GIỚI HẠN SAI SỐ

Ví dụ

Khảo sát: du khách có thích du lịch tại Việt Nam không?

- Kết quả: 60% yêu thích.
- Giới hạn sai số: 10%

Ý nghĩa:
Nếu khảo sát toàn bộ du khách,
kết quả thuộc khoảng 50% - 70%

GIỚI THIỆU VỀ GIỚI HẠN SAI SỐ

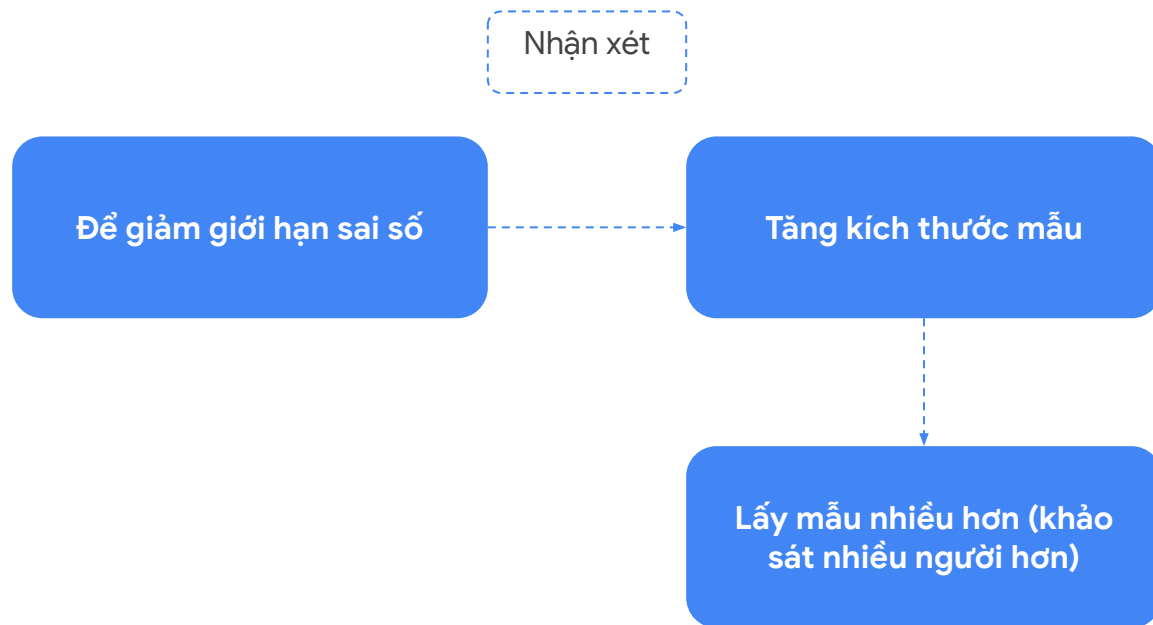
Ví dụ (tiếp theo)

- Kết quả: 60% yêu thích.
- Giới hạn sai số: 10%
- Độ tin cậy: 95%

Ý nghĩa:
nếu khảo sát toàn bộ du khách, xác
suất 95% kết quả sẽ thuộc khoảng
50% - 70%



GIỚI THIỆU VỀ GIỚI HẠN SAI SỐ



NỘI DUNG



Giới thiệu về giới hạn sai số



Tính toán giới hạn sai số



Tổng kết

TÍNH TOÁN GIỚI HẠN SAI SỐ

Tính: giới hạn sai số

Kích thước tổng thể

Kích thước mẫu

Độ tin cậy

Giới hạn sai số

- Tiếp theo, chúng tôi xin phép trình bày khái niệm và ý nghĩa, mà tạm thời bỏ qua phân toán học và xem như đó là phần mở rộng cho người học.

TÍNH TOÁN GIỚI HẠN SAI SỐ

Ví dụ: kiểm tra tính hiệu quả của loại thuốc mới

Kích thước tổng thể:
80 000 000 người

Kích thước mẫu:
500 người

Độ tin cậy:
99 %

Giới hạn sai số:
5.77 %

Xác định mức độ sai số của khảo sát

NỘI DUNG



Giới thiệu về giới hạn sai số



Tính toán giới hạn sai số



Tổng kết

TỔNG KẾT

Khái niệm được học:

- Kiểm tra tính toàn vẹn của dữ liệu
- Mối quan hệ giữa dữ liệu và mục tiêu kinh doanh
- Lấy mẫu ngẫu nhiên
- Kích thước mẫu
- Giới hạn sai số



TIẾP THEO

Tiếp theo: phương pháp làm sạch dữ liệu





7 GIỚI THIỆU VỀ DỮ LIỆU SẠCH



NỘI DUNG



Giới thiệu dữ liệu sạch



Tầm quan trọng của dữ liệu sạch



Nhận dạng và sửa chữa dữ liệu không sạch

GIỚI THIỆU DỮ LIỆU SẠCH

Dữ liệu sạch (clean data): đầy đủ, chính xác và liên quan đến vấn đề đang giải quyết



Dữ liệu không sạch (dirty data): không đầy đủ, chưa chính xác hoặc không liên quan đến vấn đề đang giải quyết.



GIỚI THIỆU DỮ LIỆU SẠCH

Ví dụ: dữ liệu không sạch

Nhập thiếu

Họ tên	Nghề nghiệp
Nguyễn Văn A	Giáo viên
Trần Thị B	
Võ Văn C	Ngân hàng

Còn được gọi là
dữ liệu NULL

Nhập sai

“Bến Te”

Sai định dạng

Thu nhập (triệu đồng)
25
100%
30

GIỚI THIỆU DỮ LIỆU SẠCH

Ví dụ: Tính tổng thu nhập

Thu nhập (triệu đồng)
25
100%
30

Dữ liệu sai

Dừng chương trình

NỘI DUNG



Giới thiệu dữ liệu sạch

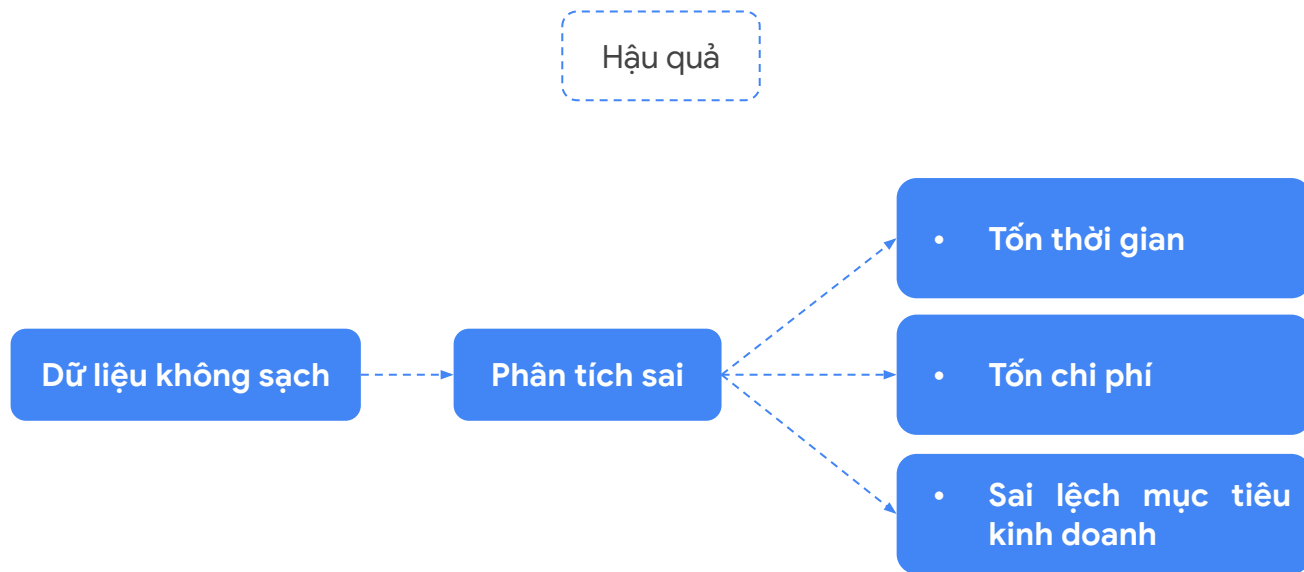


Tầm quan trọng của dữ liệu sạch



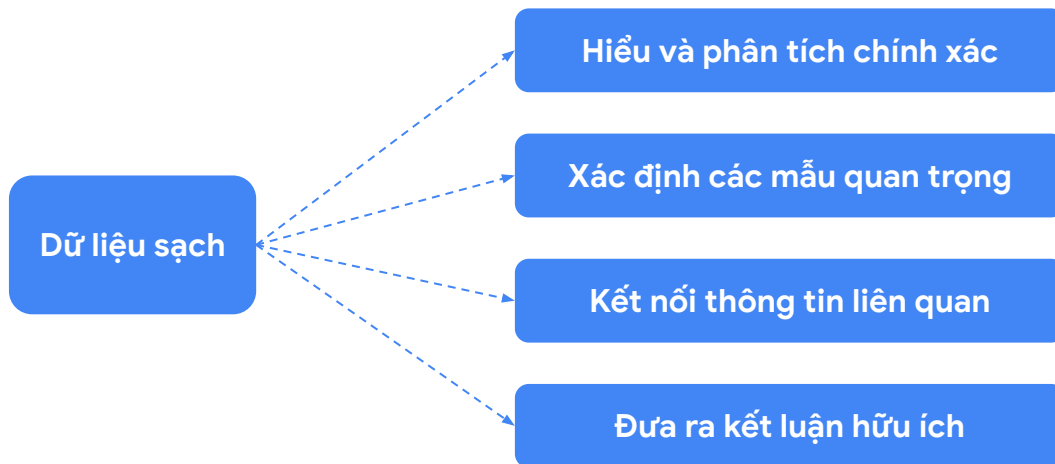
Nhận dạng và sửa chữa dữ liệu không sạch

TẦM QUAN TRỌNG CỦA DỮ LIỆU SẠCH



TẦM QUAN TRỌNG CỦA DỮ LIỆU SẠCH

Dữ liệu sạch là cực kỳ quan trọng để phân tích hiệu quả



TẦM QUAN TRỌNG CỦA DỮ LIỆU SẠCH

Ví dụ: đếm tổng số khách hàng

Họ tên	Nghề nghiệp
Nguyễn Văn A	Giáo viên
Trần Thị B	Quân đội
Võ Văn C	Ngân hàng
Nguyễn Văn A	Giáo viên

Tổng: 4 khách hàng

Loại bỏ dữ liệu trùng

Họ tên	Nghề nghiệp
Nguyễn Văn A	Giáo viên
Trần Thị B	Quân đội
Võ Văn C	Ngân hàng

Tổng: 3 khách hàng

➤ Nhận xét: dữ liệu không sạch, dẫn đến kết quả phân tích sai

TẦM QUAN TRỌNG CỦA DỮ LIỆU SẠCH

Nhờ sự trợ giúp:

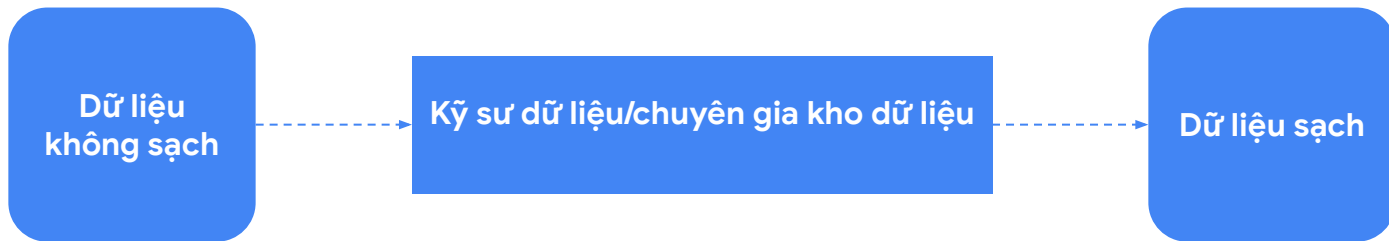
- **Kỹ sư dữ liệu (Data engineer)** chuyển đổi dữ liệu thành định dạng hữu ích để phân tích và xây dựng cơ sở hạ tầng đáng tin cậy.
- Ngoài ra, **Kỹ sư dữ liệu (Data engineer)**: phát triển, duy trì và kiểm tra cơ sở dữ liệu, bộ xử lý dữ liệu và các hệ thống liên quan



TẦM QUAN TRỌNG CỦA DỮ LIỆU SẠCH

Nhờ sự trợ giúp:

- **Chuyên gia kho dữ liệu (Data warehousing specialists)** phát triển quy trình và thủ tục để lưu trữ và tổ chức dữ liệu hiệu quả.
- Ngoài ra, **Chuyên gia kho dữ liệu** đảm bảo rằng dữ liệu có sẵn, an toàn và được sao lưu để tránh mất mát



NỘI DUNG



Giới thiệu dữ liệu sạch



Tầm quan trọng của dữ liệu sạch



Nhận dạng và sửa chữa dữ liệu không sạch

NHẬN DẠNG VÀ SỬA CHỮA DỮ LIỆU KHÔNG SẠCH

Dữ liệu không sạch phổ biến

- Dữ liệu không chính xác
- Dữ liệu không nhất quán
- Dữ liệu bị trống
- Dữ liệu trùng lặp



NHẬN DẠNG VÀ SỬA CHỮA DỮ LIỆU KHÔNG SẠCH

- **Giải pháp:** kiểm tra trong từ điển hoặc cơ sở dữ liệu.
 - Dữ liệu không chính xác

Ví dụ

Nhập sai

“Bến Te”

“Bến Tre”



NHẬN DẠNG VÀ SỬA CHỮA DỮ LIỆU KHÔNG SẠCH

Dữ liệu không nhất quán

➤ **Giải pháp:** chỉnh sửa lại cho đúng định dạng được yêu cầu.

Ví dụ: dữ liệu không nhất quán (% và triệu đồng)

Thu nhập (triệu đồng)	Định dạng loại dữ liệu	Thu nhập (triệu đồng)
25		25
1000%		10
30		30

NHẬN DẠNG VÀ SỬA CHỮA DỮ LIỆU KHÔNG SẠCH

Dữ liệu bị trống

➤ **Giải pháp:** truy xuất thông tin và tìm nội dung thiếu.

Ví dụ:

Họ tên	Nghề nghiệp
Nguyễn Văn A	Giáo viên
Trần Thị B	
Võ Văn C	Ngân hàng

truy xuất thông tin thiếu

Họ tên	Nghề nghiệp
Nguyễn Văn A	Giáo viên
Trần Thị B	Quân đội
Võ Văn C	Ngân hàng

NHẬN DẠNG VÀ SỬA CHỮA DỮ LIỆU KHÔNG SẠCH

Dữ liệu trùng lặp

➤ **Giải pháp:** tìm và xóa dữ liệu trùng

Ví dụ: “Nguyễn Văn A” lặp 2 lần.

Họ tên	Nghề nghiệp
Nguyễn Văn A	Giáo viên
Trần Thị B	Quân đội
Võ Văn C	Ngân hàng
Nguyễn Văn A	Giáo viên



Họ tên	Nghề nghiệp
Nguyễn Văn A	Giáo viên
Trần Thị B	Quân đội
Võ Văn C	Ngân hàng

NHẬN DẠNG VÀ SỬA CHỮA DỮ LIỆU KHÔNG SẠCH

Một số loại dữ liệu không sạch khác:

- Dữ liệu gán nhãn sai
- Dữ liệu có chiều dài không nhất quán
- ...



NHẬN DẠNG VÀ SỬA CHỮA DỮ LIỆU KHÔNG SẠCH

Ví dụ

- Dữ liệu gắn nhãn sai



Mèo



Mèo



Mèo

NHẬN DẠNG VÀ SỬA CHỮA DỮ LIỆU KHÔNG SẠCH

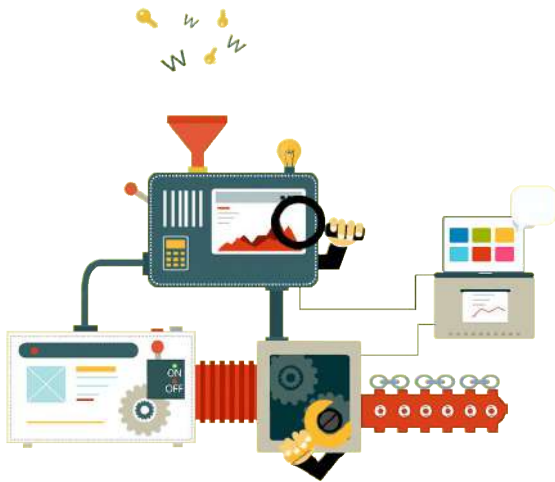
Ví dụ: mã bưu chính ở Việt Nam bắt buộc có 6 chữ số

- **Trường dữ liệu (field):** thông tin từ một hàng hoặc cột của bảng tính
 - **Chiều dài trường (field length):** bao nhiêu ký tự có thể nhập vào một trường
- Chiều dài trường dữ liệu không nhất quán



NHẬN DẠNG VÀ SỬA CHỮA DỮ LIỆU KHÔNG SẠCH

- Các kỹ thuật trên, còn được gọi: xác thực dữ liệu (data validation).
- **Xác thực dữ liệu (Data validation)** kiểm tra tính chính xác và chất lượng dữ liệu trước khi thêm hoặc nhập.





8 BẮT ĐẦU LÀM SẠCH DỮ LIỆU



NỘI DUNG



Một số kỹ thuật làm sạch dữ liệu



Gom nhóm và làm sạch dữ liệu từ nguồn

MỘT SỐ KỸ THUẬT LÀM SẠCH DỮ LIỆU

- Bảng tính (spreadsheet): hỗ trợ làm sạch.
- Làm sạch phụ thuộc vào loại dữ liệu.
- **Sao lưu (backup) trước khi xử lý, phòng trường hợp gặp lỗi.**



MỘT SỐ KỸ THUẬT LÀM SẠCH DỮ LIỆU

Một số kỹ thuật phổ biến:

- Xóa dữ liệu không cần thiết
- Sửa lỗi chính tả
- Xóa định dạng



MỘT SỐ KỸ THUẬT LÀM SẠCH DỮ LIỆU

Một số kỹ thuật phổ biến:

- Xóa dữ liệu không cần thiết
- Sửa lỗi chính tả
- Xóa định dạng

Dữ liệu không cần thiết (unwanted data): dữ liệu không phù hợp với vấn đề đang giải quyết cần xóa.

Lợi ích:

- Dữ liệu gọn nhẹ.
- Dễ quan sát và xử lý nhanh.



MỘT SỐ KỸ THUẬT LÀM SẠCH DỮ LIỆU

Một số kỹ thuật phổ biến:

- Xóa dữ liệu không cần thiết
- Sửa lỗi chính tả
- Xóa định dạng

Ví dụ: phân tích khách hàng hiện tại (năm 2022)

➤ xóa khách hàng trong quá khứ.

Họ tên	Năm hoạt động
Nguyễn Văn A	2021
Trần Thị B	2020
Võ Văn C	2022
Phan Văn D	2022

Họ tên	Năm hoạt động
Võ Văn C	2022
Phan Văn D	2022

MỘT SỐ KỸ THUẬT LÀM SẠCH DỮ LIỆU

Một số kỹ thuật phổ biến:

- Xóa dữ liệu không cần thiết
- Sửa lỗi chính tả
- Xóa định dạng

Cách thực hiện

1. Xóa thủ công:

- Ví dụ: tìm khách hàng không phải năm 2022 và xóa.

2. Dùng bảng tính:

- Bôi đen bảng cần chọn.
- Chọn “tạo bộ lọc”
- Chọn theo năm 2022



MỘT SỐ KỸ THUẬT LÀM SẠCH DỮ LIỆU

Một số kỹ thuật phổ biến:

- Xóa dữ liệu không cần thiết
 - Sửa lỗi chính tả
 - Xóa định dạng



Xóa khoảng trắng dư thừa

- **Khoảng trắng dư thừa (extra space)** gây ra kết quả không mong muốn bạn sắp xếp, lọc, tìm kiếm trong dữ liệu.
- **Ví dụ:** hai kết quả tìm kiếm sau đây sẽ khác nhau

“Bến Tre” và “Bến Tre ”

MỘT SỐ KỸ THUẬT LÀM SẠCH DỮ LIỆU

Một số kỹ thuật phổ biến:

- Xóa dữ liệu không cần thiết
- Sửa lỗi chính tả
- Xóa định dạng

Cách xóa khoảng trắng dư thừa

1. Xóa thủ công:

- tìm khoảng trắng và xóa

2. Dùng bảng tính:

- Dùng lệnh TRIM
- Ví dụ:

Câu lệnh

=trim("Bến Tre ")

Kết quả

"Bến Tre"

MỘT SỐ KỸ THUẬT LÀM SẠCH DỮ LIỆU

Một số kỹ thuật phổ biến:

- Xóa dữ liệu không cần thiết
- **Sửa lỗi chính tả**
- Xóa định dạng



Cách thực hiện

1. Xóa thủ công:

- Tìm lỗi chính tả và sửa

2. Dùng bảng tính:

- Bôi đen ô hoặc cột cần kiểm tra.
- Chọn “Công cụ” → chọn “Chỉnh tả và ngữ pháp” → chọn “Kiểm tra chính tả”

MỘT SỐ KỸ THUẬT LÀM SẠCH DỮ LIỆU

Một số kỹ thuật phổ biến:

- Xóa dữ liệu không cần thiết
- **Sửa lỗi chính tả**
- Xóa định dạng

Ví dụ: sửa lỗi chính tả tự động với bảng tính

Thay đổi Te thành: ×

Bến Te	Tre	Thay đổi
	Tre	Bỏ qua
		Thêm vào từ điển

MỘT SỐ KỸ THUẬT LÀM SẠCH DỮ LIỆU

Một số kỹ thuật phổ biến:

- Xóa dữ liệu không cần thiết
- Sửa lỗi chính tả
- Xóa định dạng

Xóa định dạng giúp cho bảng tính dễ theo dõi, phân tích hơn.

Ví dụ:

Họ tên	Năm hoạt động
Nguyễn Văn A	2021
Trần Thị B	2020
Võ Văn C	2022
Phan Văn D	2022

Họ tên	Năm hoạt động
Nguyễn Văn A	2021
Trần Thị B	2020
Võ Văn C	2022
Phan Văn D	2022

NỘI DUNG



Một số kỹ thuật làm sạch dữ liệu

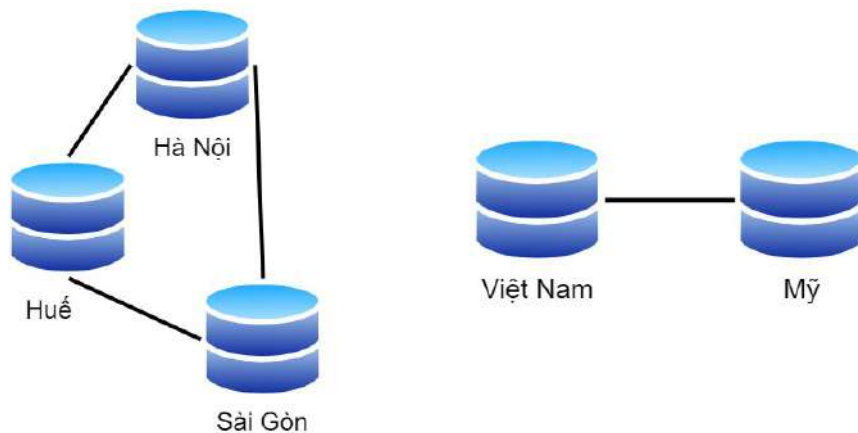


Gom nhóm và làm sạch dữ liệu từ nguồn

GOM NHÓM VÀ LÀM SẠCH DỮ LIỆU TỪ NHIỀU NGUỒN

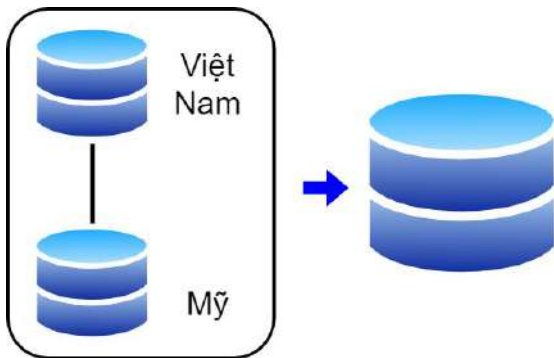
Dữ liệu có thể ở nhiều nguồn.

Ví dụ:



GOM NHÓM VÀ LÀM SẠCH DỮ LIỆU TỪ NHIỀU NGUỒN

Sự hợp nhất (merger): hợp hai hay nhiều tổ chức thành một tổ chức mới.



Hợp nhất dữ liệu (data merging):
quá trình hợp hai hoặc nhiều tập dữ liệu thành một tập dữ liệu duy nhất.

GOM NHÓM VÀ LÀM SẠCH DỮ LIỆU TỪ NHIỀU NGUỒN

Hợp nhất dữ liệu: thường không nhất quán

Ví dụ: định dạng ngày:

ở Việt Nam: ngày / tháng / năm

ở Mỹ: tháng / ngày / năm

Ví dụ: định dạng địa chỉ:

ở TP HCM: phường, quận, thành

ở tỉnh: xã, huyện, tỉnh

GOM NHÓM VÀ LÀM SẠCH DỮ LIỆU TỪ NHIỀU NGUỒN

Khi gom nhóm, đặt các câu hỏi:

- Đặt câu hỏi giúp tránh dư thừa và để xác nhận rằng các tập dữ liệu đó tương thích.
- Khả năng tương thích (compatibility) mô tả hai hoặc nhiều bộ dữ liệu có thể hoạt động cùng nhau tốt như thế nào.



GOM NHÓM VÀ LÀM SẠCH DỮ LIỆU TỪ NHIỀU NGUỒN

- **Giải pháp cho sự không nhất quán:**
 - Tìm hiểu vấn đề và sửa chữa phù hợp.
 - Sử dụng hỗ trợ các công cụ: bảng tính, SQL, ...





9 LÀM SẠCH DỮ LIỆU VỚI BẢNG TÍNH NÂNG CAO



NỘI DUNG



Tính năng làm sạch dữ liệu trong bảng tính



Tối ưu hóa quy trình làm sạch dữ liệu



Làm sạch dữ liệu với bảng tính năng cao

TÍNH NĂNG LÀM SẠCH DỮ LIỆU TRONG BẢNG TÍNH

Kỹ thuật làm sạch dữ liệu

- Định dạng có điều kiện
- Xóa dữ liệu trùng lặp
- Định dạng ngày tháng
- Phân tách văn bản thành các cột



TÍNH NĂNG LÀM SẠCH DỮ LIỆU TRONG BẢNG TÍNH

Một số kỹ thuật phổ biến:

- Định dạng có điều kiện
 - Xóa dữ liệu trùng lặp
 - Định dạng ngày tháng
 - Phân tách văn bản thành các cột

Định dạng có điều kiện (conditional formatting): thay đổi cách các ô xuất hiện khi các giá trị đáp ứng điều kiện cụ thể.

Ví dụ: xác định các ô có điều kiện là “ô trống”

Việt Nam
Thái Lan

→ Thỏa điều kiện

TÍNH NĂNG LÀM SẠCH DỮ LIỆU TRONG BẢNG TÍNH

Một số kỹ thuật phổ biến:

- Định dạng có điều kiện
 - Xóa dữ liệu trùng lặp
 - Định dạng ngày tháng
 - Phân tách văn bản thành các cột

Việt Nam
Thái Lan

Xác định ô trống với bảng tính:

- Bôi đen bảng tính.
- Chọn “Định dạng” → “Định dạng có điều kiện”
- Tại mục “Định dạng ô nếu” chọn “Trống”.
- Chọn “Đã xong”.

Kết quả: bảng tính sẽ tô màu các ô theo định dạng là ô trống.

TÍNH NĂNG LÀM SẠCH DỮ LIỆU TRONG BẢNG TÍNH

Một số kỹ thuật phổ biến:

- Định dạng có điều kiện
- Xóa dữ liệu trùng lặp
- Định dạng ngày tháng
- Phân tách văn bản thành các cột

Xóa dữ liệu trùng lặp (Remove duplicates):
công cụ tự động tìm kiếm và xóa các ô trùng lặp khỏi bảng tính.

Ví dụ: xóa tên người trùng lặp

Họ và tên
Nguyễn Văn A
Trần Thị B
Võ Văn C
Nguyễn Văn A

TÍNH NĂNG LÀM SẠCH DỮ LIỆU TRONG BẢNG TÍNH

Một số kỹ thuật phổ biến:

- Định dạng có điều kiện
- Xóa dữ liệu trùng lặp
- Định dạng ngày tháng
- Phân tách văn bản thành các cột

Họ và tên
Nguyễn Văn A
Trần Thị B
Võ Văn C

Xóa dữ liệu trùng lặp với bảng tính:

- Bôi đen bảng tính.
- Chọn “Dữ liệu” → “Dọn sạch dữ liệu” → “Xóa bản trùng lặp”
- Chọn “Dữ liệu có hàng tiêu đề”
- Chọn “Xóa hàng trùng lặp”

Kết quả: bảng tính sẽ xóa họ và tên người trùng lặp.

TÍNH NĂNG LÀM SẠCH DỮ LIỆU TRONG BẢNG TÍNH

Một số kỹ thuật phổ biến:

- Định dạng có điều kiện
- Xóa dữ liệu trùng lặp
- Định dạng ngày tháng
- Phân tách văn bản thành các cột

Ví dụ: Một số dữ liệu, có thể có ngày tháng không thống nhất.

Ngày sinh
20/09/1999
30/10/2000
ngày 10 tháng 11 năm 2000

TÍNH NĂNG LÀM SẠCH DỮ LIỆU TRONG BẢNG TÍNH

Một số kỹ thuật phổ biến:

- Định dạng có điều kiện
- Xóa dữ liệu trùng lặp
- Định dạng ngày tháng
- Phân tách văn bản thành các cột

Ngày sinh
20/09/1999
30/10/2000
10/11/2000

Định dạng ngày tháng với bảng tính:

- Bôi đen bảng tính.
- Chọn “Định dạng” → “Số” → “Ngày”

Kết quả: bảng tính sẽ định dạng ngày dạng: ngày/tháng/năm

TÍNH NĂNG LÀM SẠCH DỮ LIỆU TRONG BẢNG TÍNH

Một số kỹ thuật phổ biến:

- Định dạng có điều kiện
- Xóa dữ liệu trùng lặp
- Định dạng ngày tháng
- Phân tách văn bản thành các cột



Loading Segments

Phân tách (split): chia một chuỗi văn bản dựa vào dấu phân cách (Delimiter) và đặt mỗi phân đoạn vào một ô mới.

Chuỗi văn bản (text string): các ký tự trong một ô, thường bao gồm các chữ cái.

Chuỗi con (Substring): chuỗi nhỏ nằm bên trong chuỗi văn bản.

TÍNH NĂNG LÀM SẠCH DỮ LIỆU TRONG BẢNG TÍNH

Một số kỹ thuật phổ biến:

- Định dạng có điều kiện
- Xóa dữ liệu trùng lặp
- Định dạng ngày tháng
- Phân tách văn bản thành các cột

Sử dụng bảng tính:

- Bôi đen bảng tính.
- Chọn “Dữ liệu” → “Phân tách văn bản thành các cột”.
- Chọn ký tự phân tách là “dấu cách”

Ví dụ: tách họ và tên ra.

Họ và tên
Nguyễn Văn A
Trần Thị B
Võ Văn C

Họ và tên		
Nguyễn	Văn	A
Trần	Thị	B
Võ	Văn	C

TÍNH NĂNG LÀM SẠCH DỮ LIỆU TRONG BẢNG TÍNH

Một số kỹ thuật phổ biến:

- Định dạng có điều kiện
- Xóa dữ liệu trùng lặp
- Định dạng ngày tháng
- Phân tách văn bản thành các cột

Để tách văn bản, ta còn có thể dùng hàm SPLIT của bảng tính.

Cú pháp hàm: SPLIT(văn bản, dấu phân cách)

Ví dụ: Nguyễn Văn A, được lưu ở ô A1.

Câu lệnh: =SPLIT(A1, “ ”)

Họ và tên
Nguyễn Văn A

Họ và tên		
Nguyễn	Văn	A

NỘI DUNG



Tính năng làm sạch dữ liệu trong bảng tính



Tối ưu hóa quy trình làm sạch dữ liệu



Làm sạch dữ liệu với bảng tính năng cao

TỐI ƯU HÓA QUY TRÌNH LÀM SẠCH DỮ LIỆU

Hàm (function): tập hợp các bước thực hiện phép tính cụ thể bằng cách sử dụng dữ liệu trong bảng tính.

- Mỗi hàm có cú pháp riêng.

Cú pháp (syntax): cấu trúc xác định trước gồm thông tin được yêu cầu và vị trí của nó trong hàm.



TỐI ƯU HÓA QUY TRÌNH LÀM SẠCH DỮ LIỆU

Một số hàm của bảng tính

- COUNTIF
- LEN
- LEFT
- RIGHT
- MID
- CONCATENATE



TỐI ƯU HÓA QUY TRÌNH LÀM SẠCH DỮ LIỆU

Một số hàm của bảng tính

- COUNTIF
- LEN
- LEFT
- RIGHT
- MID
- CONCATENATE

Chức năng: trả về số ô phù hợp với điều kiện được chỉ định.

Cú pháp: COUNTIF(bảng tính, "điều kiện")

Ví dụ: đến số người thu nhập lớn hơn hoặc bằng 20 triệu.

Câu lệnh: =COUNTIF(A2:A5, ">=20")

Thu nhập (Triệu VND)
10
20
30
25

Kết quả trả về: 3

TỐI ƯU HÓA QUY TRÌNH LÀM SẠCH DỮ LIỆU

Một số hàm của bảng tính

- COUNTIF
- **LEN**
- LEFT
- RIGHT
- MID
- CONCATENATE

Chức năng: đếm số ký tự của chuỗi văn bản.

Cú pháp: LEN(ô hoặc bảng tính)

Ví dụ: đếm xem ô A1 có bao nhiêu ký tự

Câu lệnh: LEN(A1)

Nguyễn Văn A

Kết quả trả về: 12

TỐI ƯU HÓA QUY TRÌNH LÀM SẠCH DỮ LIỆU

Một số hàm của bảng tính

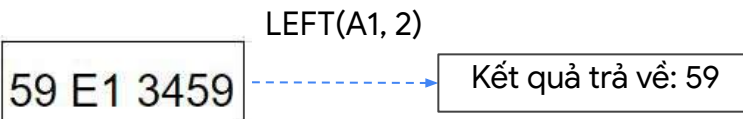
- COUNTIF
- LEN
- **LEFT**
- RIGHT
- MID
- CONCATENATE

Chức năng: trả về một tập hợp ký tự từ phía bên trái chuỗi văn bản.

Cú pháp: LEFT(ô hoặc bảng tính, số ký tự)

Ví dụ: biển số xe của Việt Nam

Câu lệnh: LEFT(A1, 2)



TỐI ƯU HÓA QUY TRÌNH LÀM SẠCH DỮ LIỆU

Một số hàm của bảng tính

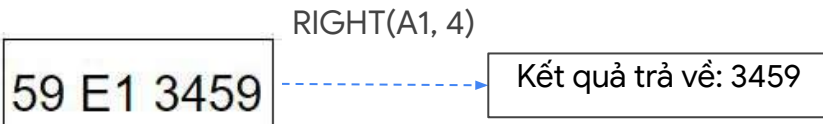
- COUNTIF
- LEN
- LEFT
- **RIGHT**
- MID
- CONCATENATE

Chức năng: trả về một tập hợp ký tự từ phía bên phải chuỗi văn bản.

Cú pháp: RIGHT(ô hoặc bảng tính, số ký tự)

Ví dụ: biển số xe của Việt Nam

Câu lệnh: RIGHT(A1, 4)



TỐI ƯU HÓA QUY TRÌNH LÀM SẠCH DỮ LIỆU

Một số hàm của bảng tính

- COUNTIF
- LEN
- LEFT
- RIGHT
- **MID**
- CONCATENATE

Chức năng: trả về một tập hợp ký tự ở giữa chuỗi văn bản

Cú pháp: MID(ô hoặc bảng tính, ký tự bắt đầu, độ dài)

Ví dụ: biển số xe của Việt Nam

Câu lệnh: MID(A1, 4, 2)

59 E1 3459

MID(A1, 4, 2)

Kết quả trả về: E1

TỐI ƯU HÓA QUY TRÌNH LÀM SẠCH DỮ LIỆU

Một số hàm của bảng tính

- COUNTIF
- LEN
- LEFT
- RIGHT
- MID
- **CONCATENATE**

Chức năng: kết hợp hai hoặc nhiều chuỗi văn bản với nhau.

Cú pháp: CONCATENATE(chuỗi 1, chuỗi 2, ...)

Ví dụ: biển số xe của Việt Nam

Câu lệnh: CONCATENATE(A1, A2, A3)



NỘI DUNG



Tính năng làm sạch dữ liệu trong bảng tính



Tối ưu hóa quy trình làm sạch dữ liệu



Làm sạch dữ liệu với bảng tính năng cao

LÀM SẠCH DỮ LIỆU VỚI BẢNG TÍNH NÂNG CAO

Kỹ thuật làm sạch dữ liệu nâng cao:

- VLOOKUP
- Trực quan hóa
- Ánh xạ dữ liệu



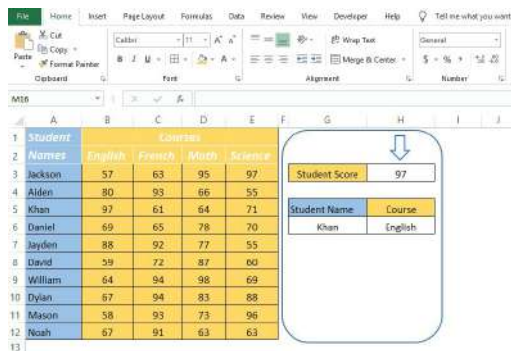
LÀM SẠCH DỮ LIỆU VỚI BẢNG TÍNH NÂNG CAO

Kỹ thuật làm sạch dữ liệu nâng cao:

- VLOOKUP
- Trực quan hóa
- Ánh xạ dữ liệu

Cú pháp: VLOOKUP(khóa tìm kiếm, ô hoặc bảng tính, chỉ mục, false)

VLOOKUP (Vertical Lookup): hàm tìm kiếm giá trị nhất định trong một cột để trả về thông tin tương ứng.



The screenshot shows an Excel spreadsheet with a table of student scores. The table has columns for Student Names, English, French, Math, and Science. A VLOOKUP formula is entered in cell G3, and the result, 97, is displayed in cell H3. A callout box highlights the formula and the result.

Student Names	English	French	Math	Science
Jackson	57	63	95	97
Aiden	80	93	66	55
Khan	97	61	64	71
Daniel	69	65	78	70
Jayden	88	92	77	55
David	59	72	87	60
William	64	94	98	69
Dylan	67	94	83	88
Mason	58	93	73	96
Noah	67	91	63	63

Formula: =VLOOKUP("Khan", A3:D12, 2, FALSE)

Result: 97

LÀM SẠCH DỮ LIỆU VỚI BẢNG TÍNH NÂNG CAO

Kỹ thuật làm sạch dữ liệu nâng cao:

- VLOOKUP
- Trực quan hóa
- Ánh xạ dữ liệu

Ví dụ: cho hai bảng dữ liệu. Ta tính lương của mỗi người nhận được.

Câu lệnh: =VLOOKUP(C3:C6, \$F\$5:\$G\$7, 2, false)

➤ Ta được kết quả

Họ tên	Chức vụ
Nguyễn Văn A	GD
Trần Thị B	TP
Võ Văn C	NV
Trương Thị D	NV

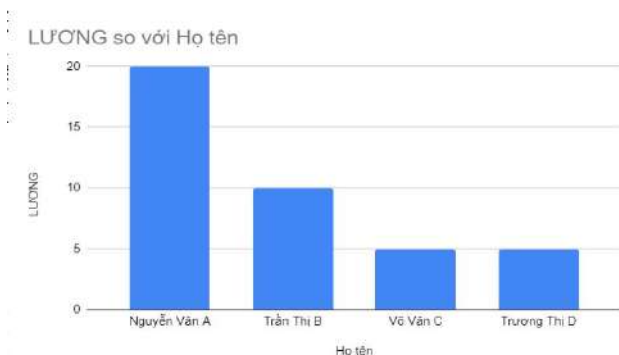
	F	G
4	CHỨC VỤ	LƯƠNG
5	GD	20
6	TP	10
7	NV	5

Họ tên	Chức vụ	LƯƠNG
Nguyễn Văn A	GD	20
Trần Thị B	TP	10
Võ Văn C	NV	5
Trương Thị D	NV	5

LÀM SẠCH DỮ LIỆU VỚI BẢNG TÍNH NÂNG CAO

Kỹ thuật làm sạch dữ liệu nâng cao:

- VLOOKUP
- **Trực quan hóa**
- Ánh xạ dữ liệu



Trực quan hóa: biểu diễn số liệu thành đồ thị. Giúp dễ dàng quan sát hơn.

Ví dụ: biểu diễn biểu đồ cột bảng lương (ở slide trước).

Trực quan hóa với bảng tính:

- Chọn bảng dữ liệu.
- Chọn “chèn” → “Biểu đồ”.
- Chọn loại biểu đồ: “Biểu đồ cột”

Kết quả trực quan hóa

LÀM SẠCH DỮ LIỆU VỚI BẢNG TÍNH NÂNG CAO

Kỹ thuật làm sạch dữ liệu nâng cao:

- VLOOKUP
- Trực quan hóa
- **Ánh xạ dữ liệu**

Ánh xạ dữ liệu (Data mapping): quá trình kết hợp (matching) các trường từ cơ sở dữ liệu này sang cơ sở dữ liệu khác.



Chương 3



10 SỬ DỤNG SQL LÀM SẠCH DỮ LIỆU



NỘI DUNG



Giới thiệu nội dung chương



Giới thiệu về SQL



Ưu điểm SQL khi xử lý dữ liệu lớn

GIỚI THIỆU NỘI DUNG CHƯƠNG

Nhắc lại:

- Dữ liệu sạch
- Làm sạch dữ liệu với bảng tính.

- Làm sạch dữ liệu với SQL
- Truy vấn dữ liệu với SQL
- Áp dụng SQL biến đổi dữ liệu



NỘI DUNG



Giới thiệu nội dung chương



Giới thiệu về SQL



Ưu điểm SQL khi xử lý dữ liệu lớn

GIỚI THIỆU VỀ SQL

SQL (structured query language) – ngôn ngữ truy vấn có cấu trúc, dùng thao tác với cơ sở dữ liệu (CSDL):

- Tốc độ truy vấn của SQL nhanh.
- SQL dùng truy vấn trên CSDL quan hệ.
- **CSDL quan hệ (Relational Database):** cơ sở dữ liệu chứa các bảng có thể được kết nối với nhau tạo ra các quan hệ.



NỘI DUNG



Giới thiệu nội dung chương



Giới thiệu về SQL



Ưu điểm SQL khi xử lý dữ liệu lớn

ƯU ĐIỂM SQL XỬ LÝ DỮ LIỆU LỚN

Điểm chung SQL và Bảng tính

- Thực hiện các phép tính
- Áp dụng công thức có sẵn
- Kết hợp dữ liệu từ nhiều bảng (join data)



ƯU ĐIỂM SQL XỬ LÝ DỮ LIỆU LỚN

Ưu điểm của SQL:

- Tốc độ truy vấn nhanh
- Thao tác với CSDL lớn
- Có thể lấy thông tin từ các nguồn khác nhau trong CSDL
- Sử dụng dữ liệu trên nhiều nơi (và cả đám mây)



ƯU ĐIỂM SQL XỬ LÝ DỮ LIỆU LỚN

- Một số biến thể (Dialects) của SQL như: Transact-SQL, PostgreSQL, ..
- SQL chuẩn và biến thể gần giống nhau.
- Một số trường hợp, biến thể của SQL phù hợp với loại phần mềm/loại dữ liệu đặc biệt.





11 TRUY VẤN VỚI SQL



NỘI DUNG



Truy vấn với SQL



Sử dụng truy vấn để làm sạch

TRUY VẤN VỚI SQL

Truy vấn (query): yêu cầu đưa vào CSDL để yêu cầu CSDL làm việc.

- Làm quen với SELECT và các mở rộng.
- Dùng CSDL khách hàng, với các thuộc tính:

(ID, ten_khach_hang, thu_nhap, que_quan)



TRUY VẤN VỚI SQL

SELECT FORM

- Chỉ định dữ liệu muốn trích xuất từ CSDL.
- Ví dụ: xem tất cả khách hàng.
- Câu lệnh: `SELECT * FROM Khách_hang;`

	ID	ten_khach_hang	thu_nhap	que_quan
1	100	Nguyen Van A	1000000	Ben Tre
2	101	Tran Thi B	20000000	TP HCM
3	102	Vo Van C	30000000	TP HCM
4	100	Nguyen Van A	1000000	Ben Tre
5	104	Ho Van E	-10000000	Ha Noi

1. TRUY VẤN VỚI SQL

INSERT INTO

Chèn các bản ghi mới trong một bảng.

Cú pháp:

INSERT INTO tên_bảng (cột 1, cột 2, cột 3, ...)

VALUES (giá trị 1, giá trị 2, giá trị 3, ...);

Hoặc:

INSERT INTO tên_bảng

VALUES (giá trị 1, giá trị 2, giá trị 3, ...);



TRUY VẤN VỚI SQL

INSERT INTO

Ví dụ: thêm dòng gồm giá trị: (105, Pham Minh D, 10000000, Hue)

Câu lệnh:

`INSERT INTO` Khach_hang

`VALUES` (105, 'Pham Minh D', 10000000, 'Hue');

	ID	ten_khach_hang	thu_nhap	que_quan
1	100	Nguyen Van A	1000000	Ben Tre
2	101	Tran Thi B	20000000	TP HCM
3	102	Vo Van C	30000000	TP HCM
4	100	Nguyen Van A	1000000	Ben Tre
5	104	Ho Van E	-10000000	Ha Noi
6	105	Pham Minh D	10000000	Hue

TRUY VẤN VỚI SQL

UPDATE

Sửa đổi các bản ghi hiện có trong một bảng.

Cú pháp:

UPDATE tên_bảng

SET cột 1 = giá trị 1, cột 2 = giá trị 2, ...

WHERE điều kiện



TRUY VẤN VỚI SQL

Dùng để tạo bảng mới trong CSDL.

Cú pháp:

```
CREATE TABLE tên_bảng  
( cột1 kiểu_dữ_liệu,  
  cột2 kiểu_dữ_liệu,  
  cột3 kiểu_dữ_liệu, )
```

CREATE TABLE

hoặc

CREATE TABLE IF NOT EXISTS

TRUY VẤN VỚI SQL

Dùng để xóa bảng trong CSDL

- Cú pháp:
`DROP TABLE` tên_bảng

`DROP TABLE`

hoặc

`DROP TABLE IF EXISTS`

TRUY VẤN VỚI SQL

Lưu ý khi truy vấn:

- Không tạo ra một bảng dữ liệu. Chỉ lưu trữ kết quả truy vấn trong bộ nhớ tạm..
- Để lưu kết quả truy vấn, cần tải nó xuống dưới dạng bảng tính hoặc lưu kết quả vào một bảng mới.



NỘI DUNG



Truy vấn với SQL

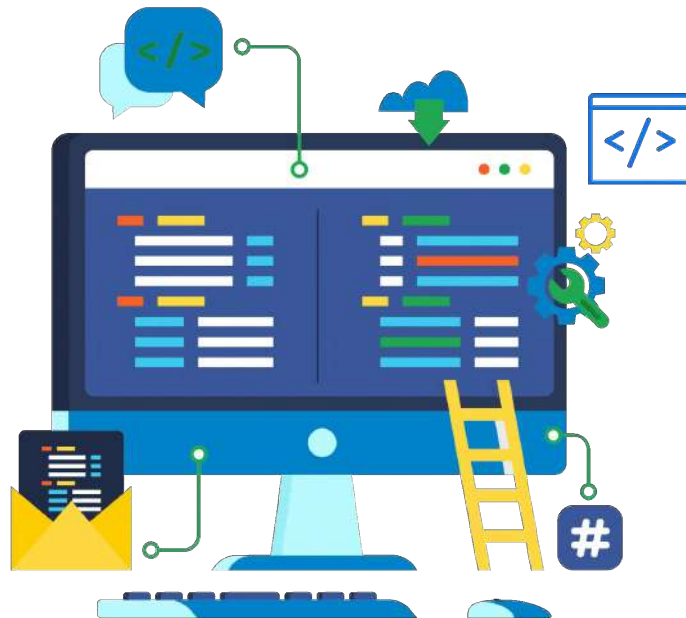


Sử dụng truy vấn để làm sạch

SỬ DỤNG TRUY VẤN ĐỂ LÀM SẠCH DỮ LIỆU

Một số cách làm sạch dữ liệu SQL

- Xóa dữ liệu trùng
- Xử lý chuỗi
- Kiểm tra điều kiện



SỬ DỤNG TRUY VẤN ĐỂ LÀM SẠCH DỮ LIỆU

Một số cách làm sạch dữ liệu SQL

- Xóa dữ liệu trùng
- Xử lý chuỗi
- Kiểm tra điều kiện

- Ví dụ: xem tất cả ID khách hàng
- Câu lệnh: `SELECT ID FROM`
Khach_hang

ID '100' xuất hiện 2 lần

	ID
1	100
2	101
3	102
4	100
5	104
6	105

SỬ DỤNG TRUY VẤN ĐỂ LÀM SẠCH DỮ LIỆU

Một số cách làm sạch dữ liệu SQL

- Xóa dữ liệu trùng
- Xử lý chuỗi
- Kiểm tra điều kiện

- Sử dụng từ khóa **DISTINCT** trong câu lệnh **SELECT**
- Câu lệnh: **SELECT DISTINCT ID FROM Khách_hang**

	ID
1	100
2	101
3	102
4	104
5	105

SỬ DỤNG TRUY VẤN ĐỂ LÀM SẠCH DỮ LIỆU

Một số cách làm sạch dữ liệu SQL

- Xóa dữ liệu trùng
- **Xử lý chuỗi**
- Kiểm tra điều kiện

Ví dụ:

- Khách hàng có quê quán ở TP HCM: “Tran Thi B” và “Vo Van C”
- Cột que_quan của “Tran Thi B” bị dư thừa khoảng trắng.

Giải pháp: dùng từ khóa **TRIM**

	ID	ten_khach_hang	thu_nhap	que_quan
1	100	Nguyen Van A	1000000	Ben Tre
2	101	Tran Thi B	20000000	TP HCM
3	102	Vo Van C	30000000	TP HCM
4	100	Nguyen Van A	1000000	Ben Tre
5	104	Ho Van E	-10000000	Ha Noi
6	105	Pham Minh D	10000000	Hue

SỬ DỤNG TRUY VẤN ĐỂ LÀM SẠCH DỮ LIỆU

Một số cách làm sạch dữ liệu SQL

- Xóa dữ liệu trùng
- Xử lý chuỗi
- Kiểm tra điều kiện

Câu lệnh:

```
SELECT * FROM Khách_hang  
WHERE TRIM (que_quan) = 'TP HCM'
```

	ID	ten_khach_hang	thu_nhap	que_quan
1	101	Tran Thi B	20000000	TP HCM
2	102	Vo Van C	30000000	TP HCM

SỬ DỤNG TRUY VẤN ĐỂ LÀM SẠCH DỮ LIỆU

Một số cách làm sạch dữ liệu SQL

- Xóa dữ liệu trùng
- Xử lý chuỗi
- Kiểm tra điều kiện

Giới thiệu thêm, SQL còn có một số hàm xử lý chuỗi:

- SUBSTR (hoặc SUBSTRING): trích xuất một số ký tự từ một chuỗi.
- CONCAT: nối hai hoặc nhiều chuỗi với nhau.
- LOWER: chuyển đổi một chuỗi thành chữ thường.



SỬ DỤNG TRUY VẤN ĐỂ LÀM SẠCH DỮ LIỆU

Một số cách làm sạch dữ liệu SQL

- Xóa dữ liệu trùng
- Xử lý chuỗi
- Kiểm tra điều kiện

Ví dụ:

- Có khách hàng có thu nhập là số âm.
- Cần lọc ra các khách hàng đó.

	ID	ten_khach_hang	thu_nhap	que_quan
1	100	Nguyen Van A	1000000	Ben Tre
2	101	Tran Thi B	20000000	TP HCM
3	102	Vo Van C	30000000	TP HCM
4	100	Nguyen Van A	1000000	Ben Tre
5	104	Ho Van E	-10000000	Ha Noi
6	105	Pham Minh D	10000000	Hue

SỬ DỤNG TRUY VẤN ĐỂ LÀM SẠCH DỮ LIỆU

Một số cách làm sạch dữ liệu SQL

- Xóa dữ liệu trùng
- Xử lý chuỗi
- Kiểm tra điều kiện

Câu lệnh:

```
SELECT * FROM Khách_hang WHERE thu_nhap < 0
```

	ID	ten_khach_hang	thu_nhap	que_quan
1	104	Ho Van E	-10000000	Ha Noi

SỬ DỤNG TRUY VẤN ĐỂ LÀM SẠCH DỮ LIỆU

Tổng kết

- Cơ bản về SQL
- Sử dụng SQL để làm sạch dữ liệu
- Tiếp theo: truy vấn và làm sạch nâng cao với SQL





12 TRUY VẤN VÀ LÀM SẠCH VỚI SQL NÂNG CAO



MỘT SỐ HÀM SQL NÂNG CAO

ORDER BY

CAST

COALESCE

CONCAT

LENGTH



MỘT SỐ HÀM SQL NÂNG CAO

Ta sẽ dùng CSDL khách hàng, với các thuộc tính:

(ID, ten_khach_hang, thu_nhap, que_quan)

	ID	ten_khach_hang	thu_nhap	que_quan
1	100	Nguyen Van A	1000000	Ben Tre
2	101	Tran Thi B	20000000	TP HCM
3	102	Vo Van C	30000000	TP HCM
4	100	Nguyen Van A	1000000	Ben Tre
5	104	Ho Van E	-10000000	Ha Noi
6	105	Pham Minh D	10000000	Hue
7	107	Pham Van F	10000000	NULL

MỘT SỐ HÀM SQL NÂNG CAO

ORDER BY

CAST

COALESCE

CONCAT

LENGTH

Chức năng: sắp xếp kết quả theo thứ tự tăng dần hoặc giảm dần

Cú pháp:

SELECT cột1, cột2, ...

FROM tên_bảng

ORDER BY cột1, cột1, ... ASC|DESC;

Ví dụ: sắp xếp danh sách theo thu nhập tăng dần.

Câu lệnh: **SELECT** * **FROM** Khách_hang **ORDER BY** thu_nhap **ASC**;

	ID	ten_khach_hang	thu_nhap	que_quan
1	104	Ho Van E	-10000000	Ha Noi
2	100	Nguyen Van A	1000000	Ben Tre
3	100	Nguyen Van A	1000000	Ben Tre
4	105	Pham Minh D	10000000	Hue
5	107	Pham Van F	10000000	NULL
6	101	Tran Thi B	20000000	TP HCM
7	102	Vo Van C	30000000	TP HCM

MỘT SỐ HÀM SQL NÂNG CAO

ORDER BY

CAST

COALESCE

CONCAT

LENGTH

Chức năng: chuyển đổi giá trị (thuộc bất kỳ kiểu dữ liệu nào) thành một kiểu dữ liệu được chỉ định.

Cú pháp: `CAST(biểu_thức AS kiểu_dữ_liệu);`

Ví dụ: `CAST('2017-08-25' AS datetime)`

2017-08-25

Một số kiểu dữ liệu thông dụng của SQL:

- INT: kiểu số nguyên. Ví dụ: 100
- FLOAT: kiểu số thực. Ví dụ: 10.5
- DATETIME: kiểu ngày tháng (định dạng YYYY-MM-DD). Ví dụ: 2020-12-29

MỘT SỐ HÀM SQL NÂNG CAO

ORDER BY

CAST

COALESCE

CONCAT

LENGTH

Chức năng: trả về biểu thức có giá trị khác NULL đầu tiên trong những biểu thức được truyền vào.

Cú pháp: `COALESCE (bieuthuc_1, bieuthuc_2,... bieuthuc_n)`

Ví dụ: truy vấn que_quan của khách hàng. Trường hợp que_quan NULL, trả về tên của người đó.

Câu lệnh: `SELECT COALESCE (que_quan, ten_khach_hang) FROM Khach_hang;`

	ID	ten_khach_hang	thu_nhap	que_quan
1	104	Ho Van E	-10000000	Ha Noi
2	100	Nguyen Van A	1000000	Ben Tre
3	100	Nguyen Van A	1000000	Ben Tre
4	105	Pham Minh D	10000000	Hue
5	107	Pham Van F	10000000	NULL
6	101	Tran Thi B	20000000	TP HCM
7	102	Vo Van C	30000000	TP HCM

	(No column name)
1	Ben Tre
2	TP HCM
3	TP HCM
4	Ben Tre
5	Ha Noi
6	Hue
7	Pham Van F

MỘT SỐ HÀM SQL NÂNG CAO

ORDER BY

CAST

COALESCE

CONCAT

LENGTH

Chức năng: nối hai hoặc nhiều chuỗi với nhau.

Cú pháp: `CONCAT(string1, string2, ..., string_n)`

Ví dụ: `SELECT CONCAT('Khoa hoc', ' du lieu');`

'Khoa hoc du lieu'

MỘT SỐ HÀM SQL NÂNG CAO

ORDER BY

CAST

COALESCE

CONCAT

LENGTH

Chức năng: trả về số ký tự của chuỗi.

Cú pháp: `LENGTH(chuỗi)`

Ví dụ: `SELECT LENGTH('SQL');`

3

TỔNG KẾT:

- Giới thiệu công thức và hàm SQL.
- Ví dụ SQL làm sạch dữ liệu.

CHƯƠNG TIẾP THEO:

- Xác minh quá trình làm sạch
- Báo cáo quá trình làm sạch





13 XÁC MINH QUÁ TRÌNH LÀM SẠCH DỮ LIỆU



NỘI DUNG



Xác minh và báo cáo kết quả làm sạch



Dữ liệu làm sạch và dữ liệu mong đợi



Ví dụ quá trình xác minh

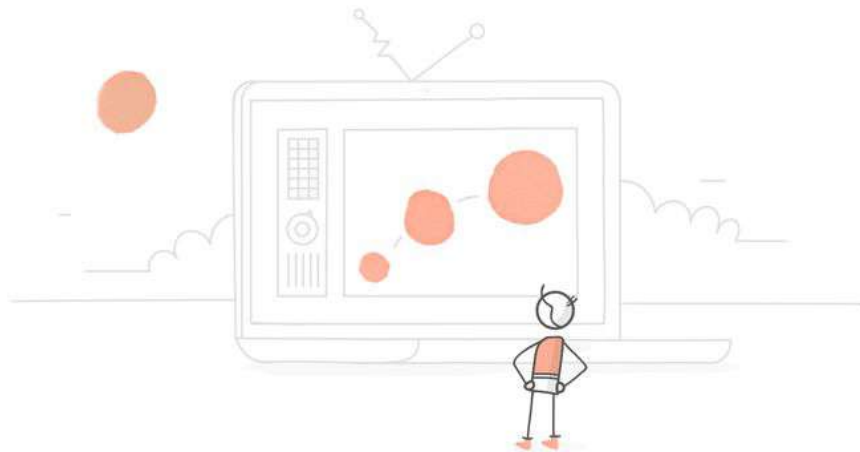
XÁC MINH VÀ BÁO CÁO KẾT QUẢ LÀM SẠCH

Nhắc lại:

- Làm sạch dữ liệu với SQL

Tiếp theo:

- Xác minh và báo cáo về quá trình làm sạch dữ liệu



XÁC MINH VÀ BÁO CÁO KẾT QUẢ LÀM SẠCH

Sự xác minh (verification): xác nhận quá trình làm sạch dữ liệu đã được thực hiện tốt và dữ liệu là chính xác và đáng tin cậy.

Kiểm tra lần hai: quá trình làm sạch là chính xác

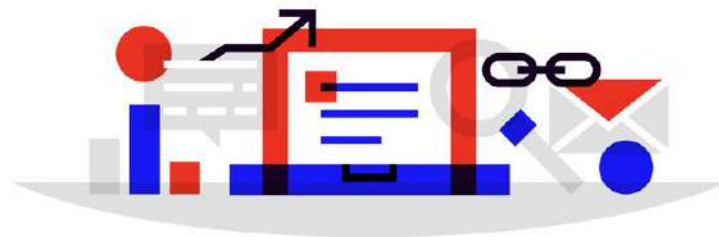
Dữ liệu đáng tin cậy và phù hợp với mục tiêu kinh doanh



XÁC MINH VÀ BÁO CÁO KẾT QUẢ LÀM SẠCH

Nhiệm vụ quan trọng của việc báo cáo:

- Hiểu về quá trình làm sạch dữ liệu.
- Xây dựng lòng tin với khách hàng và đối tác.
- Chuẩn bị sẵn sàng cho bước phân tích tiếp.



XÁC MINH VÀ BÁO CÁO KẾT QUẢ LÀM SẠCH

Các cách báo cáo:

- Tạo bảng ghi thay đổi (changelog)
- Ghi chép tài liệu quá trình làm sạch



NỘI DUNG



Xác minh và báo cáo kết quả làm sạch



Dữ liệu làm sạch và dữ liệu mong đợi



Ví dụ quá trình xác minh

DỮ LIỆU LÀM SẠCH VÀ DỮ LIỆU MONG ĐỢI

Cho một cái nhìn toàn cảnh hơn, quan sát bức tranh lớn hơn.

- Xác nhận lại vấn đề:

Vấn đề ban đầu đặt ra?

Mục tiêu kinh doanh?

Dữ liệu và mục tiêu kinh doanh?

DỮ LIỆU LÀM SẠCH VÀ DỮ LIỆU MONG ĐỢI

Quá trình xác minh (verification): nhờ sự trợ giúp của bảng ghi thay đổi.

Có thể tham khảo bảng ghi thay đổi trong quá trình xác minh nếu có sai sót hoặc thắc mắc



DỮ LIỆU LÀM SẠCH VÀ DỮ LIỆU MONG ĐỢI

Kiểm tra lại các vấn đề nghi ngờ?

Ví dụ: thực hiện cuộc khảo sát

Thực hiện bởi 1000 người

Thu về nhiều hơn 1000 kết quả!!!

➤ *Có dấu hiệu sai phạm!*

DỮ LIỆU LÀM SẠCH VÀ DỮ LIỆU MONG ĐỢI

Quá trình xác minh đảm bảo:

- Kết quả phân tích có thể tin được.
- Hạn chế nguy cơ gây sai lầm.
- Tiết kiệm thời gian.



NỘI DUNG



Xác minh và báo cáo kết quả làm sạch



Dữ liệu làm sạch và dữ liệu mong đợi



Ví dụ quá trình xác minh

VÍ DỤ QUÁ TRÌNH XÁC MINH

Một số cách xác minh phổ biến:

- Thủ công
- Dùng bảng tổng hợp (pivot table)
- Dùng câu lệnh CASE



VÍ DỤ QUÁ TRÌNH XÁC MINH

Một số cách xác minh phổ biến:

- Thủ công
- Dùng bảng tổng hợp (pivot table)
- Dùng câu lệnh CASE

Duyệt thủ công dữ liệu và tìm điểm sai.

Ví dụ: lỗi chính tả (Bến Te, thay vì Bến Tre)

Quê quán
TP HCM
Huế
Hà Nội
Bến Te

➤ Sử dụng “Tìm và thay thế” (find and replace)

VÍ DỤ QUÁ TRÌNH XÁC MINH

Một số cách xác minh phổ biến:

- Thủ công
- Dùng bảng tổng hợp (pivot table)
- Dùng câu lệnh CASE

Các bước thực hiện “tìm và thay thế”

- Chọn “Chỉnh sửa” → “Tìm và thay thế”
- Nhập vào ô “Tìm” và ô “Thay thế bằng”
- Chọn “Thay thế tất cả” → “Đã xong”

Tìm

Bến Te

Thay thế bằng

Bến Tre



Quê quán

TP HCM

Huế

Hà Nội

Bến Tre

VÍ DỤ QUÁ TRÌNH XÁC MINH

Một số cách xác minh phổ biến:

- Thủ công
- Dùng bảng tổng hợp (pivot table)
- Dùng câu lệnh CASE

- Ví dụ: lỗi chính tả (Bến Te, thay vì Bến Tre)
- Dùng bảng tổng hợp (pivot table)

Quê quán
TP HCM
Huế
Bến Tre
Bến Te
Huế
TP HCM
TP HCM

VÍ DỤ QUÁ TRÌNH XÁC MINH

Một số cách xác minh phổ biến:

- Thủ công
- Dùng bảng tổng hợp (pivot table)
- Dùng câu lệnh CASE

Dùng bảng tổng hợp:

- Bôi đen dữ liệu.
- Chọn “Chèn” → “Bảng tổng hợp” → “Tạo”
- Chọn “Cột” → “Quê quán”
- Chọn “Giá trị” → “Quê quán”

Ta thấy cột “Bến Te” và “Bến Tre” bất thường, cần được sửa chữa.

Bến Te	Bến Tre	Huế	TP HCM	Grand Total
1	1	2	3	7

VÍ DỤ QUÁ TRÌNH XÁC MINH

Một số cách xác minh phổ biến:

- Thủ công
- Dùng bảng tổng hợp (pivot table)
- Dùng câu lệnh CASE

- Bảng tổng hợp mặc định dùng câu lệnh COUNTA.
- **COUNTA**: đếm tổng số giá trị trong một phạm vi được chỉ định
- Bảng tổng hợp cũng có nhiều hàm khác (COUNT, COUNTIF, ...).



VÍ DỤ QUÁ TRÌNH XÁC MINH

Một số cách xác minh phổ biến:

- Thủ công
- Dùng bảng tổng hợp (pivot table)
- Dùng câu lệnh CASE



Chức năng: đi qua các điều kiện và trả về một giá trị khi thỏa điều kiện

Cú pháp:

CASE

WHEN điều_kiện_1 THEN kết_quả_1

WHEN điều_kiện_2 THEN kết_quả_2

WHEN điều_kiện_n THEN kết_quả_n

ELSE result

END;

VÍ DỤ QUÁ TRÌNH XÁC MINH

Một số cách xác minh phổ biến:

- Thủ công
- Dùng bảng tổng hợp (pivot table)
- Dùng câu lệnh CASE

Quê quán
TP HCM
Huế
Bến Tre
Bến Tre
Huế
TP HCM
TP HCM

Ví dụ: cột que_quan trong bảng Khách_hang

Câu lệnh:

```
SELECT *,  
CASE  
  WHEN que_quan = 'Ben Te' THEN 'Ben Tre'  
  ELSE que_quan  
END  
FROM Khách_hang
```


VÍ DỤ QUÁ TRÌNH XÁC MINH

Hiện tại: đã biết cách dùng bảng tính và SQL để tự động sửa lỗi.

- Tiếp theo: khám phá cách theo dõi các thay đổi.





14 GHI NHẬN KẾT QUẢ CỦA QUÁ TRÌNH LÀM SẠCH



NỘI DUNG



Ghi lại các bước quá trình làm sạch



Tầm quan trọng của tài liệu



Nhận phản hồi từ khách hàng

GHI LẠI CÁC BƯỚC QUÁ TRÌNH LÀM SẠCH

Bài học thảo luận:

- Tại sao theo dõi các thay đổi?
- Cách ghi nhận thay đổi trong quá trình làm sạch.



GHI LẠI CÁC BƯỚC QUÁ TRÌNH LÀM SẠCH

Tài liệu (document): quá trình theo dõi các thay đổi, bổ sung, xóa và các lỗi liên quan đến quá trình làm sạch dữ liệu.



GHI LẠI CÁC BƯỚC QUÁ TRÌNH LÀM SẠCH

Có tài liệu giúp:

- Phục hồi lỗi trong quá trình làm sạch.
- Thông báo cho người khác về thay đổi đã thực hiện.
- Xác định chất lượng của dữ liệu.



GHI LẠI CÁC BƯỚC QUÁ TRÌNH LÀM SẠCH

Bản ghi thay đổi (changelog): tệp chứa danh sách sửa đổi theo thứ tự thời gian được thực hiện đối với một dự án:

- Bảng tính
- SQL
- Phần mềm công ty



GHI LẠI CÁC BƯỚC QUÁ TRÌNH LÀM SẠCH

Tạo bản ghi thay đổi tự động của “Bảng tính”:

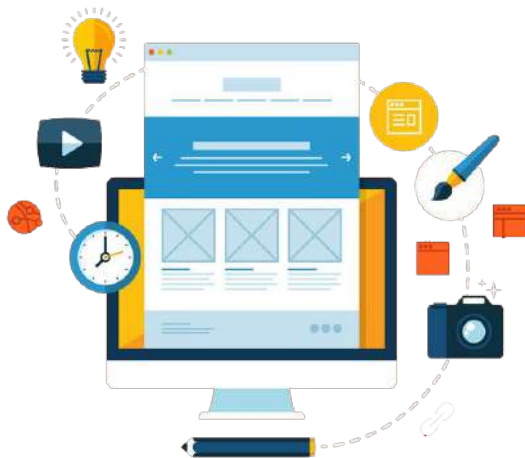
- Chọn “Tập”
- Chọn “Lịch sử phiên bản”
- Chọn “Xem lịch sử phiên bản”



GHI LẠI CÁC BƯỚC QUÁ TRÌNH LÀM SẠCH

Bản ghi thay đổi của SQL:

- Tùy thuộc vào nền tảng: BigQuery, Microsoft sql server management studio, ...



GHI LẠI CÁC BƯỚC QUÁ TRÌNH LÀM SẠCH

- Bản ghi thay đổi: ghi nhận các bản cập nhật theo thời gian.
- Ngoài ra, còn có báo cáo (sẽ được trình bày ở phần sau)



NỘI DUNG



Ghi lại các bước quá trình làm sạch



Tầm quan trọng của tài liệu



Nhận phản hồi từ khách hàng

TẦM QUAN TRỌNG CỦA TÀI LIỆU

Tài liệu (document): theo dõi sự thay đổi.

Tài liệu giúp ích trong tương lai:

- Phân tích dữ liệu giống nhau
- Sửa lỗi giống nhau.



TẦM QUAN TRỌNG CỦA TÀI LIỆU



Lập trình viên/nhà khoa học dữ liệu
có thể hiểu bản ghi thay đổi



Khách hàng

không hiểu bản ghi thay đổi.

- Dựa vào báo cáo.
- Ví dụ: tạo tệp văn bản (file word).

TẦM QUAN TRỌNG CỦA TÀI LIỆU

Ví dụ bản ghi thay đổi

```
SELECT *  
FROM Khach_hang  
ORDER BY thu_nhap ASC;
```

Ví dụ báo cáo

Truy vấn tất cả khách hàng có trong bảng Khach_hang và sắp xếp chúng dựa theo thu nhập tăng dần.

TẦM QUAN TRỌNG CỦA TÀI LIỆU

- Ta còn có thể chú thích trong mã nguồn, để diễn giải mã nguồn. Xem như là bảng ghi thay đổi.
- Chú thích được đặt phía ở giữa dấu /* và */

/* truy vấn khách hàng theo thu nhập tăng dần */

SELECT *

FROM Khách_hang

ORDER BY thu_nhap ASC;

TẦM QUAN TRỌNG CỦA TÀI LIỆU

Tầm quan trọng của tài liệu hoặc báo cáo:

- Chứng minh tính minh bạch về quá trình làm sạch.
- Mọi người đều nắm tình hình.
- Cho khách hàng lòng tin và trách nhiệm



NỘI DUNG



Ghi lại các bước quá trình làm sạch



Tầm quan trọng của tài liệu



Nhận phản hồi từ khách hàng

NHẬN PHẢN HỒI TỪ KHÁCH HÀNG

Một số lợi ích khi nhận phản hồi từ khách hàng:

- Chứng minh: tôi làm đúng.
 - Quá trình làm sạch dữ liệu
- cái nhìn bên trong về kinh doanh.



NHẬN PHẢN HỒI TỪ KHÁCH HÀNG

Xin chúc mừng!

Bạn đã có nền tảng cần thiết để xác minh thành công báo cáo về kết quả làm sạch của mình.



Congratulations!



TỔNG KẾT



NHỮNG Ý CHÍNH CẦN NẮM

- Biết cách kiểm tra tính toàn vẹn của dữ liệu.
- Hiểu các kỹ thuật làm sạch dữ liệu bằng bảng tính.
- Thực hiện được các truy vấn SQL cơ bản để sử dụng trên cơ sở dữ liệu.
- Áp dụng được các hàm cơ bản của SQL để làm sạch và chuyển đổi dữ liệu.
- Mô tả được cách xác minh kết quả của việc làm sạch dữ liệu.
- Trình bày được các yếu tố và tầm quan trọng của việc báo cáo quá trình làm sạch dữ liệu.





THANK YOU

