Tài liệu đọc

Phân Tích Dữ Liệu

Khóa 4: Xử lý và làm sạch dữ liệu

Phần 1: Tài liệu đọc bổ trợ	
Bài đọc 1	Sự toàn vẹn và tuân thủ dữ liệu
	1.1 Sự toàn vẹn dữ liệu Giới thiệu về toàn vẹn dữ liệu Nguyên nhân của sự không toàn vẹn và giải pháp Ví dụ về toàn vẹn dữ liệu Tuân thủ dữ liệu Giới thiệu về tuân thủ dữ liệu Ví dụ về tuân thủ dữ liệu 1.2 Ví dụ về tầm quan trọng của sự toàn vẹn và tuân thủ dữ liệu 1.3 Ràng buộc dữ liệu Nội dung: giới thiệu một số ràng buộc về dữ liệu, định nghĩa và ví dụ 1.4 References
Bài đọc 2	Dữ liệu và mục tiêu kinh doanh
	2.1 Giới thiệu về sự liên quan giữa dữ liệu và mục tiêu kinh doanh 2.2 Ví dụ về sự liên quan giữa dữ liệu và mục tiêu kinh doanh Ví dụ 1 Ví dụ 2 Ví dụ 3 2.3 Những điều cần ghi nhớ 2.4 References

	Mô tả dữ liệu lỗi thời Nguyên nhân của dữ liệu lỗi thời Hậu quả của dữ liệu lỗi thời 5.4 Dữ liệu không đầy đủ Mô tả dữ liệu không đầy đủ Nguyên nhân của dữ liệu không đầy đủ Hậu quả của dữ liệu không đầy đủ 5.5 Dữ liệu không chính xác Mô tả dữ liệu không chính xác Nguyên nhân của dữ liệu không chính xác Hậu quả của dữ liệu không chính xác 5.6 Dữ liệu không nhất quán Mô tả dữ liệu không nhất quán Nguyên nhân của dữ liệu không nhất quán Hậu quả của dữ liệu không nhất quán 5.7 References
Bài đọc 6	Một số lỗi phổ biến khi làm sạch dữ liệu
	6.1 Một số lỗi phổ biến Không kiểm tra lỗi chính tả Không ghi nhận lại quá trình chỉnh sửa Không kiểm tra giá trị bị điền sai Bỏ qua giá trị bị thiếu Chỉ xem xét một phần của dữ liệu Không quan tâm đến mục tiêu kinh doanh Không kiểm tra nguyên nhân gây ra lỗi Không sửa chữa nguyên nhân gây ra lỗi Không sao lưu dữ liệu trước khi làm sạch 6.2 References
Bài đọc 7	Giới thiệu SQL
	7.1 Giới thiệu SQL 7.2 Ví dụ dùng SQL truy vấn dữ liệu 7.3 So sánh SQL và bảng tính 7.4 Các biến thể của SQL 7.5 References

Bài đọc 8	Giới thiệu về BigQuery
	8.1 Giới thiệu về BigQuery 8.2 Hướng dẫn cơ bản tải dữ liệu với BigQuery 8.3 References
Bài đọc 9	Tiêu chí để xác minh dữ liệu sạch
	9.1 Một số lỗi cần được kiểm tra Nguyên nhân gây ra lỗi Dữ liệu NULL Dữ liệu chuỗi nhập sai chính tả Dữ liệu số nhập sai Dư thừa các ký tự hoặc khoảng trắng Dữ liệu bị trùng lặp Kiểu dữ liệu không chính xác Giá trị nhập không nhất quán Nhãn của dữ liệu gây hiểu nhầm Dữ liệu bị cắt bớt Dữ liệu phù hợp với mục tiêu kinh doanh 9.2 Đánh giá lại mục tiêu kinh doanh và dữ liệu Khi đã hoàn thành các nhiệm vụ làm sạch dữ liệu, ta nên xem lại mục tiêu kinh doanh và xác nhận rằng dữ liệu vẫn phù hợp với mục tiêu phân tích được đặt ra từ ban đầu. 9.3 References
Bài đọc 10	Ghi lại thay đổi trong quá trình làm sạch
	10.1 Giới thiệu một số công cụ hỗ trợ việc ghi lại lịch sử thay đổi 10.2 Bảng ghi thay đổi (Changelog) Giới thiệu về changelog Những thông tin thường có của một changelog 10.3 Hệ thống kiểm soát phiên bản (version control system) trong thực tế Giới thiệu về hệ thống kiểm soát phiên bản Các bước của hệ thống kiểm soát phiên bản 10.4 References
Bài đọc 11	Các hàm nâng cao để tăng tốc quá trình làm sạch dữ liệu

Phần 2: Hướng dẫn trả lời câu hỏi - Quiz

Phần 1
TÀI LIỆU ĐỌC BỔ TRỢ

Bài đọc 1: Sự toàn vẹn và tuân thủ dữ liệu

Bài đọc giới thiệu về sự toàn vẹn dữ liệu là gì, nêu lên tầm quan trọng của hai yếu tố trên. Đồng thời đưa ra những ví dụ cụ thể để diễn giải vấn đề trên.

1. Sự toàn vẹn dữ liệu

Giới thiệu về toàn vẹn dữ liệu

Toàn vẹn dữ liệu (tiếng anh: data integrity) là quá trình duy trì và đảm bảo tính chính xác (accuracy), nhất quán (consistency), hoàn chỉnh (completeness) và đáng tin cậy (trustworthiness) của dữ liệu. Toàn vẹn dữ liệu là một khía cạnh quan trọng trong việc thiết kế, cài đặt và sử dụng bất kỳ hệ thống nào nhằm phục vụ việc lưu trữ, xử lý và truy vấn dữ liệu. [1]

Duy trì tính toàn vẹn của dữ liệu có thể giúp bạn và công ty/tổ chức của bạn tiết kiệm được thời gian và chi phí trong quá trình đưa ra quyết định dựa trên dữ liệu. Ví dụ, do bất cẩn dữ liệu của một số người dùng bị xóa bỏ, điều này sẽ dẫn đến sai sót trong quá trình thống kê và phân tích dữ liệu sau này.

Nguyên nhân gây ra sự không toàn vẹn, ví dụ và giải pháp

Trong quá trình phân tích và xử lý dữ liệu, một yếu tố như: nhân bản dữ liệu (data replication), chuyển đổi dữ liệu (data transfer), thao tác dữ liệu (data manipulation) có thể làm mất đi tính toàn vẹn của dữ liệu.

- Nhân bản dữ liệu (data replication) gây ra sự không toàn vẹn dữ liệu.

o Nhân bản dữ liệu là quá trình lưu trữ dữ liệu ở nhiều nơi. Điều này cho phép tất cả người dùng có thể chia sẻ cùng một dữ liệu. Ngoài ra, nhân bản dữ liệu còn đảm bảo lưu trữ bản sao lưu dữ liệu, điều này cần thiết trong trường hợp các tai nạn bất ngờ xảy ra, ta vẫn không bị mất hoàn toàn dữ liệu.

o Bên cạnh đó vẫn còn một số rủi ro trong quá trình nhân bản dữ liệu có thể dẫn đến sự không toàn vẹn. Ví dụ, dữ liệu được nhân bản ở cả Mỹ và Việt Nam. Ở Việt Nam, họ và tên bao gồm: họ, chữ lót, tên. Ở Mỹ, họ và tên bao gồm: first name và last name. Những yếu tố không giống nhau như vậy sẽ dẫn đến dữ liệu không còn nhất quán, từ đó vi phạm tính toàn vẹn của dữ liệu.

o Một giải pháp cho vấn đề này, là thống nhất các thông tin tuân theo một định dạng có sẵn. Hoặc xác thực dữ liệu có cùng định dạng hay chưa trước khi nhập hoặc trước khi phân tích. Để đảm rằng tính toàn vẹn của dữ liệu không bị vi phạm.

- Chuyển đổi dữ liệu (transfer data) gây ra sự không toàn vẹn dữ liệu.

o Chuyển đổi dữ liệu là quá trình thu thập, sao chép và truyền tải dữ liệu từ nơi này sang nơi khác.

o Quá trình chuyển đổi dữ liệu này có thể dẫn đến sự không toàn vẹn dữ liệu. Ví dụ: một công ty có các trụ sở ở Hà Nội và Thành Phố Hồ Chí Minh, công ty này đang thực hiện quá trình tải dữ liệu từ Hà Nội vào Thành Phố Hồ Chí Minh, tuy nhiên do kết nối internet không ổn định, quá trình truyền tải dữ liệu bị ngắt dẫn đến dữ liệu bị mất mát. Điều này ảnh hưởng tính toàn vẹn của dữ liệu.

o Một giải pháp cho vấn đề trên là sau khi thực hiện chuyển đổi dữ liệu, ta cần kiểm tra xem dữ liệu sau khi chuyển đổi đã chính xác và hoàn chỉnh hay chưa.

- Thao tác dữ liệu (data manipulation) gây ra sự không toàn vẹn dữ liệu.

o Thao tác dữ liệu là quá trình thay đổi hoặc chỉnh sửa dữ liệu để làm cho dữ liệu dễ đọc và có tổ chức hơn. Ví dụ: bạn có thể sắp xếp dữ liệu theo thứ tự bảng chữ cái để giúp tìm kiếm dễ dàng hơn.

o Quá trình thao tác dữ liệu có thể dẫn đến sự không toàn vẹn dữ liệu. Ví dụ, một nhà phân tích dữ liệu phát hiện có hai bản ghi trùng lặp của người dùng trong cơ sở dữ liệu và loại bỏ bản ghi trùng lặp đó. Tuy nhiên, hóa ra đó không phải là hai bản ghi trùng lặp mà thật sự là hai người dùng khác nhau có cùng tên và ngày sinh. Lúc này dữ liệu hiện đang bị thiếu và cần được khôi phục cho hoàn chỉnh.

o Một giải pháp cho vấn đề này, là có dữ liệu sao lưu để phòng trường hợp xảy ra lỗi ta có thể khôi phục dữ liệu bất kỳ lúc nào. Ngoài ra, ta có thể dùng nhật ký để theo dõi thời điểm dữ liệu được thêm, sửa đổi hoặc xóa.

2. Ví dụ về tầm quan trọng của sự toàn vẹn dữ liệu

Tiếp theo, ta sẽ minh họa tầm quan trọng của toàn vẹn dữ liệu bằng cách sử dụng ví dụ về dữ liệu của một công ty đa quốc gia.

Định dạng về ngày tháng của các quốc gia khác nhau sẽ khác nhau. Ví dụ, cùng là ngày 10, tháng 12, năm 2020, ta sẽ có nhiều cách thể hiện khác nhau.

- Ở Việt Nam dùng định dạng dd/mm/yyyy, do đó kết quả sẽ là 10/12/2020.
- Ở Mỹ dùng định dạng mm/dd/yyyy, do đó kết quả sẽ là 12/10/2020.
- Ở một số nước, dùng định dạng yyyy-mm-dd, do đó kết quả sẽ là 2020-12-10.

Những sự không thống nhất như vậy sẽ dẫn đến việc dữ liệu không trùng khớp và tính toàn vẹn của dữ liệu sẽ bị ảnh hưởng. Hãy tưởng tượng thời gian được cho là tháng 12 trong khi nó thực sự là vào tháng 10!

Một phân tích tốt phụ thuộc vào tính toàn vẹn của dữ liệu và tính toàn vẹn của dữ liệu thường phụ thuộc vào việc sử dụng một định dạng chung. Vì vậy, điều quan trọng là phải kiểm tra kỹ cách định dạng ngày tháng để đảm bảo đó là tháng 10 hay tháng 12. Do đó, ta cần đưa ra các định dạng chuẩn và yêu cầu tất cả mọi người và hệ thống tuân thủ theo định dạng đã được đưa ra nhằm duy trì tính toàn vẹn của dữ liệu.

Cho dù dữ liệu đến từ nguồn nào đi nữa, hãy luôn đảm bảo kiểm tra xem dữ liệu đó có hợp lệ, đầy đủ và sạch trước khi bạn bắt đầu bất kỳ phân tích nào hay không.

3. Ràng buộc dữ liệu

Khi tiến hành xử lý dữ liệu, bạn sẽ gặp nhiều loại ràng buộc dữ liệu (tiếng anh data constraint) hoặc tiêu chí xác định tính hợp lệ của dữ liệu. Bảng dưới đây cung cấp các định nghĩa và ví dụ về các thuật ngữ về ràng buộc dữ liệu mà bạn có thể gặp phải. [2]

Ràng buộc dữ liệu	Định nghĩa	Ví dụ
Kiểu dữ liệu (tiếng anh: data type)	Giá trị của dữ liệu phải thuộc loại nhất định, ví dụ: số nguyên, số thực, chuỗi, ...	Nếu kiểu dữ liệu là số nguyên, giá trị 1.5 sẽ không được chấp nhận.
Phạm vi dữ liệu (tiếng anh: data range)	Giá trị phải nằm giữa giá trị lớn nhất và giá trị nhỏ nhất được xác định trước.	Nếu phạm vi dữ liệu là [10-20], giá trị 30 sẽ không hợp lệ.
Bắt buộc (tiếng anh: mandatory)	Giá trị không được để rỗng hoặc để trống.	Ví dụ giá trị tuổi của người dùng là bắt buộc, giá trị đó phải được điền vào.
Duy nhất (tiếng anh: unique)	Giá trị không được trùng lặp	Hai người dùng không thể có cùng số điện thoại

Biểu thức chính quy (tiếng anh: regular expression – viết tắt: regex)	Các giá trị phải khớp với một mẫu quy định	Ví dụ, số điện thoại phải có định dạng ### - ### - #### (không cho phép ký tự chữ)
Xác thực trường chéo (tiếng anh: Cross-field validation)	Các điều kiện nhất định cho nhiều trường phải được thỏa mãn	Giá trị là tỷ lệ phần trăm và tổng các giá trị cộng lại tối đa 100%.
Khóa chính (tiếng anh: primary-key)	(Cơ sở dữ liệu) giá trị phải là duy nhất trên mỗi cột	Một bảng cơ sở dữ liệu không được có hai hàng có cùng giá trị khóa chính. Khóa chính là một mã định danh trong cơ sở dữ liệu tham chiếu đến một cột trong đó mỗi giá trị là duy nhất. Thông tin thêm về khóa chính và khóa ngoại được cung cấp ở phần sau của chương trình.
Tập hợp thành phần. (tiếng anh: Set-membership)	(Cơ sở dữ liệu) Giá trị cho một cột phải đến từ một tập hợp các giá trị rời rạc cho sẵn.	Giới tính chỉ có thể lựa chọn 'nam' hoặc 'nữ'
Khóa ngoại (tiếng anh: Foreign-key)	(Cơ sở dữ liệu) giá trị cho một cột phải là giá trị duy nhất đến từ một cột trong bảng khác.	Trong cơ sở dữ liệu về người đóng thuế ở Việt Nam, cột "tỉnh thành" phải là một tỉnh thành hợp lệ với tập hợp các giá trị được xác định trong bảng "Tỉnh thành"
Sự hoàn chỉnh (tiếng anh: Completeness)	Mức độ dữ liệu chứa tất cả các thành phần hoặc thước đo mong muốn	Nếu dữ liệu cho hồ sơ cá nhân yêu cầu tóc và màu mắt và cả hai đều được thu thập, thì dữ liệu được gọi là hoàn chỉnh.
Tính nhất quán (tiếng anh: Consistency)	Mức độ mà dữ liệu giống nhau giữa các	Nếu trong cả hai cơ sở dữ liệu, là "Danh sách khách hàng" và "Bán hàng", tất cả dữ liệu của

	cơ sở dữ liệu khác nhau.	một khách hàng đều thống nhất và chính xác, thì dữ liệu đó gọi là nhất quán.

4. References

[1]https://en.wikipedia.org/wiki/Data_integrity
[2]https://www.coursera.org/learn/process-data/supplement/uWR9E/more-about-data-integrity-and-compliance

Bài đọc 2: Dữ liệu và mục tiêu kinh doanh

Bài đọc giới thiệu về mối liên hệ giữa dữ liệu và mục tiêu kinh doanh, trước khi thực hiện phân tích dữ liệu ta cần xác minh xem mục tiêu kinh doanh là gì, liệu sau khi phân tích dữ liệu ta có trả lời được các câu hỏi mong muốn hay không?

1. Giới thiệu về sự liên quan giữa dữ liệu và mục tiêu kinh doanh

Trong quá trình phân tích dữ liệu, ta có thể đưa ra các phân tích và dự đoán chính xác khi và chỉ khi dữ liệu phù hợp với mục tiêu kinh doanh. Là một nhà phân tích dữ liệu, sự liên quan là điều bạn sẽ cần phải đánh giá. Sự liên kết tốt có nghĩa là dữ liệu có liên quan và có thể giúp bạn giải quyết vấn đề kinh doanh để đạt được mục tiêu kinh doanh nhất định.

Ví dụ, khi cần phân tích về nhu cầu và xu hướng tiêu dùng của khách hàng, thì điều cần thiết ở đây là cơ sở dữ liệu về khách hàng chứ không phải là dữ liệu về nhà sản xuất hay dữ liệu về nhà phân phối.

Trong bài đọc này, bạn sẽ xem xét sự liên quan giữa dữ liệu và mục tiêu kinh doanh dựa trên ba tình huống. Bạn sẽ khám phá cách dữ liệu sạch và các mục tiêu kinh doanh phù hợp có thể giúp bạn đưa ra kết luận chính xác. Ngoài ra, bạn sẽ tìm hiểu cách các biến mới được phát hiện trong quá trình phân tích dữ liệu có thể khiến bạn thiết lập các ràng buộc dữ liệu để bạn có thể giữ cho dữ liệu phù hợp với mục tiêu kinh doanh.

2. Ví dụ về sự liên quan giữa dữ liệu và mục tiêu kinh doanh

Tình huống 1

Dữ liệu sạch (clean data) + phù hợp với mục tiêu kinh doanh (alignment to business objective) = kết luận chính xác (accurate conclusions)

Tình huống kinh doanh: “một công ty muốn phân tích xem, sau bao lâu một mặt hàng sẽ được mua kể từ khi mở bán?”.

Để bắt đầu, nhà phân tích dữ liệu xác minh rằng dữ liệu được xuất sang bảng tính là sạch và xác nhận rằng dữ liệu cần thiết để phân tích đã có (thời gian khách hàng mua sản phẩm). Biết được điều này, nhà phân tích quyết định rằng: có sự phù hợp tốt giữa dữ liệu với mục tiêu kinh doanh. Tất cả những gì còn lại là tìm ra chính xác thời gian khách hàng mua sản phẩm kể từ khi nó được bán.

Dưới đây là các bước xử lý dữ liệu mà nhà phân tích thực hiện cho sản phẩm đồng hồ. (Các bước này sẽ được lặp lại cho mỗi sản phẩm và mỗi người dùng của sản phẩm đó.)

Bước 1:

Xử lý dữ liệu (data processing)	Nguồn dữ liệu (source of data)
Tìm kiếm ngày ra mắt của sản phẩm đồng hồ	Bảng tính về "Sản phẩm"

C	D
Sản phẩm	**Ngày ra mắt**
Đồng hồ	10/12/2020
Xe điện	6/15/2021
Điện thoại	7/20/2021

Kết quả: tháng 10, ngày 12, năm 2020 (lưu ý, ở Microsoft Excel hoặc Google bảng tính, ngày tháng thường được dùng ở định dạng mm/dd/yyyy)

Bước 2:

Xử lý dữ liệu (data processing)	Nguồn dữ liệu (source of data)
Tìm kiếm những khách hàng mua sản phẩm đồng hồ	Bảng tính về "Khách hàng"

Sản phẩm	Khách hàng	Ngày mua hàng
Đồng hồ	Nguyễn Văn A	10/22/2020
Đồng hồ	Trần Thị B	10/30/2020
Đồng hồ	Võ Văn C	10/30/2020

Kết quả: có 3 khách hàng. Nhà phân tích sẽ lần lượt tính với tất cả 3 khách hàng đó. Tuy nhiên ở các ví dụ tiếp theo, chỉ trình bày ví dụ với khách hàng Nguyễn Văn A, những khách hàng khác được tính tương tự.

Bước 3:

Xử lý dữ liệu (data processing)	Nguồn dữ liệu (source of data)
Tìm ngày mua sản phẩm của khách hàng Nguyễn Văn A. Tính khoảng cách từ ngày sản phẩm được ra mắt cho đến ngày Nguyễn Văn A mua hàng.	Tạo ra bảng tính mới, là sự kết hợp giữa bảng tính Khách hàng và bảng tính Sản phẩm

A	B	C	D	E
Sản phẩm	**Ngày ra mắt**	**Khách hàng**	**Ngày mua hàng**	**Số ngày chênh lệch**
Đồng hồ	10/12/2020	Nguyễn Văn A	10/22/2020	10

Kết quả: 10 ngày

Một số lời khuyên trong quá trình xử lý bên trên:

- Lời khuyên 1:

o Trong quá trình trên, nhà phân tích có thể sử dụng hàm VLOOKUP để tra cứu dữ liệu trong các bước 1 và 2 để điền các giá trị vào bảng tính mới ở bước 3. VLOOKUP là một hàm bảng tính tìm kiếm một giá trị nhất định trong một cột để trả về một phần thông tin có liên quan. Sử dụng hàm VLOOKUP có thể tiết kiệm rất nhiều thời gian; nếu không có nó, bạn phải tra cứu ngày tháng và tên theo cách thủ công.

o Tham khảo trang VLOOKUP trong trợ giúp của Google để biết cách sử dụng hàm trong Google Trang tính.

- Lời khuyên 2:

o Trong bước 3 của quy trình trên, nhà phân tích có thể sử dụng hàm DATEDIF để tự động tính toán sự khác biệt giữa các ngày trong cột B và cột D.

o Tham khảo trang DATEDIF Hỗ trợ của Microsoft để biết cách sử dụng hàm trong Excel.

o Tham khảo trang DATEDIF trong Trợ giúp của Google để biết cách sử dụng hàm trong Google Trang tính.

Tình huống 2

Phù hợp với mục tiêu kinh doanh (Alignment to business objective) + làm sạch dữ liệu bổ sung (additional data cleaning) = kết luận chính xác (accurate conclusions)

Tình huống kinh doanh: một công ty phần mềm tổ chức một hội chợ việc làm nhằm quảng cáo hình ảnh công ty và thu hút nhân tài. Công ty muốn xác định các trường đại học có trên 100 sinh viên tham gia chương trình này và xem như đây là các trường có nhiều tiềm năng để thu hút nhân tài.
Nhiệm vụ của nhà phân tích dữ liệu là tìm ra danh sách các trường đó?

Để tham gia hội chợ việc làm, sinh viên cần điền thông tin dưới đây:

Họ và tên	...	Bắt buộc
Email sinh viên	...	Bắt buộc
Trường học	...	Tự chọn

Dữ liệu tham dự hội chợ có vẻ phù hợp với mục tiêu kinh doanh. Nhưng nhà phân tích dữ liệu và người quản lý chương trình quyết định rằng cần phải làm sạch dữ liệu trước khi phân tích. Họ cho rằng cần phải làm sạch dữ liệu vì:

- Tên trường học không phải là một thông tin bắt buộc. Nếu tên trường học để trống, ta có thể tìm được trường học từ địa chỉ email. Ví dụ: nếu địa chỉ email là username@hcmus.edu.vn, thì ta có thể biết đây là trường đại học Khoa Học Tự Nhiên TP HCM.
- Sau buổi hội chợ việc làm, sẽ là phần rút thăm trúng thưởng. Để tăng cơ hội nhận thưởng, nhiều sinh viên đã điền tên của mình nhiều lần, từ đó là cho dữ liệu có thể bị trùng lặp. Để ngăn chặn điều ngày, ta cần kiểm tra địa chỉ email duy nhất để xác định các sinh viên bị trùng nhau. Ví dụ, Nguyễn Tuấn Anh tham gia và điền tên mình nhiều lần tuy nhiên địa chỉ email lại là cùng một người ntanh@hcmus.edu.vn.

Sau khi thực hiện làm sạch, ta có thể phân tích dữ liệu và trả lời các câu hỏi kinh doanh ban đầu đặt ra.

Tình huống 3

Phù hợp với mục tiêu kinh doanh (Alignment to business objective) + các biến mới được phát hiện (newly discovered variables) + ràng buộc (constraints) = kết luận chính xác (accurate conclusions)

Tình huống kinh doanh: một trung tâm dạy tiếng anh muốn tìm hiểu xem, cần bao nhiêu giờ học để tăng 1 band điểm IELTS. Với dữ liệu là thời gian học của các học viên ở trung tâm và điểm kiểm tra IELTS của các học viên trong quá trình học.
Nhà phân tích dữ liệu cho rằng có sự phù hợp tốt giữa dữ liệu có sẵn và mục tiêu kinh doanh vì:

- Danh sách điểm danh học sinh mỗi buổi học và số giờ học được ghi lại.
- Trung tâm thường xuyên tổ chức các buổi thi thử IELTS với nội dung giống kỳ thi thật.

Ràng buộc dữ liệu cho các biến mới:
Sau khi xem xét dữ liệu, nhà phân tích dữ liệu phát hiện ra rằng có những biến số khác cần xem xét. Một số học sinh có các buổi học hàng tuần cố định trong khi các học sinh khác có các buổi học theo lịch trình ngẫu nhiên và có những buổi vắng học, mặc dù tổng số giờ đi học của họ là như nhau. Điều này ảnh hưởng đến quá trình đưa ra dự đoán. Vì vậy, nhà phân tích thêm ràng buộc dữ liệu để chỉ tập trung vào các sinh viên với các buổi học hàng tuần cố định. Việc sửa đổi này giúp có được bức tranh chính xác hơn về thời gian ghi danh và từ đó đạt được sự cải thiện trong việc đánh giá học viên.

3. Những điều cần ghi nhớ

Hy vọng rằng những ví dụ trên sẽ cho bạn cái nhìn tổng quan về những gì cần để đánh giá liệu dữ liệu của bạn có phù hợp với mục tiêu kinh doanh hay không.

- Khi có dữ liệu sạch và phù hợp với mục tiêu kinh doanh, bạn có thể có được thông tin chính xác và đưa ra kết luận.
- Nếu có sự liên kết tốt nhưng dữ liệu cần được làm sạch, hãy làm sạch dữ liệu trước khi bạn thực hiện phân tích.
- Nếu dữ liệu chỉ phù hợp một phần với mục tiêu, hãy nghĩ về cách bạn có thể sửa đổi mục tiêu hoặc sử dụng các ràng buộc dữ liệu để đảm bảo rằng tập hợp con dữ liệu phù hợp hơn với mục tiêu kinh doanh.

4. References

[1]https://www.coursera.org/learn/process-data/supplement/Rhj9e/well-aligned-objectives-and-data

Bài đọc 3: Chuẩn bị dữ liệu cho quá trình làm sạch

Trong quá trình làm sạch và phân tích dữ liệu, đôi khi ta sẽ gặp các vấn đề như: không có dữ liệu, không đủ dữ liệu hoặc dữ liệu bị sai lệch. Bài đọc sẽ nêu lên các vấn đề có thể sẽ gặp phải và chỉ ra những giải pháp khả thi nhằm giải quyết chúng.

1. Không có dữ liệu

Trong thống kê không có dữ liệu là tình trạng không có giá trị nào được lưu trữ đối với biến được quan sát. Có một cách hiểu khác (liên quan đến bài đọc số 2 – dữ liệu và mục tiêu kinh doanh), không có dữ liệu là tình trạng mục tiêu kinh doanh đã được đặt ra tuy nhiên ta không có dữ liệu để có thể phân tích nhằm trả lời cho câu hỏi kinh doanh.
Khi gặp tình trạng này, ta có một số cách khả thi để có thể giải quyết chúng:

Giải pháp khả thi	Ví dụ
Thu thập thêm dữ liệu. Tuy nhiên, việc thu thập đủ lượng dữ liệu cần thiết trong một thời gian ngắn để phân tích là một công việc không dễ dàng. Do đó, ta có thể làm bằng cách: "Thu thập dữ liệu ở quy mô nhỏ để thực hiện phân tích sơ bộ và sau đó yêu cầu thêm thời gian để hoàn thành phân tích sau khi bạn đã thu thập đủ dữ liệu".	Nếu bạn đang khảo sát nhân viên về nhận xét của họ về kế hoạch làm việc và lương thưởng mới. Hãy sử dụng một một lượng nhỏ dữ liệu để phân tích sơ bộ. Sau đó, yêu cầu thêm 3 tuần nữa để thu thập dữ liệu từ tất cả nhân viên.
Tìm dữ liệu thay thế. Nếu không có thời gian để thu thập dữ liệu, hãy thực hiện phân tích bằng cách sử dụng dữ liệu từ các nguồn khác có tính chất tương tự với vấn đề đang xét. *Đây là cách giải quyết phổ biến nhất.*	Nếu bạn muốn phân tích tình hình du lịch của một thành phố cụ thể nhưng lại không có dữ liệu về thành phố đó. Hãy sử dụng dữ liệu từ một thành phố khác có quy mô, dân số và điều kiện tự nhiên tương tự.

2. Không đủ dữ liệu

Không đủ dữ liệu để phân tích là tình trạng: khi câu hỏi kinh doanh được đặt ra, ta có dữ liệu để phân tích, tuy nhiên lượng dữ liệu này quá ít và không đủ để đưa ra kết luận có thể tin cậy được.
Khi gặp tình trạng không đủ dữ liệu để phân tích, ta có một số cách khả thi để có thể giải quyết chúng:

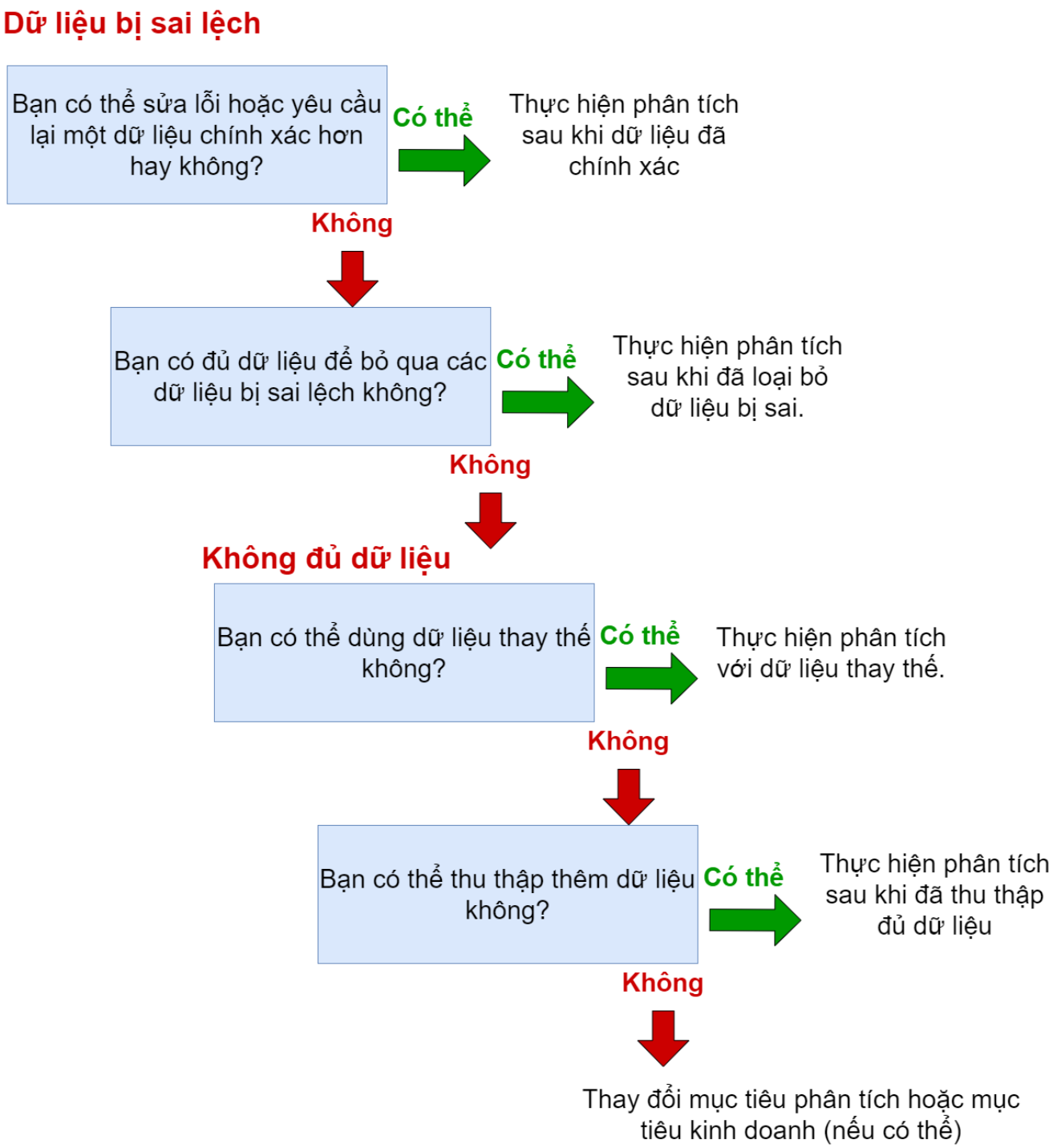
Giải pháp khả thi	Ví dụ
Thực hiện phân tích bằng cách sử dụng dữ liệu thay thế kết hợp cùng với dữ liệu thực tế.	Nếu bạn đang phân tích xu hướng của du lịch Việt Nam, hãy làm cho tập dữ liệu của bạn lớn hơn bằng cách kết hợp dữ liệu của công ty bạn với dữ liệu được công bố bởi các công ty hoặc tổ chức khác có liên quan.
Điều chỉnh phân tích của bạn để phù hợp với dữ liệu bạn đã có.	Nếu bạn thiếu dữ liệu cho thanh niên từ 18 đến 24 tuổi, hãy thực hiện phân tích nhưng lưu ý trong báo cáo của bạn rằng: kết luận này chỉ áp dụng cho người lớn từ 25 tuổi trở lên.

3. Dữ liệu bị sai lệch

Dữ liệu bị sai lệch có nhiều dạng, ví dụ như: dữ liệu bị thiếu các yếu tố chính, dữ liệu không liên quan đến mục đích sử dụng, dữ liệu bị trùng lặp, dữ liệu được nhập vào không chính xác hoặc không đúng định dạng,...
Lưu ý: đôi khi dữ liệu có lỗi có thể là dấu hiệu cảnh báo rằng dữ liệu đó không đáng tin cậy. Hãy cẩn thận khi sử dụng và phân tích.
Khi gặp tình trạng dữ liệu bị sai lệch, ta có một số cách khả thi để có thể giải quyết như sau:

Giải pháp khả thi	Ví dụ
Nếu bạn có dữ liệu sai do các bên liên quan hoặc đối tác hiểu nhầm, hãy thông báo và trình bày với các bên liên quan.	Nếu bạn cần dữ liệu về các công dân nữ và nhận dữ liệu về các công dân nam, hãy trình bày lại nhu cầu của bạn.
Xác định lỗi trong dữ liệu và sửa lại chúng. Nếu có thể, hãy tìm hiểu nguồn gốc gây ra lỗi và sửa chữa chúng.	Nếu dữ liệu của bạn nằm trong bảng tính và có câu lệnh điều kiện khiến các phép tính bị sai, hãy thay đổi câu lệnh điều kiện thay vì chỉ sửa các giá trị đã được tính toán.
Nếu bạn không thể tự sửa lỗi dữ liệu, bạn có thể bỏ qua dữ liệu sai và tiếp tục phân tích nếu kích thước dữ liệu của bạn vẫn đủ lớn và việc bỏ qua một phần của dữ liệu sẽ không gây ra sai lệch hệ thống.	Nếu tập dữ liệu của bạn được dịch từ một ngôn ngữ khác và một số bản dịch không có ý nghĩa hoặc chưa chính xác, hãy bỏ qua các bản dịch không tốt và tiếp tục phân tích các bản dịch còn lại.

Sử dụng cây quyết định sau để nhắc nhở về cách xử lý khi dữ liệu bị sai lệch hoặc không đủ dữ liệu:


Dữ liệu bị sai lệch
Bạn có thể sửa lỗi hoặc yêu cầu lại một dữ liệu chính xác hơn hay không?
Có thể
Thực hiện phân tích sau khi dữ liệu đã chính xác
Không
Bạn có đủ dữ liệu để bỏ qua các dữ liệu bị sai lệch không?
Có thể
Thực hiện phân tích sau khi đã loại bỏ dữ liệu bị sai.
Không
Không đủ dữ liệu
Bạn có thể dùng dữ liệu thay thế không?
Có thể
Thực hiện phân tích với dữ liệu thay thế.
Không
Bạn có thể thu thập thêm dữ liệu không?
Có thể
Thực hiện phân tích sau khi đã thu thập đủ dữ liệu
Không
Thay đổi mục tiêu phân tích hoặc mục tiêu kinh doanh (nếu có thể)

4. References

[1]https://www.coursera.org/learn/process-data/supplement/NQPE4/what-to-do-when-you-find-an-issue-with-your-data

Bài đọc 4: Kích thước của dữ liệu được lấy mẫu

Khi làm sạch và phân tích dữ liệu, ta sẽ đặt câu hỏi: ta cần bao nhiêu điểm dữ liệu cho quá trình phân tích? 10 điểm, 20 điểm hay 1000 điểm.
Bài đọc sẽ nêu ra một số khái niệm liên quan đến quá trình lấy mẫu: lấy mẫu, giới hạn sai số, khoảng tin cậy, độ tin cậy, ý nghĩa thống kê. Một số lưu ý và ví dụ cụ thể về quá trình lấy mẫu.

1. Một số khái niệm

Trước khi tìm hiểu sâu hơn về kích thước mẫu, ta sẽ làm quen với các thuật ngữ và định nghĩa sau:

Thuật ngữ chuyên ngành (terminology)	Định nghĩa
Tổng thể (population)	Toàn bộ nhóm mà bạn quan tâm đến cho nghiên cứu của bạn. Ví dụ: nếu bạn đang khảo sát mọi người trong công ty của mình, tổng thể sẽ là tất cả nhân viên trong công ty của bạn.
Mẫu (sample)	Một tập hợp con của tập tổng thể. Ví dụ, nếu công ty của bạn quá lớn để khảo sát từng cá nhân, bạn có thể khảo sát một mẫu nhỏ hơn và mẫu này có nhiệm vụ đại diện cho tổng thể công ty.
Giới hạn sai số (margin of error)	Vì một mẫu được sử dụng để đại diện cho một tổng thể, nên kết quả của mẫu sẽ khác với kết quả nếu khảo sát tổng thể. Sự khác biệt này được gọi là giới hạn sai số. Giới hạn sai số càng nhỏ, kết quả của mẫu càng gần với kết quả nếu khảo sát tổng thể.
Độ tin cậy (confidence level)	Mức độ tin cậy của bạn vào kết quả khảo sát. Ví dụ: mức độ tin cậy 95% có nghĩa là nếu bạn chạy cùng một cuộc khảo sát 100 lần, bạn sẽ nhận được kết quả tương tự 95 lần trên tổng số 100 lần. Độ tin cậy được nhắm tới trước khi bạn bắt

	đầu nghiên cứu vì nó sẽ ảnh hưởng đến mức độ sai số của bạn khi kết thúc nghiên cứu.
Khoảng tin cậy (confidence interval)	Phạm vi giá trị có thể có mà kết quả của tổng thể sẽ rơi vào với xác suất là độ tin cậy. Phạm vi này là kết quả mẫu +/- giới hạn sai số.
Ý nghĩa thống kê (Statistical significance)	Việc xác định kết quả của bạn có thể là do ngẫu nhiên hay không. Ý nghĩa thống kê càng lớn, kết luận do ngẫu nhiên càng ít mà thay vào đó là do một nguyên nhân cụ thể.

Ví dụ, khi thực hiện khảo sát: "bạn có thích du lịch ở Việt Nam hay không?". Ta có các ý sau:

- Tổng thể: toàn bộ dân số Việt Nam thực hiện khảo sát. Ước lượng khoảng 100 triệu người, điều này rõ ràng không khả thi để thực hiện khảo sát.
- Mẫu: thay vì phải khảo sát toàn bộ dân số Việt Nam, ta chỉ thực hiện khảo sát ngẫu nhiên một số người dân Việt Nam. Ví dụ: thực hiện khảo sát ngẫu nhiên 1000 người, với mọi lứa tuổi, thu nhập, điều kiện sống,... Điều này là hoàn toàn khả thi.
- Sau khi thực hiện lấy mẫu và thống kê, ta được các giá trị:

Thuật ngữ	**Giá trị**	**Giải thích ý nghĩa**
Kích thước tổng thể	100 000 000	Một trăm triệu dân số Việt Nam.
Kích thước mẫu	1068	Khảo sát ngẫu nhiên 1068 người.
Kết quả thống kê của mẫu	75%	Có 75% người được khảo sát trả lời thích du lịch ở Việt Nam.
Giới hạn sai số	3%	Kết quả của mẫu sẽ khác bao nhiêu phần trăm so với giá trị tổng thể.
Độ tin cậy	95%	Nếu thực hiện khảo sát nhiều lần, thì 95% giá trị của cuộc khảo sát sẽ rơi vào khoảng [72% - 78%]
Khoảng tin cậy	[72% - 78%]	

2. Lưu ý trong quá trình lấy mẫu

Khi quyết định kích thước mẫu phù hợp với mục đích phân tích, đây là những điều bạn cần lưu ý:

- Không sử dụng kích thước mẫu nhỏ hơn 30. Đã được chứng minh rằng 30 là kích thước mẫu tối thiểu mà một mẫu có thể đại diện cho kết quả của một tổng thể. (Nếu bạn đọc muốn tìm hiểu thêm về lý do tại sao kích thước mẫu tối thiểu là 30, có thể tham khảo thêm các tài liệu: Central Limit Theorem (CLT), Sample Size Formula)
- Mức độ tin cậy thường được sử dụng nhất là 95%, nhưng 90% có thể hoạt động tốt trong một số trường hợp.

Một số ghi chú khi tăng kích thước mẫu để đáp ứng nhu cầu cụ thể của dự án của bạn:
- Để có độ tin cậy cao hơn, hãy sử dụng kích thước mẫu lớn hơn.
- Để giảm giới hạn sai số, hãy sử dụng kích thước mẫu lớn hơn.
- Để có ý nghĩa thống kê lớn hơn, hãy sử dụng kích thước mẫu lớn hơn.

Kích thước mẫu lớn hơn sẽ tốn chi phí cao hơn.
- Bạn cũng phải cân nhắc chi phí tốn kém so với lợi ích nhận được khi kết quả chính xác hơn. Lấy ví dụ về khảo sát du lịch bên trên. Để có độ tin cậy là 95%, ta chỉ cần thực hiện khảo sát cho 1068 người. Tuy nhiên, để có độ tin cậy là 99%, ta cần thực hiện khảo sát của 1849 người.

Lưu ý: tính toán cụ thể kích thước mẫu sẽ được trình bày trong các phần sau của khóa học!

3. Ví dụ quá trình lấy mẫu

Kích thước mẫu sẽ khác nhau tùy thuộc vào loại vấn đề kinh doanh mà bạn đang cố gắng giải quyết.

Ví dụ: nếu bạn sống ở một thành phố với dân số 200.000 người và có 180.000 người trả lời khảo sát, thì đó là cỡ mẫu lớn. Nhưng nếu không làm vậy, kích thước mẫu nhỏ hơn có thể chấp nhận được sẽ như thế nào?

200 người có ổn không nếu những người được khảo sát là đại diện cho mọi quận trong thành phố?

Trả lời: Nó phụ thuộc vào việc ta đang khảo sát là gì?
- Kích thước mẫu 200 có thể đủ lớn nếu vấn đề phân tích của bạn là tìm hiểu xem cư dân cảm thấy như thế nào về thư viện mới.
- Kích thước mẫu 200 có thể không đủ lớn nếu vấn đề phân tích của bạn là xác định cách cư dân bỏ phiếu bầu cử.

Bạn có thể chấp nhận một sai số lớn hơn khi khảo sát cảm nhận của cư dân về thư viện mới so với khảo sát cư dân về cách họ sẽ bỏ phiếu bầu cử. Vì lý do đó, bạn rất có thể sẽ sử dụng cỡ mẫu lớn hơn cho cuộc khảo sát cử tri.

4. References

[1]https://www.coursera.org/learn/process-data/supplement/blyd3/calculating-sample-size

Bài đọc 5: Dữ liệu không sạch

Bài đọc sẽ củng cố về khái niệm của dữ liệu không sạch (dirty data). Đồng thời nêu lên một số loại dữ liệu không sạch thường gặp (dữ liệu trùng lặp, lỗi lời, không đầy đủ, không chính xác, không nhất quán), nguyên nhân và hậu quả.

1. Giới thiệu về dữ liệu không sạch

Dữ liệu không sạch (dirty data) là dữ liệu không đầy đủ, không chính xác hoặc không liên quan đến vấn đề bạn đang cố gắng giải quyết.
Dữ liệu không sạch có thể ảnh hưởng và thậm chí là làm sai lệch kết quả phân tích và dự đoán của nhà phân tích dữ liệu. Trong bài đọc này, ta sẽ lần lượt tìm hiểu các loại dữ liệu không sạch thường gặp, nguyên nhân và hậu quả.
Một số loại dữ liệu không sạch thường gặp:

Dữ liệu trùng lặp
(duplicate data)

Dữ liệu lỗi thời
(outdated data)

Dữ liệu không đầy đủ
(incomplete data)

Dữ liệu không chính xác
(incorrect/inaccurate data)

Dữ liệu không nhất quán
(inconsistent data)

2. Dữ liệu trùng lặp

Mô tả	Nguyên nhân có thể	Hậu quả
Bất kỳ bản ghi (dòng) dữ liệu nào hiển thị nhiều lần	Nhập dữ liệu thủ công, nhập dữ liệu hàng loạt hoặc di chuyển dữ liệu.	Các chỉ số hoặc phân tích sai lệch, số lượng hoặc dự đoán bị thổi phồng hoặc không chính xác hoặc nhầm lẫn trong quá trình truy xuất dữ liệu.

Ví dụ: một công ty có nhu cầu đếm xem có bao nhiêu khách hàng sử dụng dịch vụ của mình. Con số đếm được là 2000 khách hàng, nhưng thật ra chỉ có 1500 khách hàng, do có nhiều khách hàng được đếm hai lần (dữ liệu bị trùng lặp).

3. Dữ liệu lỗi thời

Mô tả	Nguyên nhân có thể	Hậu quả
Là dữ liệu đã không còn chính xác hoặc là không còn sử dụng nữa. Dữ liệu cũ này cần được thay thế bằng thông tin mới hơn và chính xác hơn	Nhân viên trong công ty thay đổi nhiệm vụ hoặc rời khỏi công ty. Khách hàng không sử dụng dịch vụ nữa. Đều làm cho phần mềm/hệ thống trở nên lỗi thời.	Thông tin chi tiết, ra quyết định và phân tích không chính xác.

Ví dụ: một công ty muốn phân tích tình hình du lịch Việt Nam hiện tại (năm 2022) để đưa ra quyết định kinh doanh. Nhưng nhà phân tích lại dùng dữ liệu năm 2015 để phân tích và đưa ra dự đoán. Điều đó làm cho kết quả dự đoán bị sai lệch và không đáng tin cậy, ảnh hưởng đến mục tiêu kinh doanh.

4. Dữ liệu không đầy đủ

Mô tả	Nguyên nhân có thể	Hậu quả
Bất kỳ dữ liệu nào bị thiếu các trường (cột) quan trọng	Thu thập dữ liệu không đúng cách hoặc nhập dữ liệu không chính xác	Năng suất giảm, thông tin chi tiết không chính xác hoặc

		không có khả năng hoàn thành nhu cầu phân tích.

Ví dụ: một nhà hàng khi phục vụ khách hàng không ghi nhận lại thời gian khách hàng dùng bữa. Sau này, khi nhà hàng cần phân tích khung giờ cao điểm trong ngày, để tăng số lượng nhân viên phục vụ. Lúc này, ta không có dữ liệu để phân tích.

5. Dữ liệu không chính xác

Mô tả	Nguyên nhân có thể	Hậu quả
Bất kỳ dữ liệu nào đầy đủ nhưng không chính xác	Lỗi do con người trong quá trình nhập dữ liệu, thông tin giả mạo hoặc dữ liệu giả	Thông tin chi tiết không chính xác hoặc ra quyết định dựa trên thông tin không đúng dẫn đến mất doanh thu.

Ví dụ, trong lĩnh vực y tế khi thông tin một bệnh nhân bị sai lệch, có thể dẫn đến những hậu quả vô cùng nghiêm trọng, thay vì "mắt khỏe" lại nhầm lẫn thành "cận thị"

6. Dữ liệu không nhất quán

Mô tả	Nguyên nhân có thể	Hậu quả
Bất kỳ dữ liệu nào sử dụng các định dạng khác nhau để đại diện cho cùng một thứ.	Dữ liệu được lưu trữ không chính xác hoặc lỗi xảy ra trong quá trình truyền/nhập dữ liệu.	Các điểm dữ liệu mâu thuẫn dẫn đến nhầm lẫn hoặc không thể phân loại hoặc phân khúc khách hàng

Ví dụ, định dạng ngày được dùng ở Việt Nam có dạng ngày/tháng/năm còn ở Mỹ dùng định dạng ngày tháng/ngày/năm. Điều này dẫn tới sự không nhất quán khi phân tích dữ liệu ở hai quốc gia.

Tác động của dữ liệu không sạch đến kinh doanh (đọc thêm):

- Ngân hàng: Chi phí không chính xác khiến các công ty mất từ 15% đến 25% doanh thu (nguồn).
- Thương mại điện tử: Có tới 25% địa chỉ liên hệ trong cơ sở dữ liệu B2B chứa thông tin không chính xác (nguồn).

- Tiếp thị và bán hàng: 8 trong số 10 công ty đã nói rằng dữ liệu không sạch cản trở các chiến dịch bán hàng (nguồn).
- Chăm sóc sức khỏe: Hồ sơ trùng lặp có thể là 10% và thậm chí lên đến 20% hồ sơ sức khỏe điện tử của bệnh viện (nguồn).

7. References

[1]https://www.coursera.org/learn/process-data/supplement/b90nY/what-is-dirty-data

Bài đọc 6: Một số lỗi phổ biến khi làm sạch dữ liệu

Bài đọc sẽ nêu lên tầm quan trọng của việc làm sạch dữ liệu và cách xác định các lỗi thường gặp trong quá trình làm sạch.

1. Một số lỗi phổ biến

Một số lỗi bạn có thể gặp phải khi làm sạch dữ liệu của mình có thể bao gồm:

- Không kiểm tra lỗi chính tả
- Không ghi nhận lại quá trình sửa lỗi
- Không kiểm tra giá trị bị điền sai
- Bỏ qua giá trị bị thiếu
- Chỉ xem xét một phần của dữ liệu
- Không quan tâm đến mục tiêu kinh doanh
- Không chỉnh sửa nguyên nhân gây ra lỗi
- Không phân tích hệ thống trước khi làm sạch dữ liệu
- Không sao lưu dữ liệu trước khi làm sạch
- Không quan tâm đến hạn chót của quá trình

Ta sẽ tìm hiểu chi tiết về các lỗi trên

Lỗi	Chi tiết
Không kiểm tra lỗi chính tả (Not checking for spelling errors)	- Lỗi chính tả có thể đơn giản như lỗi đánh máy. Hầu hết các lỗi chính tả hoặc lỗi ngữ pháp phổ biến đều có thể được phát hiện, nhưng điều này càng khó hơn với những thứ như tên hoặc địa chỉ. - Ví dụ: nếu bạn đang làm việc với một bảng tính dữ liệu khách hàng, bạn có thể bắt gặp một khách hàng tên là "John" có tên đã được nhập không chính xác thành "Jon" ở một số nơi. - Kiểm tra chính tả của bảng tính có thể sẽ không thông báo điều này, vì vậy nếu bạn không kiểm tra kỹ lỗi chính tả và nắm bắt lỗi này, kết quả phân tích của bạn sẽ có lỗi trong đó.
Không ghi nhận lại quá trình sửa lỗi.	- Ghi lại quá trình sửa lỗi có thể là một cách tiết kiệm thời gian, vì nó giúp bạn tránh những lỗi đó trong tương lai bằng cách cho bạn biết cách bạn đã giải quyết chúng.

(Forgetting to document errors)	- Ví dụ: bạn có thể tìm thấy lỗi trong một công thức trong bảng tính của mình. Bạn phát hiện ra rằng một vài ngày trong một trong các cột của bạn không được định dạng chính xác. Nếu bạn ghi chú về cách sửa lỗi này, bạn có thể tham khảo nó vào lần tiếp theo khi công thức của bạn bị hỏng và khắc phục sự cố. - Ghi lại các lỗi của bạn cũng giúp bạn theo dõi các thay đổi trong công việc của mình, do đó bạn có thể quay lại nếu bản sửa lỗi không hoạt động.
Không kiểm tra giá trị bị điền sai (Not checking for misfielded values)	- Giá trị bị điền sai thường xảy ra khi các giá trị được nhập vào sai cột. Những giá trị này có thể vẫn được định dạng chính xác, điều này khiến chúng khó nhận ra nếu không cẩn thận. - Ví dụ: bạn có thể có một tập dữ liệu với các cột thành phố và cột quốc gia. Đây là cùng một loại dữ liệu, vì vậy chúng rất dễ trộn lẫn. Ví dụ bạn đang tìm tất cả khách hàng có địa chỉ là TP Hồ Chí Minh, nhưng TP Hồ Chí Minh lại được nhập sai vào cột quốc gia. Điều này dẫn đến sai lệch khi phân tích. - Đảm bảo rằng dữ liệu của bạn đã được nhập chính xác là chìa khóa để phân tích chính xác và đầy đủ.
Bỏ qua giá trị bị thiếu (Overlooking missing values)	- Các giá trị bị thiếu trong tập dữ liệu của bạn có thể tạo ra lỗi và đưa ra kết luận không chính xác. - Ví dụ: nếu bạn đang cố gắng lấy tổng số lần bán hàng từ ba tháng trước, nhưng dữ liệu bị thiếu một tuần, thì tính toán của bạn sẽ không chính xác. - Cách tốt nhất là hãy cố gắng giữ cho dữ liệu của bạn sạch nhất có thể bằng cách duy trì tính hoàn chỉnh và nhất quán.
Chỉ xem xét một phần của dữ liệu (Only looking at a subset of the data)	- Điều quan trọng là phải suy nghĩ về tất cả các dữ liệu liên quan khi bạn đang làm sạch. Điều này giúp đảm bảo rằng bạn hiểu toàn bộ câu chuyện mà dữ liệu đang kể và bạn đang chú ý đến tất cả các lỗi có thể xảy ra. - Ví dụ: nếu bạn đang làm việc với dữ liệu về phân tích tình hình du lịch ở một địa phương nhất định. Bạn không thể chỉ tập trung vào

	cảnh vật của nơi đó, mà bạn còn phải xem xét các yếu tố như: cơ sở hạ tầng, ẩm thực, môi trường, ...
Không quan tâm đến mục tiêu kinh doanh (Losing track of business objectives)	- Khi đang làm sạch dữ liệu, bạn có thể có những khám phá mới và thú vị về tập dữ liệu của mình. Nhưng bạn không muốn những khám phá đó làm bạn xao nhãng khỏi nhiệm vụ hiện tại. - Ví dụ: nếu bạn đang làm việc với dữ liệu thời tiết để tìm số ngày mưa trung bình trong thành phố của mình, bạn cũng có thể nhận thấy một số mô hình thú vị về tuyết rơi. Điều đó thực sự thú vị, nhưng nó không liên quan đến câu hỏi bạn đang cố gắng trả lời ngay bây giờ. - Tò mò là điều tuyệt vời để sáng tạo và khám phá! Nhưng hãy cố gắng đừng để nó làm bạn phân tâm khỏi nhiệm vụ.
Không chỉnh sửa nguyên nhân gây ra lỗi (Not fixing the source of the error)	- Việc tự sửa lỗi là rất quan trọng. Nhưng nếu lỗi đó thực sự là một phần của một vấn đề lớn hơn, bạn cần phải tìm ra nguồn gốc của vấn đề. Nếu không, bạn sẽ phải tiếp tục sửa lỗi đó lặp đi lặp lại. - Ví dụ: hãy tưởng tượng bạn có một bảng tính nhóm theo dõi tiến trình của mọi người. Bảng tính liên tục bị lỗi vì những người khác nhau đang nhập các giá trị khác nhau. Thay vì khắc phục từng vấn đề này một cách đơn lẻ, bạn có thể thiết lập một định dạng chung để các thành viên trong nhóm nhập dữ liệu thống nhất và chính xác hơn. - Giải quyết nguồn gốc của các lỗi trong dữ liệu của bạn sẽ giúp bạn tiết kiệm rất nhiều thời gian về lâu dài.
Không phân tích hệ thống trước khi làm sạch dữ liệu (Not analyzing the system prior to data cleaning)	- Nếu bạn muốn làm sạch dữ liệu của mình và tránh các lỗi trong tương lai, bạn cần phải hiểu nguyên nhân gốc rễ của dữ liệu không sạch bạn. - Hãy tưởng tượng bạn là một thợ sửa ô tô. Bạn sẽ tìm ra nguyên nhân của vấn đề trước khi bắt đầu sửa xe, phải không? Đối với dữ liệu cũng vậy. Đầu tiên, bạn tìm ra lỗi đến từ đâu. Có thể do lỗi nhập dữ liệu, không thiết lập kiểm tra chính tả, thiếu định dạng hoặc do trùng lặp. Sau đó, khi bạn hiểu dữ liệu không sạch đến từ đâu, bạn có thể kiểm soát nó và giữ cho dữ liệu của mình sạch sẽ.

Không sao lưu dữ liệu trước khi làm sạch (Not backing up your data prior to data cleaning)	- Bạn nên chủ động và tạo bản sao lưu dữ liệu trước khi bắt đầu dọn dẹp dữ liệu. Nếu chương trình của bạn gặp sự cố hoặc nếu các thay đổi của bạn gây ra sự cố trong tập dữ liệu, bạn luôn có thể quay lại phiên bản đã lưu và khôi phục nó. Quy trình sao lưu dữ liệu đơn giản của bạn có thể giúp bạn tiết kiệm rất nhiều thời gian làm việc.
Không quan tâm đến hạn chót của quá trình (Not accounting for data cleaning in your deadlines/process)	- Tất cả công việc đều cần thời gian và điều đó bao gồm cả việc làm sạch dữ liệu. Điều quan trọng là phải ghi nhớ điều đó khi xem xét quá trình của bạn và xem xét hạn chót thời gian của bạn. - Khi bạn dự trữ thời gian cho việc dọn dẹp dữ liệu, điều này sẽ giúp bạn có được ước tính chính xác hơn về thời gian dự kiến cho các bên liên quan và có thể giúp bạn biết khi nào cần yêu cầu thời gian dự kiến cần được điều chỉnh.

Một số nguồn tham khảo (đọc thêm):

- Mười cách hàng đầu để làm sạch dữ liệu của bạn: Xem lại hướng dẫn để làm sạch dữ liệu trong Microsoft Excel. (nguồn)
- 10 mẹo làm sạch dữ liệu của Google Workspace: Tìm hiểu các phương pháp hay nhất để làm sạch dữ liệu trong Google Trang tính. (nguồn)

2. References

[1]https://www.coursera.org/learn/process-data/supplement/m3iWu/common-data-cleaning-pitfalls

Bài đọc 7: Giới thiệu SQL

Bài đọc với mục đích củng cố lại kiến thức về SQL, cách truy xuất dữ liệu cơ bản với SQL thông qua câu lệnh SELECT, FROM, WHERE. Đồng thời so sánh về ưu điểm của SQL so với bảng tính.

1. Giới thiệu SQL

Trong bài đọc này, bạn sẽ tìm hiểu thêm về cách quyết định khi nào sử dụng SQL (Structured Query Language – dịch: ngôn ngữ truy vấn có cấu trúc). Là một nhà phân tích dữ liệu, bạn sẽ được giao nhiệm vụ xử lý rất nhiều dữ liệu, và SQL là một trong những công cụ có thể giúp công việc của bạn trở nên dễ dàng hơn rất nhiều.

SQL là cách phổ biến để các nhà phân tích dữ liệu trích xuất dữ liệu từ cơ sở dữ liệu. Là một nhà phân tích dữ liệu, bạn sẽ làm việc với cơ sở dữ liệu rất nhiều, đó là lý do tại sao SQL là một kỹ năng quan trọng. Hãy cùng theo dõi khi một nhà phân tích dữ liệu sử dụng SQL để giải quyết một công việc kinh doanh.

SQL có thể:

- Tạo cơ sở dữ liệu mới
- Thực thi các truy vấn đối với cơ sở dữ liệu
- Tạo mới, chèn, cập nhật, xóa, bảng trong cơ sở dữ liệu

2. Ví dụ dùng SQL truy vấn dữ liệu

Trong phần này, ta sẽ xem xét bối cảnh kinh doanh cụ thể, xem xét các câu hỏi và đưa ra các câu truy vấn để trả lời câu hỏi.

Bối cảnh kinh doanh: “Nhà phân tích dữ liệu làm việc cho một công ty truyền thông. Một mô hình kinh doanh mới đã được triển khai vào ngày 15 tháng 2 năm 2020 và công ty muốn hiểu mức độ tăng trưởng người dùng của họ so với trước khi triển khai mô hình kinh doanh mới sẽ như thế nào.”

Với bảng khách hàng (customer) gồm các cột:

customer_id	customer_name	join_date

Ta sẽ tìm hiểu các câu hỏi và các câu truy vấn để trả lời câu hỏi

Câu hỏi	Câu truy vấn SQL
In ra danh sách toàn bộ khách hàng	SELECT * FROM customer;
Đếm số lượng toàn bộ khách hàng	SELECT COUNT(DISTINCT customer_id) FROM customer;
Danh sách người dùng đã tham gia kể từ ngày 15 tháng 2 năm 2020.	SELECT * FROM customer WHERE join_date >= '2020-02-15';
Đếm có bao nhiêu người dùng đã tham gia kể từ ngày 15 tháng 2 năm 2020.	SELECT COUNT(DISTINCT customer_id) FROM customer WHERE join_date >= '2020-02-15';

3. So sánh SQL và bảng tính

Trước khi thực hiện phân tích, nhà phân tích dữ liệu cần chọn công cụ để sử dụng (SQL, bảng tính, ... ?). Đầu tiên, ta phải xem xét dữ liệu được lưu trữ ở đâu. Nếu nó được lưu trữ trong cơ sở dữ liệu, thì SQL là công cụ tốt nhất cho công việc. Nhưng nếu nó được lưu trữ trong một bảng tính, thì ta sẽ phải thực hiện phân tích của mình trong bảng tính đó.

Ngoài ra, SQL và bảng tính đều có ưu điểm riêng, bao gồm:

Đặc trưng của bảng tính	Đặc trưng của SQL
Phù hợp với dữ liệu nhỏ.	Phù hợp với dữ liệu lớn.

Nhập dữ liệu theo cách thủ công. Cấu trúc linh hoạt hơn: bất kỳ ô nào cũng có thể thuộc bất kỳ kiểu dữ liệu nào, bất kể nó nằm trong cột nào.	Bảng SQL nghiêm ngặt hơn về các kiểu dữ liệu nhất quán và hạn chế người dùng nếu họ cố gắng nhập sai kiểu.
Tạo đồ thị và hình ảnh hóa trong cùng một chương trình	Việc tạo đồ thị trong SQL còn hạn chế. Tuy nhiên có thể tích hợp với các phần mềm/chương trình trực quan hóa khác.
Tích hợp kiểm tra chính tả và các chức năng hữu ích khác	Tốc độ thực thi các hàm có sẵn nhanh. Dễ dàng viết và lập trình ra các hàm.
Làm việc độc lập trong một dự án	Tuyệt vời cho công việc cộng tác và theo dõi các truy vấn do tất cả người dùng thực hiện. Sử dụng hết tiềm năng của cơ sở dữ liệu quan hệ.

4. Các biến thể của SQL

Mặc dù, SQL trở thành tiêu chuẩn của American National Standards Institute (dịch: Viện Tiêu chuẩn Quốc gia Hoa Kỳ, viết tắt: ANSI) vào năm 1986 và của Tổ chức Tiêu chuẩn hóa Quốc tế (ISO) vào năm 1987. SQL vẫn có các biến thể (còn gọi là phương ngữ - tiếng anh: dialects) của nó.

Một số công ty, phát triển cơ sở dữ liệu của riêng họ dựa trên SQL và các loại SQL khác nhau này là thứ giúp bạn giao tiếp với từng sản phẩm cơ sở dữ liệu khác nhau. Những biến thể này sẽ khác nhau giữa các công ty và có thể thay đổi theo thời gian nếu công ty chuyển sang hệ thống cơ sở dữ liệu khác. Vì vậy, rất nhiều nhà phân tích bắt đầu với SQL chuẩn và sau đó điều chỉnh dựa trên cơ sở dữ liệu mà họ đang làm việc. SQL tiêu chuẩn hoạt động với phần lớn cơ sở dữ liệu và yêu cầu một số thay đổi cú pháp nhỏ để thích ứng với các biến thể khác.

Là một nhà phân tích dữ liệu, điều quan trọng là phải biết rằng có sự khác biệt nhỏ giữa các biến thể. Nhưng bằng cách thành thạo SQL chuẩn, bạn sẽ sẵn sàng sử dụng SQL trong bất kỳ cơ sở dữ liệu nào.

Bạn có thể không cần biết mọi biến thể của SQL, nhưng sẽ rất hữu ích nếu biết rằng những biến thể khác nhau này tồn tại. Nếu bạn muốn tìm hiểu thêm về biến thể của SQL và khi nào chúng được sử dụng, bạn có thể xem các thông tin sau để biết thêm thông tin:

- Bài viết: What Is a SQL Dialect, and Which One Should You Learn?

- Bài viết Differences Between SQL Vs MySQL vs SQL Server
- Hướng dẫn What is SQL

5. References

[1]https://www.coursera.org/learn/process-data/supplement/GzgMb/using-sql-as-a-junior-data-analyst
[2]https://www.coursera.org/learn/process-data/supplement/gTnKd/sql-dialects-and-their-uses

Bài đọc 8: Giới thiệu về BigQuery

Bài đọc giới thiệu về BigQuery và ví dụ về việc tải dữ liệu với BigQuery.

1. Giới thiệu về BigQuery

Bạn phải có tài khoản BigQuery để có thể thao tác và sử dụng. Bạn có thể tham khảo lại cách sử dụng BigQuery trong khóa học Chuẩn bị dữ liệu để khám phá (Prepare Data for Exploration) sẽ bao gồm cách thiết lập tài khoản BigQuery.

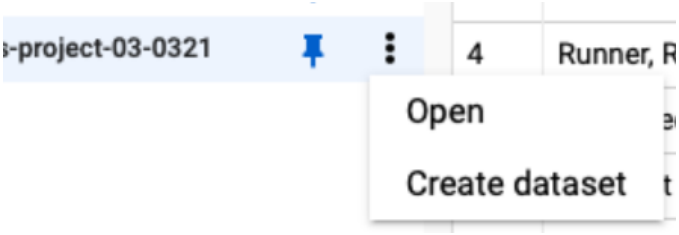
BigQuery có thể một kho dữ liệu (data warehouse) trên nền tảng điện toán đám mây của Google(Google Cloud) cho phép bạn phân tích và chạy các câu lệnh truy vấn nhanh đặc biệt là trên các tập dữ liệu lớn.

2. Hướng dẫn cơ bản tải dữ liệu với BigQuery

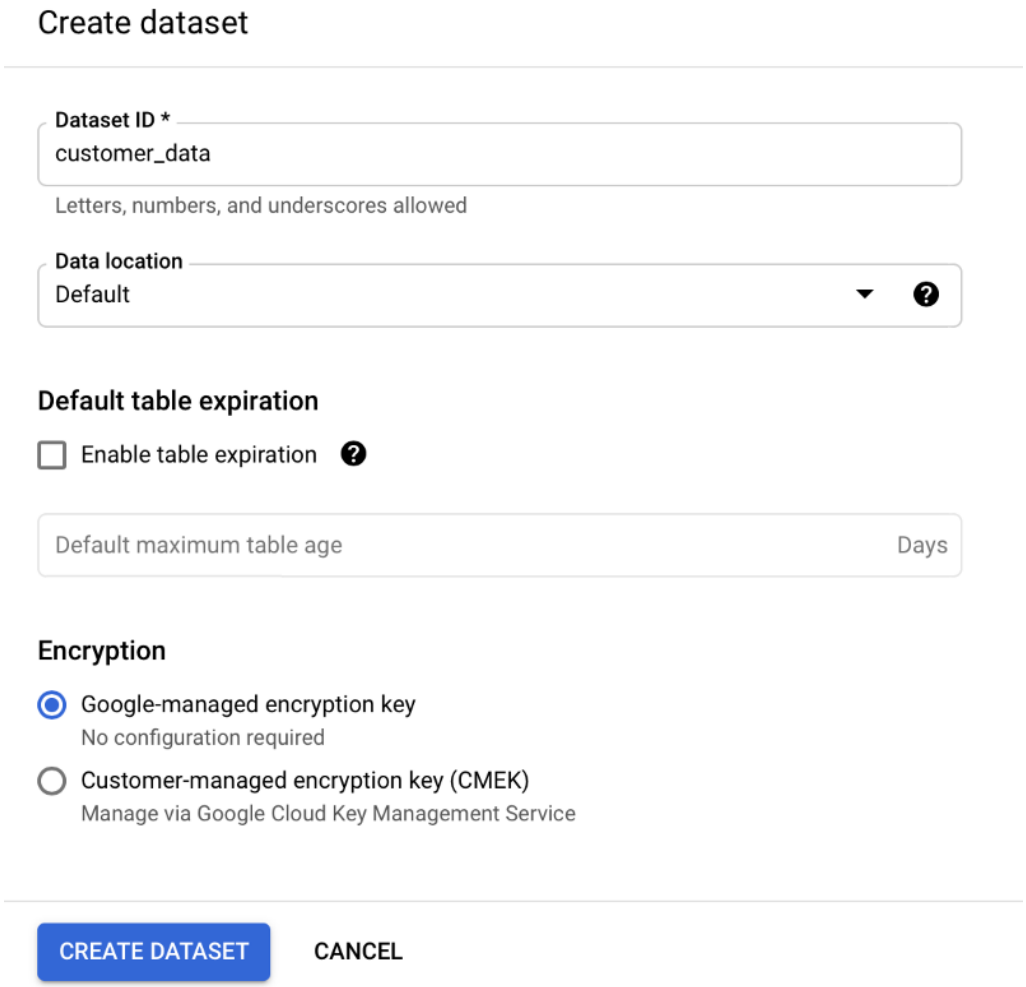
Để thao tác với dữ liệu, trước tiên ta cần có dữ liệu. Bạn có thể tải dữ liệu ở đây.

Các bước để thao tác với BigQuery gồm:

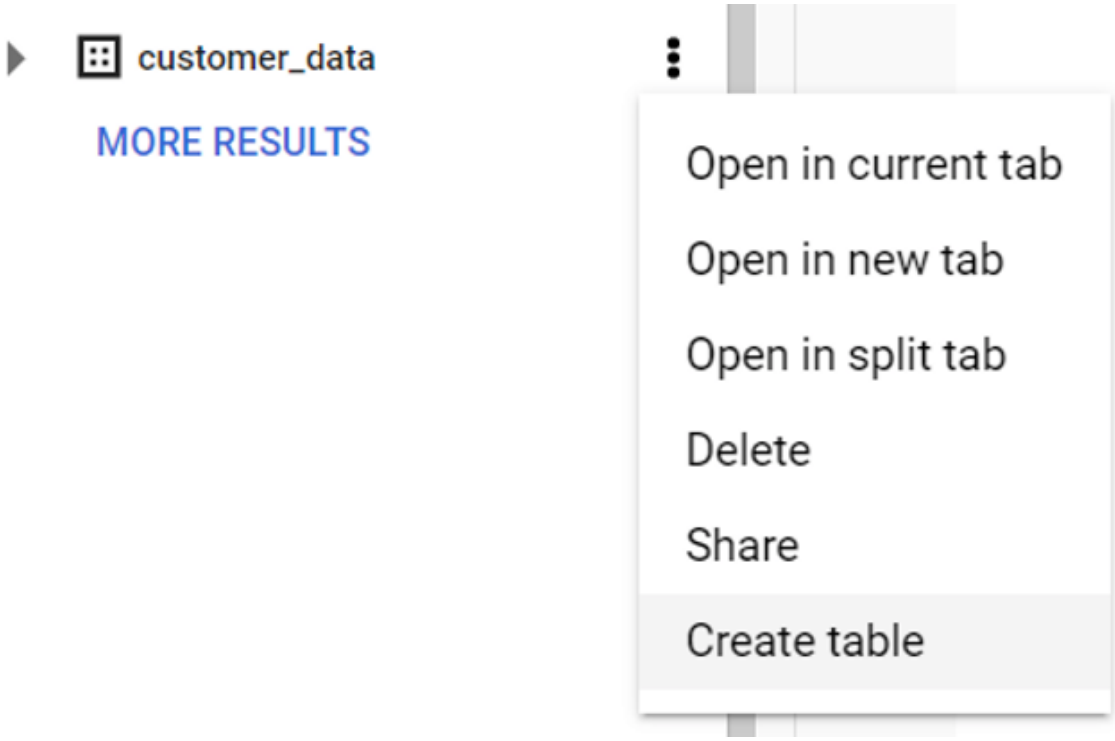
- Bước 1: Mở bảng console của BigQuery và nhấp vào dự án bạn muốn tải dữ liệu lên.
- Bước 2: Trong Explorer ở bên trái, nhấp vào biểu tượng hành động (ba dấu chấm dọc) bên cạnh tên dự án của bạn và chọn Create dataset.



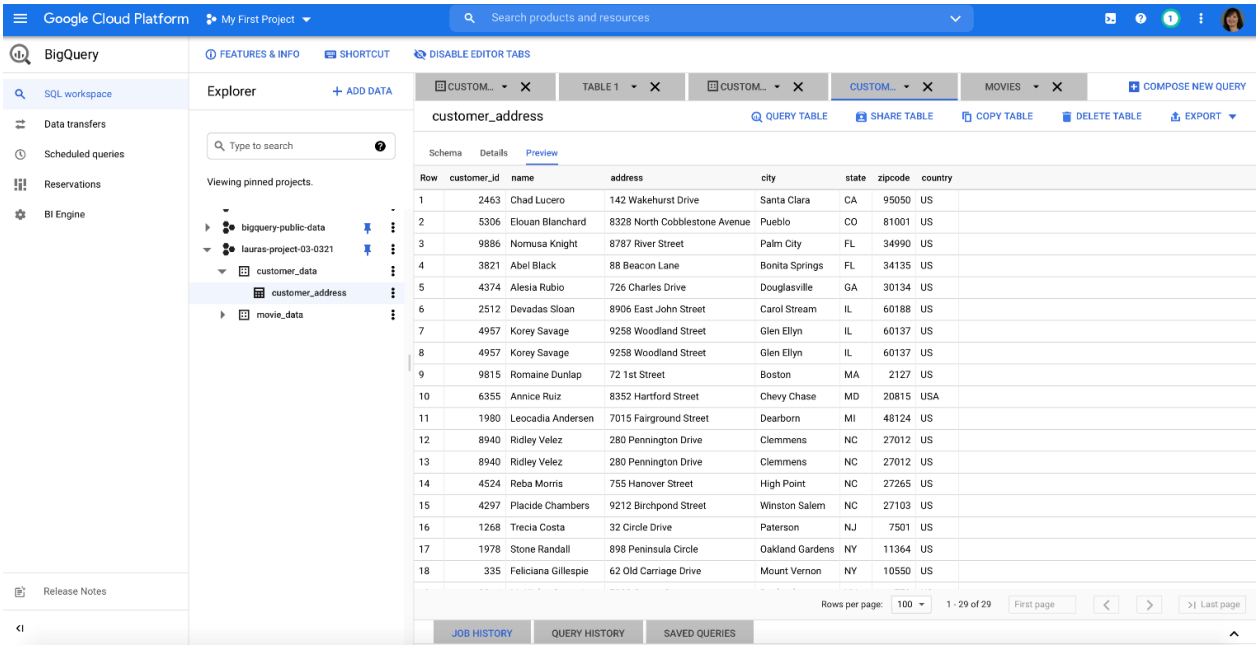
- Bước 3: bạn có thể đặt tên cho tập dữ liệu. Trong trường hợp này, ta sẽ đặt ‘customer_data’



- Bước 4: Nhấp vào CREATE DATASET (nút màu xanh) để thêm tập dữ liệu vào dự án của bạn.
- Bước 5: Trong Explorer ở bên trái, hãy nhấp để mở rộng dự án của bạn, sau đó nhấp vào tập dữ liệu customer_data mà bạn vừa tạo.
- Bước 6: Nhấp vào biểu tượng hành động (ba dấu chấm dọc) bên cạnh customer_data và chọn Create table.



- Bước 8: trong phần Source, trong phần Create table from:
  - Chọn Upload
  - Click chọn BROWSE để chọn tệp CSV mà bạn đã tải xuống khi nãy.
  - Trong phần File format, chọn định dạng CSV.
- Bước 9 (tùy chọn): bạn có thể đặt tên cho bảng bằng cách điền vào phần Table, ví dụ, ta điền vào **customer_address**
- Bước 10: trong phần Scheme, click chọn Auto detect.
- Bước 11: nhấp vào CREATE TABLE (nút màu xanh). Bây giờ bạn sẽ thấy bảng customer_address bên dưới tập dữ liệu customer_data trong dự án của bạn.
- Bước 12: nhấp vào customer_address và sau đó chọn thẻ Preview. Xác nhận rằng bạn thấy dữ liệu được hiển thị bên dưới.



Và bây giờ bạn đã biết cách tải dữ liệu lên BigQuery và có thể thực hành truy vấn dữ liệu của riêng bạn.

3. References

[1]https://www.coursera.org/learn/process-data/supplement/kt8qL/optional-upload-the-customer-dataset-to-bigquery

Bài đọc 9: Tiêu chí để xác minh dữ liệu sạch

Bài đọc này sẽ cung cấp danh sách các tiêu chí phổ biến mà bạn nên xem xét khi thực hiện xác minh quá trình làm sạch dữ liệu, bất kể bạn đang sử dụng công cụ làm sạch nào.

1. Một số lỗi cần được kiểm tra

Khi nói đến xác minh quá trình làm sạch dữ liệu, không có phương pháp tiếp cận chung cho tất cả các trường hợp hoặc một danh sách các tiêu chí có thể được áp dụng cho tất cả các dự án. Mỗi dự án có tổ chức và yêu cầu dữ liệu riêng biệt dẫn đến một danh sách những thứ cần chạy thử để xác minh.

Hãy nhớ rằng khi bạn nhận được nhiều dữ liệu hơn hoặc hiểu rõ hơn về các mục tiêu của dự án, bạn có thể muốn xem lại một số hoặc tất cả các bước này. Một số vấn đề phổ biến nhất bao gồm:

Vấn đề	Mô tả
Nguyên nhân gây ra lỗi (Sources of errors)	- Bạn đã sử dụng các công cụ và hàm phù hợp để tìm ra nguồn gốc gây lỗi trong tập dữ liệu của mình chưa? - Ví dụ: lỗi do công thức tính toán bị sai, thì ngoại trừ việc sửa các giá trị bị sai đó, ta cần làm đúng công thức đó để không gặp lại lỗi này trong tương lai.
Dữ liệu NULL (Null data)	- Bạn có tìm kiếm NULL bằng cách sử dụng định dạng có điều kiện (conditional formatting) hoặc bộ lọc (filter) không? - Ví dụ trong Google bảng tính, để tìm các ô không có giá trị (NULL), ta có thể làm các bước: o Bôi đen ô cần kiểm tra o Chọn “Định dạng”

	o Chọn "Định dạng có điều kiện" o Tại mục "Quy tắt định dạng" chọn "Trống" o Chọn "Đã xong" Google bảng tính sẽ làm nổi bật các ô NULL.
Dữ liệu nhập sai chính tả (Misspelled words)	- Bạn đã xác định được tất cả các lỗi chính tả chưa? - Ví dụ trong Google bảng tính, để kiểm tra lỗi chính tả ta có thể làm các bước: o Bôi đen ô cần kiểm tra o Chọn "Công cụ" o Chọn "Chính tả và ngữ pháp" o Chọn "Kiểm tra chính tả" Google bảng tính sẽ đưa ra các đề xuất chỉnh sửa nếu gặp phải lỗi chính tả.
Dữ liệu số nhập sai (Mistyped numbers)	- Bạn đã kiểm tra kỹ xem dữ liệu số của mình đã được nhập chính xác chưa? - Ví dụ: ta muốn nhập là 88%, nhưng trong quá trình tính toán và truyền tải, dữ liệu đã biến thành 0.88 - Trong trường hợp như trên, ta có thể xử lý bằng cách: o Bôi đen ô cần kiểm tra o Chọn "Định dạng" o Chọn "Số" o Chọn "Phần trăm" Google trang tính sẽ chuyển từ 0.88 về 88%
Dư thừa các ký tự hoặc khoảng trắng (Extra spaces and characters)	- Bạn đã xóa bất kỳ khoảng trắng hoặc ký tự thừa nào bằng cách sử dụng hàm TRIM không? - Nếu bị dư thừa khoảng trắng, sẽ dẫn đến sai sót trong quá trình phân tích dữ liệu. Ví dụ, hai chuỗi "khoa học" và " khoa học " là khác nhau.
Dữ liệu bị trùng lặp (Duplicates)	- Bạn đã xóa các bản sao trong bảng tính bằng cách sử dụng chức năng "xóa bản trùng lặp" hoặc từ khóa DISTINCT trong SQL?

Kiểu dữ liệu không chính xác (Mismatched data types)	- Bạn đã kiểm tra xem dữ liệu số, ngày tháng và chuỗi có chính xác hay không? - Ví dụ, giá trị cột yêu cầu là ghi vào ngày tháng năm sinh mà người dùng lại điền vào số tuổi, điều đó là chưa chính xác.
Chuỗi nhập không nhất quán (Messy/inconsistent strings)	- Bạn có đảm bảo rằng tất cả các chuỗi của mình đều nhất quán và có ý nghĩa không? - Ví dụ, khi nhập tên quốc gia thì nhập là VN hay Vina hay Vietnamese hay Việt Nam?
Định dạng ngày tháng không nhất quán (Messy/inconsistent date formats)	- Bạn có định dạng ngày tháng một cách nhất quán trong tập dữ liệu của mình không? - Ví dụ, thời gian nhập có định dạng như thế nào ngày/tháng/năm hay là tháng/ngày/năm
Nhãn của dữ liệu gây hiểu nhầm	- Bạn đã đặt tên cho các cột của mình một cách có ý nghĩa, đầy đủ và chính xác chưa? - Ví dụ, trong điều tra nhân khẩu, người dùng được yêu cầu nhập vào cột có tên là “phái”, vậy “phái” ở đây có nghĩa là gì, là “giới tính” hay “đảng phái”?
Dữ liệu bị cắt bớt (Truncated data)	- Bạn đã kiểm tra dữ liệu bị cắt ngắn hoặc bị thiếu cần sửa chữa? - Ví dụ, họ và tên của người dùng tiếng Việt có bị mất dấu (Nguyễn Văn A trở thành Nguyen Van A) trong quá trình xử lý dữ liệu không?
Logic kinh doanh	- Bạn đã kiểm tra xem dữ liệu có hợp lý với kiến thức của bạn về doanh nghiệp không? - Ví dụ: dịp lễ tết nguyên đán và số lượng người mua bánh trung thu tăng đột biến. Sự kiện trên có vẻ chưa hợp lý lắm, có thể là quá trình lưu trữ hoặc trích xuất dữ liệu có vấn đề.

2. Đánh giá lại mục tiêu kinh doanh và dữ liệu

Khi bạn đã hoàn thành các nhiệm vụ làm sạch dữ liệu, bạn nên xem lại mục tiêu của dự án và xác nhận rằng dữ liệu của bạn vẫn phù hợp với mục tiêu đó.
Đây là một quá trình liên tục mà bạn sẽ thực hiện trong suốt dự án của mình - nhưng đây là ba bước bạn có thể ghi nhớ khi nghĩ về điều này:

- Xác nhận vấn đề kinh doanh
- Xác nhận mục tiêu của dự án
- Xác minh rằng dữ liệu có thể giải quyết vấn đề và phù hợp với mục tiêu

3. References

[1]https://www.coursera.org/learn/process-data/supplement/c9vis/data-cleaning-verification-a-checklist

Bài đọc 10: Ghi lại thay đổi trong quá trình làm sạch

Bài đọc trình bày về bảng ghi thay đổi (changelog), một công cụ cho phép nhà phân tích dữ liệu ghi nhận lại những thay đổi xảy ra trong quá trình làm sạch và chuyển đổi dữ liệu. Bài đọc sẽ giới thiệu một số tiêu chí cần có cho một bảng ghi thay đổi.
Bên cạnh đó, bài đọc còn nêu lên cách ghi nhận những thay đổi đối với một hệ thống đã được triển khai ngoài thực tế, bằng cách áp dụng hệ thống quản lý phiên bản (version control system).

1. Giới thiệu một số công cụ hỗ trợ việc ghi lại lịch sử thay đổi

Các nhà phân tích dữ liệu sử dụng các bảng thay đổi để theo dõi quá trình chuyển đổi và làm sạch dữ liệu.
Hầu hết các phần mềm/chương trình đều tích hợp sẵn các cách để theo dõi lịch sử. Ví dụ: trong Google trang tính, bạn có thể kiểm tra lịch sử phiên bản của toàn bộ trang tính hoặc một ô riêng lẻ và quay lại phiên bản cũ hơn. Trong Microsoft Excel, bạn có thể sử dụng một tính năng được gọi là Track Changes. Và trong BigQuery, bạn có thể xem lịch sử để kiểm tra những gì đã thay đổi.
Một số ví dụ:

Google trang tính	1. Nhấp chuột phải vào ô và chọn Hiển thị lịch sử chỉnh sửa (Show edit history). 2. Nhấp vào mũi tên trái hoặc mũi tên phải để tiến và lùi trong lịch sử nếu cần.
Microsoft Excel	1. Nếu Track Changes đã được bật cho bảng tính: hãy nhấp vào Xem lại. 2. Trong Track Changes, hãy nhấp vào Accept/Reject Changes để chấp nhận hoặc từ chối bất kỳ thay đổi nào được thực hiện.

Bigquery	Đưa ra một phiên bản trước đó (mà không hoàn nguyên về nó) và tìm ra những gì đã thay đổi bằng cách so sánh nó với phiên bản hiện tại.

2. Bảng ghi thay đổi (changelog)

Bảng ghi thay đổi (changelog) được xây dựng dựa trên lịch sử phiên bản tự động bằng cách cung cấp cho bạn bảng ghi chi tiết về công việc của bạn. Đây là nơi các nhà phân tích dữ liệu ghi lại tất cả những thay đổi mà họ thực hiện đối với dữ liệu.

Ngoài việc ghi nhận lại thay đổi, bảng ghi thay đổi còn ghi nhận lại lý do xảy ra hay đổi. Bảng ghi thay đổi hữu ích để giúp ta hiểu lý do thay đổi được thực hiện. Bảng ghi thay đổi không có định dạng cố định và bạn có thể tạo ra cách ghi phù hợp cho cá nhân mình. Tuy nhiên nếu bạn đang sử dụng bảng thay đổi được chia sẻ với nhiều đối tác khác, tốt nhất là bạn nên thống nhất với các nhà phân tích dữ liệu khác về định dạng của bảng ghi thay đổi của bạn.

Thông thường, một bảng ghi thay đổi ghi lại loại thông tin này:

- Dữ liệu, tệp, công thức, truy vấn hoặc bất kỳ thành phần nào khác đã thay đổi.
- Mô tả về những gì đã thay đổi.
- Ngày thay đổi.
- Người thực hiện thay đổi.
- Người đã xác nhận sự thay đổi.
- Phiên bản của dữ liệu.
- Lý do thay đổi.
-

Giả sử bạn đã thực hiện thay đổi đối với một công thức trong bảng tính vì bạn đã quan sát công thức đó trong một báo cáo khác và bạn muốn dữ liệu của mình trùng khớp và nhất quán. Nếu sau đó bạn phát hiện ra rằng công thức trong báo cáo đó là sai, thì lịch sử phiên bản tự động của bảng tính sẽ giúp bạn hoàn tác thay đổi. Nhưng nếu bạn cũng ghi lại lý do thay đổi trong bảng ghi

thay đổi, bạn có thể quay lại với người tạo báo cáo và cho họ biết về công thức không chính xác đó. Bằng cách theo dõi, bạn sẽ đảm bảo tính toàn vẹn của dữ liệu của dự án. Bạn cũng sẽ thể hiện sự chính trực cá nhân như một người có thể được tin cậy. Đó là sức mạnh của một bảng ghi thay đổi (changelog)!

Cuối cùng, bảng ghi thay đổi rất quan trọng khi có nhiều thay đổi đối với bảng tính hoặc truy vấn. Hãy tưởng tượng một nhà phân tích đã thực hiện bốn thay đổi và ta muốn xóa đi kết quả của lần thay đổi #2. Thay vì nhấp vào tính năng hoàn tác ba lần để hoàn tác thay đổi #2 (và mất các thay đổi #3 và #4), nhà phân tích có thể hoàn tác thay đổi #2 và giữ nguyên tất cả các thay đổi khác. Bây giờ, ví dụ của chúng tôi chỉ là 4 thay đổi, nhưng hãy thử nghĩ xem bảng thay đổi đó sẽ quan trọng như thế nào nếu có hàng trăm thay đổi cần theo dõi.

3. Hệ thống quản lý phiên bản (version control system) trong thực tế

Nếu một nhà phân tích đang thực hiện các thay đổi đối với truy vấn SQL đang được chia sẻ trong toàn công ty, thì công ty rất có thể sử dụng cái được gọi là hệ thống quản lý phiên bản (version control system). Một ví dụ có thể là câu truy vấn dùng để tổng kết doanh thu hàng ngày để xây dựng báo cáo cho quản lý cấp cao.

Dưới đây là cách hệ thống quản lý phiên bản ảnh hưởng đến sự thay đổi đối với một truy vấn:

- Một công ty có phiên bản chính thức của các truy vấn quan trọng trong hệ thống quản lý phiên bản của họ.
- Một nhà phân tích đảm bảo rằng phiên bản họ đang nghiên cứu và chỉnh sửa là phiên bản mới nhất. Đây được gọi là đồng bộ hóa (syncing)
- Nhà phân tích thực hiện một thay đổi đối với truy vấn.
- Nhà phân tích có thể yêu cầu ai đó xem xét sự thay đổi này. Đây được gọi là code review và có thể được thực hiện không chính thức hoặc chính thức. Một đánh giá không chính thức có thể đơn giản như: yêu cầu một nhà phân tích cấp cao xem xét sự thay đổi.
- Sau khi người đánh giá phê duyệt thay đổi, nhà phân tích gửi phiên bản cập nhật của truy vấn tới kho lưu trữ trong hệ thống quản lý phiên bản của

công ty. Đây được gọi là code commit. Một phương pháp hay nhất là ghi lại chính xác những gì thay đổi là gì và tại sao nó được thực hiện.

- Sau khi thay đổi được gửi (submitted), mọi người khác trong công ty sẽ có thể truy cập và sử dụng truy vấn mới này khi họ đồng bộ hóa (sync) với các truy vấn cập nhật nhất được lưu trữ trong hệ thống quản lý phiên bản.
- Nếu truy vấn có vấn đề hoặc nhu cầu kinh doanh thay đổi, nhà phân tích có thể hoàn tác (undo) thay đổi đối với truy vấn bằng cách sử dụng hệ thống quản lý phiên bản. Nhà phân tích có thể xem danh sách theo thứ tự thời gian của tất cả các thay đổi được thực hiện đối với truy vấn và ai đã thực hiện từng thay đổi. Sau đó, sau khi tìm thấy câu truy vấn thay đổi, nhà phân tích có thể hoàn nguyên (revert) về phiên bản trước đó.
- Truy vấn trở lại như trước khi nhà phân tích thực hiện thay đổi. Và tất cả mọi người trong công ty cũng thấy truy vấn ban đầu.

4. References

[1]https://www.coursera.org/learn/process-data/supplement/FvuSF/embrace-changelogs

Bài đọc 11: Các hàm nâng cao để tăng tốc quá trình làm sạch dữ liệu

Bài đọc sẽ trình bày về một số chức năng nâng cao có thể giúp tăng tốc quá trình làm sạch dữ liệu trong bảng tính, bao gồm: QUERY và FILTER

1. Truy vấn dữ liệu từ nhiều nguồn, dùng câu lệnh QUERY

Dưới đây là bảng tóm tắt chức năng:

Chức năng	Cú pháp Google trang tính	Các bước Microsoft Excel	Cách dùng
QUERY	QUERY(dữ liệu, truy vấn, [tiêu đề])	Data → From Other Sources → From Microsoft Query	Bắt chước câu lệnh SQL để nhập dữ liệu

Hàm QUERY hữu ích khi bạn muốn lấy dữ liệu từ một bảng tính khác. Đối với một lượng lớn dữ liệu, việc sử dụng chức năng QUERY sẽ nhanh hơn so với việc lọc dữ liệu theo cách thủ công. Ví dụ: bạn có thể tạo danh sách tất cả khách hàng đã mua sản phẩm của công ty bạn trong một tháng cụ thể bằng cách sử dụng tính năng lọc thủ công. Nhưng nếu bạn cũng muốn tính toán mức tăng trưởng khách hàng theo tháng, bạn phải sao chép dữ liệu đã lọc sang một bảng tính mới, lọc dữ liệu cho doanh số bán hàng trong tháng tiếp theo, sau đó sao chép các kết quả đó để phân tích. Với chức năng QUERY, bạn có thể lấy tất cả dữ liệu cho cả hai tháng mà không cần thay đổi tập dữ liệu gốc hoặc sao chép kết quả.

Bạn nhập trang tính theo tên và phạm vi dữ liệu mà bạn muốn truy vấn, sau đó sử dụng lệnh SELECT để chọn các cột cụ thể. Bạn cũng có thể thêm các tiêu chí cụ thể sau câu lệnh SELECT bằng cách thêm câu lệnh WHERE. Nhưng hãy nhớ rằng, tất cả mã SQL bạn sử dụng phải được đặt giữa các dấu ngoặc kép.

Ví dụ về việc sử dụng QUERY.

- Ví dụ, ta có bảng khách hàng gồm 3 thuộc tính: mã khách hàng (ma_khach_hang), tên khách hàng (ten_khach_hang) và thu nhập (thu_nhap).

	A	B	C
1	ma_khach_hang	ten_khach_hang	thu_nhap
2	1	Nguyen Van A	1000
3	2	Tran Thi B	800
4	3	Vo Van C	600

- Nhiệm vụ của nhà phân tích là truy vấn ra những khách hàng có thu nhập lớn hơn 700. Câu truy vấn sẽ là

```
=QUERY(A2:C4, "select B where C > 700")
```

- Và kết quả sẽ truy xuất ra những khách hàng đó:

1	Nguyen Van A	1000
2	Tran Thi B	800

Tài liệu đọc thêm

- Hãy xem trang hỗ trợ của Google để biết hàm QUERY với cách sử dụng, cú pháp và ví dụ mẫu mà bạn có thể tải xuống trong trang tính của Google.
- Liên kết để tạo bản sao của trang tính ví dụ: ví dụ về QUERY

Giải pháp thực tế

Các nhà phân tích có thể sử dụng SQL để kéo một tập dữ liệu cụ thể vào một bảng tính. Sau đó, họ có thể sử dụng hàm QUERY để tạo nhiều tab của tập dữ liệu đó. Ví dụ: một tab có thể chứa tất cả dữ liệu bán hàng cho một tháng cụ thể và một tab khác có thể chứa tất cả dữ liệu bán hàng từ một khu vực cụ thể. Giải pháp này minh họa cách sử dụng tốt SQL và bảng tính cùng nhau.

2. Lọc dữ liệu, dùng câu lệnh FILTER

Dưới đây là bảng tóm tắt chức năng:

Chức năng	Cú pháp Google trang tính	Các bước Microsoft Excel	Cách dùng
FILTER	FILTER(dải ô, điều kiện 1, [điều kiện 2, ...])	Filter (điều kiện mỗi cột)	Chỉ hiển thị dữ liệu đáp ứng các điều kiện được chỉ định.

Hàm FILTER toàn nằm trong bảng tính và không yêu cầu sử dụng ngôn ngữ truy vấn. Hàm FILTER cho phép bạn chỉ xem các hàng (hoặc cột) trong dữ liệu nguồn đáp ứng các điều kiện đã chỉ định của bạn. Nó giúp bạn có thể lọc dữ liệu trước khi phân tích.

Hàm FILTER có thể chạy nhanh hơn hàm QUERY. Nhưng hãy nhớ rằng, hàm QUERY có thể được kết hợp với các hàm khác để thực hiện các phép tính phức tạp hơn. Ví dụ: hàm QUERY có thể được sử dụng với các hàm khác như SUM và COUNT để tóm tắt dữ liệu, nhưng hàm FILTER thì không.

Ví dụ về sử dụng FILTER

- Ta sẽ lấy ví dụ, tương tự như phần QUERY bên trên: lấy ra thông tin của những khách hàng có thu nhập lớn hơn 700
- Ta sẽ dùng hàm FILTER với cú pháp:

```
=FILTER(A2:C4,C2:C4>700)
```

- Và kết quả sẽ truy xuất ra những khách hàng đó:

1	Nguyen Van A	1000
2	Tran Thi B	800

Tài liệu đọc thêm

- Kiểm tra trang hỗ trợ của Google để biết chức năng FILTER với cách sử dụng, cú pháp và ví dụ mẫu mà bạn có thể tải xuống trong trang tính của Google.

- Liên kết để tạo bản sao của trang tính ví dụ: ví dụ về FILTER

3. References

[1]https://www.coursera.org/learn/process-data/supplement/PLnRS/advanced-functions-for-speedy-data-cleaning

Phần 2
HƯỚNG DẪN TRẢ LỜI CÂU HỎI

Chương 1

Giới thiệu về phân tích dữ liệu

Bài tập

1. Giả sử bạn là nhà phân tích dữ liệu. Cho dữ liệu bên dưới. Hãy tìm dữ liệu trùng lặp. Link dữ liệu: June 2014 Invoices

A. Valando ngày 1/1/2014

B. Valando ngày 2/18/2014

C. Symteco ngày 21/02/2014

D. Symteco ngày 20/5/2014

Đáp án: B

2. Sử dụng dữ liệu June 2014 Invoices (được cho ở câu 10). Dòng 10 và dòng 11 có dữ liệu trùng? Đúng hay sai?

A. Đúng

B. Sai

Đáp án: B

3.Sử dụng dữ liệu June 2014 Invoices (được cho ở câu 10). Dòng 8 và 9 có gì bất thường?

A. Dòng 9 là bản sao của dòng 8.

B. Dòng 8 không đúng định dạng.

C. Dòng 9 cần thêm dữ liệu.

D. Dòng 8 và dòng 9 hiển thị sai đơn vị tiền tệ.

Đáp án: A

Toàn vẹn dữ liệu

Bài tập

1.Điền vào chỗ trống: Nếu một nhà phân tích dữ liệu đang sử dụng dữ liệu bị _____, dữ liệu sẽ thiếu tính toàn vẹn và quá trình phân tích sẽ bị lỗi.

A. Làm sạch (clean)

B. Bị tổn hại (compromised)

C. Mở rộng (wide)

D. Công khai (public)

Đáp án: B

2. Điều kiện nào sau đây là cần thiết để đảm bảo tính toàn vẹn của dữ liệu? (Có thể chọn nhiều đáp án)

A. Tính chính xác (accuracy)

B. Tính hoàn chỉnh (completeness)

C. Sức mạnh thống kê (statistical power)

D. Riêng tư (privacy)

Đáp án: A, B

3. Điền vào chỗ trống: Dữ liệu _____ đề cập đến tính chính xác, đầy đủ, nhất quán và đáng tin cậy của dữ liệu trong suốt vòng đời của dữ liệu.

A. Lấy mẫu (sampling)

B. Nhân bản (replication)

C. Phân tích (analysis)

D. Toàn vẹn (integrity)

Đáp án: D

4. Một nhà phân tích tài chính nhập một tập dữ liệu vào máy tính của họ từ một

thiết bị lưu trữ. Khi nó đang được nhập, kết nối bị gián đoạn, điều này làm ảnh hưởng đến dữ liệu. Quá trình gây ra tổn hại đó tên là gì?

A. Thao tác dữ liệu (data manipulation)

B. Phân tích dữ liệu (data analysis)

C. Chuyển đổi dữ liệu (data transfer)

D. Thu thập dữ liệu (data gathering)

Đáp án: C

5. Một bệnh viện lưu các bản sao dữ liệu của họ tại một số địa điểm trên toàn quốc. Dữ liệu có vấn đề do mỗi vị trí tạo một bản sao của bản gốc vào các thời điểm khác nhau trong ngày. Quá trình gây ra tổn hại đó tên là gì?

A. Thao tác dữ liệu (data manipulation)

B. Nhân bản dữ liệu (data replication)

C. Chuyển đổi dữ liệu (data transfer)

D. Thu thập dữ liệu (data gathering)

Đáp án: B

6. Một vấn đề tiềm ẩn liên quan đến thao tác dữ liệu mà các nhà phân tích phải lưu ý là gì?

A. Thao tác dữ liệu có thể gây ra lỗi.

B. Thao tác dữ liệu có thể làm cho tập dữ liệu dễ đọc hơn.

C. Thao tác dữ liệu có thể giúp tổ chức tập dữ liệu.

D. Thao tác dữ liệu có thể tách tập dữ liệu giữa các vị trí khác nhau.

Đáp án: A

Mối quan hệ giữa dữ liệu và mục tiêu kinh doanh

1. Nguyên tắc nào sau đây là yếu tố chính của tính toàn vẹn dữ liệu? (Có thể chọn nhiều câu trả lời)

A. Tính chính xác (Accuracy)

B. Tính chọn lọc (Selectivity)

C. Tính nhất quán (Consistency)

D. Đáng tin cậy (Trustworthiness)

Đáp án: A,C,D

2. Các nhà phân tích dữ liệu sử dụng quy trình nào để làm cho dữ liệu có tổ chức hơn và dễ đọc hơn?

A. Nhân bản dữ liệu (Data replication)

B. Chuyển đổi dữ liệu (Data transfer)

C. Thao tác dữ liệu (Data manipulation)

D. Đồng nhất dữ liệu (Data uniformity)

Đáp án: C

3. Trước khi phân tích, một công ty thu thập dữ liệu từ các quốc gia sử dụng các định dạng ngày tháng khác nhau. Bản cập nhật nào sau đây sẽ cải thiện tính toàn vẹn của dữ liệu?

A. Sắp xếp dữ liệu theo quốc gia

B. Thay đổi tất cả các ngày thành cùng một định dạng

C. Xóa dữ liệu ở định dạng ngày không quen thuộc

D. Để các ngày ở định dạng hiện tại của chúng

Đáp án: B

Bài tập

1. Một nhà phân tích dữ liệu được cung cấp một tập dữ liệu để phân tích. Nó chỉ gồm dữ liệu về số lượng dân số của mỗi quốc gia trong 20 năm trước đó. Dựa trên dữ liệu có sẵn, một nhà phân tích sẽ có thể xác định lý do đằng sau sự gia tăng dân số của một quốc gia nhất định từ năm 2016 đến năm 2017.

Điều này, có hợp lý hay không?

A. Đúng

B. Sai

Đáp án: B
(Do dữ liệu chỉ bao gồm số lượng, thì không thể dùng để phân tích lý do đằng sau sự gia tăng dân số)

2. Một nhà phân tích dữ liệu được cung cấp một tập dữ liệu để phân tích. Nó bao gồm dữ liệu về số lượng dân số của mỗi quốc gia trong 20 năm (2001-2021). Nhà phân tích có thể sử dụng tập dữ liệu này để giải quyết những câu hỏi nào sau đây? (có thể chọn nhiều đáp án)

A. Dân số trung bình của một quốc gia từ năm 2015 đến năm 2020

B. Sự khác biệt về dân số giữa hai quốc gia cụ thể trong năm 2018

C. Tác động của di cư đối với dân số của một quốc gia nhất định

D. Lý do cho sự gia tăng dân số ở một quốc gia nhất định là gì

Đáp án: A, B
(Do đáp án C và D cần thêm dữ liệu để phân tích)

3.Một nhà phân tích dữ liệu được cung cấp một tập dữ liệu để phân tích. Nó bao gồm dữ liệu về số lượng dân số của mỗi quốc gia trong 20 năm (2001-2021). Câu hỏi nào sau đây mà nhà phân tích cần thêm dữ liệu để giải quyết?

A. Dân số của một quốc gia nhất định vào năm 2020 là bao nhiêu

B. Quốc gia nào có dân số đông nhất vào năm 2015

C. Quốc gia nào có dân số nhỏ nhất vào năm 2017?

D. Lý do cho sự gia tăng dân số ở một quốc gia nhất định là gì?

Đáp án: D.
(Ta cần thêm dữ liệu để biết được lý do đằng sau).

4.Quy trình nào sau đây giúp đảm bảo sự liên kết chặt chẽ giữa dữ liệu và mục tiêu kinh doanh?

A. Duy trì tính toàn vẹn của dữ liệu

B. Hoàn thành sao chép dữ liệu

C. Tự động cập nhật dữ liệu trong quá trình phân tích

D. Truyền dữ liệu nhiều lần

Đáp án: A

5.Đôi khi trong quá trình phân tích, nhà phân tích phát hiện ra rằng cần phải điều chỉnh mục tiêu kinh doanh. Khi điều này xảy ra, nhà phân tích nên chủ động thực hiện mà không liên quan đến người khác.

A. Đúng

B. Sai

Đáp án: B

(Nếu một nhà phân tích dữ liệu tin rằng mục tiêu kinh doanh cần được điều chỉnh, điều quan trọng trước tiên là phải thảo luận với các bên liên quan.)

Vấn đề về dữ liệu không đầy đủ

1. Nhà phân tích nên làm gì nếu họ không có dữ liệu cần thiết để đáp ứng mục tiêu kinh doanh?(Có thể chọn nhiều câu trả lời)

A. Tiếp tục phân tích, sử dụng dữ liệu từ các nguồn kém tin cậy.

B. Tạo và sử dụng dữ liệu giả định phù hợp với các phân tích.

C. Thu thập dữ liệu liên quan ở quy mô nhỏ và yêu cầu thêm thời gian để tìm dữ liệu đầy đủ hơn.

D. Thực hiện phân tích bằng cách tìm và sử dụng dữ liệu thay từ các nguồn dữ liệu khác.

Đáp án: C, D

2. Điều nào sau đây là những hạn chế có thể dẫn đến không đủ dữ liệu? (Có thể chọn nhiều câu trả lời)

A. Dữ liệu trùng lặp

B. Dữ liệu vẫn được cập nhật.

C. Dữ liệu bị lỗi thời.

D. Dữ liệu bị giới hạn từ một nguồn.

Đáp án: B,C,D

3. Một nhà phân tích dữ liệu muốn tìm hiểu xem có bao nhiêu gia đình ở Utah có bể bơi. Họ không khảo sát mọi người dân Utah. Mà họ chỉ khảo sát một số người đại diện cho thành phố. Điều này mô tả khái niệm phân tích dữ liệu nào?

A. Lấy mẫu (sample)

B. Ý nghĩa thống kê (Statistical significance)

C. Giới hạn sai số (Margin of error)

D. Độ tin cậy (Confidence level)

Đáp án: A

Bài tập

1. Một nhà phân tích dữ liệu đang làm việc với tập dữ liệu về một đợt gây quỹ mùa hè. Mặc dù họ có nhiều dữ liệu hữu ích, nhưng họ nhận ra rằng dữ liệu không đủ. Vì vậy, họ quyết định đợi đến cuối mùa để bắt đầu làm việc với tập dữ liệu. Ví dụ này mô tả loại không đủ dữ liệu nào?

A. Dữ liệu vẫn được cập nhật

B. Dữ liệu lỗi thời

C. Dữ liệu bị giới hạn địa lý

D. Dữ liệu bị giới hạn từ một nguồn

Đáp án: A

2.Một nhà phân tích dữ liệu đang thực hiện một dự án về chuỗi cung ứng toàn cầu. Họ có một tập dữ liệu với nhiều dữ liệu liên quan từ Châu Âu và Châu Á. Tuy nhiên, họ quyết định tạo dữ liệu mới đại diện cho tất cả các lục địa. Kịch bản này mô tả loại không đủ dữ liệu nào?

A. Dữ liệu bị giới hạn địa lý

B. Dữ liệu vẫn được cập nhật

C. Dữ liệu lỗi thời

D. Dữ liệu bị giới hạn từ một nguồn

Đáp án: A

3. Một nhà phân tích dữ liệu tại một công ty phần mềm muốn tìm hiểu thêm về các đối thủ cạnh tranh trong ngành. Bởi vì ngành công nghiệp phần mềm có rất nhiều sự phát triển. Nhà phân tích có một tập dữ liệu từ ba năm trước, và họ nhận thấy rằng nhiều công ty và sản phẩm trong tập dữ liệu đã thay đổi. Điều gì khiến nhà phân tích quyết định rằng dữ liệu không đủ?

A. Dữ liệu lỗi thời

B. Dữ liệu vẫn được cập nhật

C. Dữ liệu bị giới hạn địa lý

D. Dữ liệu bị giới hạn từ một nguồn

Đáp án: A

4. Một nhà sản xuất xe hơi muốn tìm hiểu thêm về sở thích thương hiệu của chủ sở hữu xe điện. Có hàng triệu người sở hữu ô tô điện trên thế giới. Công ty nên khảo sát ai?

A. Toàn bộ dân cư sở hữu xe điện

B. Một mẫu (sample) của tất cả các chủ sở hữu ô tô điện

C. Một mẫu (sample) chủ sở hữu ô tô gần đây đã mua một chiếc ô tô điện

D. Một mẫu (sample) chủ sở hữu ô tô đã sở hữu nhiều hơn một chiếc ô tô điện

Đáp án: B

5. Trong quá trình phân tích dữ liệu, mẫu dữ liệu (sample) có liên quan như thế nào đến tổng thể (population)?

A. Mẫu là một lựa chọn trùng lặp của dữ liệu được lấy từ tổng thể.

B. Mẫu là một phần của tổng thể và đại diện cho tổng thể.

C. Mẫu là giá trị trung bình của tất cả dữ liệu đại diện cho tổng thể.

D. Một mẫu là một ví dụ lý tưởng được lấy từ một tổng thể.

Đáp án: B

6. Khi thu thập dữ liệu thông qua một cuộc khảo sát, các công ty có thể tiết kiệm tiền bằng cách khảo sát 100% dân số.

A. Đúng

B. Sai

Đáp án: B

(Thu thập dữ liệu 100% dân số rất tốn kém)

7. Một nhà hàng muốn thu thập dữ liệu về một món ăn mới bằng cách đưa ra các mẫu thử miễn phí và yêu cầu phản hồi. Nhà hàng nên đưa mẫu cho ai?

A. Những thực khách chi nhiều tiền nhất cho bữa ăn của họ

B. Những thực khách sẵn sàng trả tiền cho các mẫu thử.

C. 80% thực khách

D. Toàn bộ thực khách

Đáp án: D

8. Một nhà hàng thu thập dữ liệu về một món ăn mới bằng cách cung cấp các mẫu thử miễn phí cho các bữa tiệc có từ sáu thực khách trở lên. Kịch bản này mô tả điều gì?

A. Lấy mẫu giới hạn địa lý.

B. Lấy mẫu không thiên vị.

C. Lấy mẫu thiên vị.

D. Lấy mẫu ngẫu nhiên.

Đáp án: C

(Kịch bản này mô tả sai lệch lấy mẫu vì các bên gồm sáu người trở lên không đại diện cho toàn bộ tổng thể.)

9. Điền vào chỗ trống: Sai lệch lấy mẫu trong thu thập dữ liệu xảy ra khi một mẫu không đại diện cho _____.

A. Tổng thể

B. Một phần của tổng thể

C. Một tập dữ liệu về tổng thể

D. Tổng thể bị ảnh hưởng nhiều nhất bởi dữ liệu

Đáp án: A

10. Dữ liệu và mục tiêu kinh doanh có thể không phù hợp vì một số lý do. Vấn đề nào sau đây có thể ngăn cản sự liên kết? (có thể chọn nhiều đáp án)

A. Dữ liệu không đầy đủ.

B. Lấy mẫu thiên vị.

C. Toàn vẹn dữ liệu.

D. Trực quan hóa dữ liệu.

Đáp án: A,B

Kiểm tra tính đúng của dữ liệu

1. Một nhóm nghiên cứu chạy thử nghiệm để xác định xem hệ thống bảo mật mới có hiệu quả hơn phiên bản trước hay không. Loại kết quả nào được mong đợi để thử nghiệm có ý nghĩa thống kê?

A. Kết quả mang tính giả định và cần thử nghiệm thêm.

B. Kết quả có thật và không phải do cơ hội ngẫu nhiên gây ra

C. Kết quả không có khả năng xảy ra lần nữa

D. Kết quả không chính xác và cần được bỏ qua

Đáp án: B

2. Để có một mức độ tin cậy cao trong một cuộc khảo sát khách hàng, kích thước mẫu phải phản ánh chính xác những gì?

A. Dự đoán của các bên liên quan.

B. Các thành viên có giá trị nhất trong tổng thể.

C. Toàn bộ người trong tổng thể.

D. Các xu hướng từ các cuộc khảo sát khách hàng khác

Đáp án: C

3. Một nhà phân tích dữ liệu xác định cỡ mẫu thích hợp cho một cuộc khảo sát. Họ có thể kiểm tra công việc của mình bằng cách đảm bảo tỷ lệ phần trăm mức độ tin cậy cộng với tỷ lệ phần trăm lỗi cộng lại lên đến 100%.

A. Đúng

B. Sai

Đáp án: B

Đánh giá sai số của dữ liệu

1. Điền vào chỗ trống: “mức độ ____ mà kết quả phân tích mẫu sẽ khác với kết quả của tổng thể”.

A. Tối đa

B. Tối thiểu

C. Trung bình

D. Trung vị

Đáp án: A

2. Trong một cuộc khảo sát về một sản phẩm tẩy rửa mới, 75% người được hỏi cho biết họ sẽ mua lại sản phẩm đó. Biên độ sai số cho cuộc khảo sát là 5%. Dựa vào biên độ sai số, hãy cho biết khoảng phần trăm nào phản ánh phản ứng thực của quần thể?

A. Giữa 70% và 75%

B. Giữa 75% và 80%

C. Giữa 70% và 80%

D. Giữa 73% và 78%

Đáp án: C

Chương 2

Giới thiệu về dữ liệu sạch

1.Mô tả sự khác biệt giữa giá trị rỗng và số 0 trong tập dữ liệu.

A. Giá trị null chỉ ra rằng một giá trị không tồn tại. Số 0 là một con số.

B. Giá trị null đại diện cho giá trị bằng không. Số 0 đại diện cho một ô trống.

C. Giá trị rỗng biểu thị dữ liệu không hợp lệ. Số 0 là thiếu dữ liệu

D. Giá trị null đại diện cho một số không có ý nghĩa. Số 0 đại diện cho số 0

Đáp án: A

2. Các quy trình và thủ tục phổ biến nhất được xử lý bởi các kỹ sư dữ liệu (data engineer) là gì? (có thể chọn nhiều đáp án)

A. Phát triển, duy trì và kiểm tra cơ sở dữ liệu và các hệ thống liên quan

B. Xác minh kết quả phân tích dữ liệu

C. Cung cấp cho dữ liệu một cơ sở hạ tầng đáng tin cậy

D. Chuyển đổi dữ liệu thành một định dạng hữu ích để phân tích

Đáp án: A,C,D

3. Các quy trình và thủ tục phổ biến nhất được xử lý bởi các chuyên gia kho dữ liệu (data warehousing specialists) là gì? (có thể chọn nhiều đáp án)

A. Đảm bảo dữ liệu được làm sạch đúng cách

B. Đảm bảo dữ liệu có sẵn

C. Đảm bảo dữ liệu được an toàn

D. Đảm bảo dữ liệu được sao lưu để tránh mất mát

Đáp án: B,C,D

4. Một nhà phân tích dữ liệu đang làm sạch một tập dữ liệu. Họ muốn xác nhận rằng người dùng đã nhập chính xác mã zip gồm năm chữ số bằng cách kiểm tra dữ liệu trong một cột bảng tính. Ta nên làm như thế nào?

A. Sử dụng công cụ độ dài trường để chỉ định số ký tự trong mỗi ô trong cột.

B. Sử dụng hàm MAX để xác định giá trị lớn nhất trong các ô trong cột

C. Thay đổi chiều rộng cột để chỉ vừa với năm chữ số

D. Định dạng các ô trong cột dưới dạng số

Đáp án: A

Bắt đầu làm sạch dữ liệu

1. Mỗi bảng tính đều có định dạng riêng, điều này có thể khiến dữ liệu có vẻ không nhất quán. Các nhà phân tích dữ liệu sử dụng công cụ _____ để làm cho dữ liệu trở về định dạng ban đầu, giúp cho dữ liệu trực quan và nhất quán cho bảng tính của họ

A. xóa định dạng (clear formats)

B. kiểm tra chính tả (spellcheck)

C. định dạng có điều kiện (conditional formatting)

D. tự sửa lỗi (autocorrect)

Đáp án: A

2. Quá trình kết hợp hai hoặc nhiều tập dữ liệu thành một tập dữ liệu là gì?

A. Tổ hợp dữ liệu (Data composition)

B. Chuyển đổi dữ liệu (Data transferring)

C. Hợp nhất dữ liệu (Data Merging)

D. Xác thực dữ liệu (Data validation)

Đáp án: C

3. Trong phân tích dữ liệu, _____ mô tả hai hoặc nhiều bộ dữ liệu có thể hoạt động cùng nhau tốt như thế nào.

A. Sự căn chỉnh (alignment)

B. Sự đồng ý (agreement)

C. Sự thích hợp (suitability)

D. Sự tương thích (compatibility)

Đáp án: D

Bài tập

1. Điền vào chỗ trống: Ánh xạ dữ liệu là quá trình _____ trường từ nguồn dữ liệu này sang nguồn dữ liệu khác.

A. Kết hợp (matching)

B. Giải nén (extracting)

C. Hợp nhất (merging)

D. Liên kết (linking)

Đáp án: C

Làm sạch dữ liệu với bảng tính nâng cao

1. Mô tả mối quan hệ giữa chuỗi văn bản và chuỗi con

A. Chuỗi văn bản là một hàng dữ liệu trong bảng. Chuỗi con là một ô trong hàng đó.

B. Chuỗi văn bản là một nhóm các ký tự trong một ô. Chuỗi con là một tập hợp con nhỏ hơn của chuỗi văn bản đó.

C. Chuỗi văn bản là một cột dữ liệu trong một bảng. Chuỗi con là một ô trong cột đó

D. Chuỗi văn bản là danh sách các thuộc tính ở đầu các cột trong bảng. Chuỗi con là một thuộc tính duy nhất trong danh sách đó.

Đáp án: B

2. Nhà phân tích dữ liệu sử dụng hàm COUNTIF để đếm số lần giá trị nhỏ hơn 5 xảy ra giữa các ô bảng tính từ A2 đến A100. Cú pháp chính xác là gì?

A. =COUNTIF(A2:A100,"<5")

B. =COUNTIF(A2:A100,">5")

C. =COUNTIF(A2:A100,<5)

D. =COUNTIF(A2:A100,>5)

Đáp án: A

3. Để loại bỏ các khoảng trống ở đầu, cuối và lặp lại trong dữ liệu, các nhà phân tích sử dụng hàm ____.

A. LEFT

B. RIGHT

C. TRIM

D. MID

Đáp án: C

Bài tập

1. Định dạng có điều kiện (conditional formatting) là một công cụ bảng tính thay đổi cách các ô xuất hiện khi các giá trị đáp ứng một điều kiện cụ thể. Nhà phân tích dữ liệu có thể sử dụng định dạng có điều kiện để thực hiện tác vụ nào sau đây?

A. Để tính toán các phương trình toán học

B. Để xác định các ô trống hoặc thông tin bị thiếu

C. Để làm cho các ô nổi bật hơn để phân tích hiệu quả hơn

D. Để sắp xếp dữ liệu trong chuỗi ô thành một thứ tự có nghĩa

Đáp án: C

2. Là một phần của quá trình làm sạch dữ liệu, nhà phân tích dữ liệu tạo ra một quy tắc để đánh dấu bất kỳ ô trống nào bằng màu xanh lam sáng. Đây là một ví dụ về trực quan hóa dữ liệu.

A. Đúng

B. Sai

Đáp án: B

(Chỉ là ví dụ về định dạng có điều kiện.)

3. Điền vào chỗ trống: Định dạng có điều kiện là một công cụ bảng tính thay đổi cách _____ xuất hiện khi các giá trị đáp ứng một điều kiện cụ thể.

A. Bộ lọc (filters)

B. Ô (cells)

C. Biểu đồ (charts)

D. Truy vấn (queries)

Đáp án: B

4. Một nhà phân tích dữ liệu sử dụng hàm SPLIT để chia một chuỗi văn bản xung quanh một ký tự được chỉ định và đặt mỗi phân đoạn vào một ô mới, riêng biệt. Ký tự được chỉ định tách từng mục được gọi là gì?

A. Dấu phân cách (Delimiter)

B. Chuỗi con (Substring)

C. Đơn vị (Unit)

D. Phân đoạn (Partition)

Đáp án: A

5. Công cụ tách văn bản thành cột sử dụng dấu phân cách để thực hiện phân chia. Vậy nhiệm vụ của dấu phân cách (delimiters) là gì?

A. Để chia một cột thành hai

B. Để chỉ định nơi để tách một chuỗi văn bản

C. Để tách các chuỗi con trùng lặp

D. Để thay đổi định dạng của một cột văn bản

Đáp án: B

6. Một nhà phân tích dữ liệu đang làm việc với bảng tính sau, có chứa dữ liệu tên thành viên trong cột C. Họ muốn chia dữ liệu này bằng cách sử dụng *dấu gạch dưới* làm dấu phân cách, để họ được lưu trữ trong một cột và họ trong một cột khác. Người phân tích nên sử dụng công cụ nào?

A. Định dạng có điều kiện (conditional formatting)

B. Hàm MID

C. Bảng tổng hợp (pivot table)

D. Hàm SPLIT

Đáp án: D

7. Điền vào chỗ trống: Khi mô tả một hàm SUM có _____ là = *SUM (giá trị 1, giá trị 2)*

A. Cú pháp (syntax)

B. Chuẩn (standard)

C. Cấu trúc (structure)

D. Tập lệnh (script)

Đáp án: A

8.Để một hàm hoạt động bình thường, các nhà phân tích dữ liệu phải tuân theo cấu trúc định trước của từng hàm. Cấu trúc này được gọi là gì?

A. Thuật toán (algorithm)

B. Sự xác thực (validation)

C. Cú pháp (syntax)

D. Tổng kết (summary)

Đáp án: C

9.Điền vào chỗ trống: Một cấu trúc xác định trước bao gồm thông tin bắt buộc của một chức năng và vị trí thích hợp của nó được gọi là _____.

A. Chuẩn (standard)

B. Cấu trúc (structure)

C. Cú pháp (syntax)

D. Tập lệnh (script)

Đáp án: C

10.Ví dụ, ô B4 có giá trị là một chuỗi “372 W. Addison Street Brandon, FL 33510”. Để trích xuất mã bưu chính (năm chữ số cuối bên phải), ta sẽ dùng hàm gì?

A. =RIGHT(B4, 5)

B. =LEFT(B4, 5)

C. =RIGHT(5, B4)

D. =LEFT(5,B4)

Đáp án: A

11. Ví dụ, ô B2 có giá trị là một chuỗi “9912 School St. North Wales, PA 19454”. Để trích xuất mã bưu chính (năm chữ số cuối bên phải), ta sẽ dùng hàm gì?

A. =RIGHT(B2, 5)

B. =RIGHT(5, B2)

C. =LEFT(5, B2)

D. =LEFT(B2, 5)

Đáp án: A

12.Ví dụ, ô B2 có giá trị là một chuỗi “8621 Glendale Dr. Burlington, MA 01803”. Để trích xuất mã bưu chính (năm chữ số cuối bên phải), ta sẽ dùng hàm gì?

A. =RIGHT(B3, 5)

B. =LEFT(B3, 5)

C. =RIGHT(5, B3)

D. =LEFT(5, B3)

Đáp án: A

13. Một công ty khi tuyển dụng một người, sẽ cấp ID theo định dạng, năm tham gia công ty và chứng minh nhân dân. Giả sử năm tham gia được lưu trong ô A4 và chứng minh nhân dân được lưu ô B5. Ta muốn nối hai chuỗi lại với nhau, ta nên dùng hàm nào?

A. =CONCATENATE(A5,B5)

B. =CONCATENATE(A5!B5)

C. =CONCATENATE(A5*B5)

D. =CONCATENATE(A5+B5)

Đáp án: A

14. Một công ty khi tuyển dụng một người, sẽ cấp ID theo định dạng, năm tham gia công ty và chứng minh nhân dân. Giả sử năm tham gia được lưu trong ô A3 và chứng minh nhân dân được lưu ô B3. Ta muốn nối hai chuỗi lại với nhau, ta nên dùng hàm nào?

A. =CONCATENATE(A3,B3)

B. =CONCATENATE(A3!B3)

C. =CONCATENATE(A3+B3)

D. =CONCATENATE(A3*B3)

Đáp án: A

15. Một công ty khi tuyển dụng một người, sẽ cấp ID theo định dạng, năm tham gia công ty và chứng minh nhân dân. Giả sử năm tham gia được lưu trong ô A4 và chứng minh nhân dân được lưu ô B4. Ta muốn nối hai chuỗi lại với nhau, ta nên dùng hàm nào?

A. =CONCATENATE(A4,B4)

B. =CONCATENATE(A4*B4)

C. =CONCATENATE(A4!B4)

D. =CONCATENATE(A4+B4)

Đáp án: A

16. Một nhà phân tích dữ liệu đang làm việc với một bảng tính chứa doanh số bán hàng. Sản phẩm đắt nhất mà công ty của họ bán có giá 49,99$, vì vậy họ muốn nhanh chóng xác nhận rằng tất cả dữ liệu trong cột Bán hàng là 49,99$ trở xuống. Họ có thể sử dụng chức năng gì?

A. COUNTIF

B. SUM

C. COUNT

D. SUMIF

Đáp án: A (countif).

Ta có thể sử dụng COUNTIF, là một hàm trả về số lượng ô phù hợp với một giá

trị được chỉ định. Cụ thể, ta sẽ đếm COUNTIF các cô có gì trị lớn hơn 49.99$. Nếu đúng, thì kết quả phải là 0.

17. Một nhà phân tích đang thực hiện một dự án liên quan đến khách hàng từ Bogota. Họ nhận được một bảng tính với 5.000 dòng chứa thông tin khách hàng. Nhà phân tích có thể sử dụng hàm nào để xác nhận rằng cột "địa chỉ" chứa chuỗi "Bogota" chính xác 5.000 lần?

A. COUNTIF

B. SUM

C. COUNT

D. SUMIF

Đáp án: A

18. Một nhà phân tích đang làm sạch một tập dữ liệu mới chứa 500 dòng. Họ muốn đảm bảo dữ liệu chứa từ ô B2 đến ô B300 không chứa một số lớn hơn 50. Cú pháp hàm COUNTIF nào sau đây có thể được sử dụng để trả lời câu hỏi này? (Có thể chọn nhiều đáp án)

A. =COUNTIF(B2:B300,">50")

B. =COUNTIF(B2:B300,"<=50")

C. =COUNTIF(B2:B300,>50)

D. =COUNTIF(B2:B300,<=50)

Đáp án: A,B

19. VLOOKUP tìm kiếm một giá trị trong một hàng để trả về một phần thông tin tương ứng

A. Đúng

B. Sai

Đáp án: B (sai).
(VLOOKUP hàm tìm kiếm giá trị nhất định trong một cột để trả về thông tin tương ứng.)

20. V trong hàm VLOOKUP là viết tắt của gì?

A. Virtual

B. Vertical

C. Variable

D. Visual

Đáp án: B

21. Một nhà phân tích dữ liệu muốn tìm kiếm một giá trị nhất định trong một cột, sau đó trả về một phần thông tin tương ứng. Họ nên sử dụng hàm gì?

A. VALUE

B. MATCH

C. VLOOKUP

D. FIND

Đáp án: C

22. Để đánh giá mức độ hoạt động của hai hoặc nhiều nguồn dữ liệu cùng nhau, các nhà phân tích dữ liệu sử dụng ánh xạ dữ liệu (data mapping).

A. Đúng

B. Sai

Đáp án: A

23. Một nhà phân tích dữ liệu cần kết hợp hai bộ dữ liệu. Mỗi tập dữ liệu đến từ một hệ thống khác nhau và các hệ thống lưu trữ dữ liệu theo những cách khác nhau. Người phân tích dữ liệu có thể làm gì để đảm bảo dữ liệu tương thích?

A. Áp dụng cấu trúc dữ liệu

B. Sử dụng trực quan hóa dữ liệu

C. Hợp nhất dữ liệu

D. Ánh xạ dữ liệu

Đáp án: D

Chương 3

Làm việc với cơ sở dữ liệu

1. Điều nào sau đây là ưu điểm của việc sử dụng SQL?

A. SQL có thể xử lý một lượng lớn dữ liệu.

B. SQL có thể được sử dụng để lập trình bộ vi xử lý trên máy chủ cơ sở dữ liệu.

C. SQL có thể được điều chỉnh và sử dụng với nhiều loại CSDL.

D. SQL cung cấp các công cụ mạnh mẽ để làm sạch dữ liệu.

Đáp án: A,C,D

2. Các nhà phân tích dữ liệu có thể thực hiện tác vụ nào sau đây bằng cách sử dụng cả bảng tính và SQL (có thể chọn nhiều đáp án)

A. Tính toán số học

B. Sử dụng công thức

C. Hợp nhất dữ liệu

D. Xử lý dữ liệu lớn hiệu quả

Đáp án: A,B,C

3SQL là một ngôn ngữ được sử dụng để giao tiếp với cơ sở dữ liệu. Giống như hầu hết các ngôn ngữ, SQL có các biến thể. Lợi ích của việc học và sử dụng SQL chuẩn là gì?

A. SQL chuẩn có thể được dịch tự động sang các biển thể khác.

B. SQL chuẩn dễ học hơn nhiều so với các phương ngữ khác.

C. SQL chuẩn hoạt động với phần lớn các cơ sở dữ liệu.

D. SQL tiêu chuẩn yêu cầu một số thay đổi cú pháp nhỏ để thích ứng với các phương ngữ khác.

Đáp án: C,D

Bài tập

1. Một nhà phân tích dữ liệu đang phân tích dữ liệu y tế cho một công ty bảo hiểm sức khỏe. Tập dữ liệu chứa hàng tỷ dòng dữ liệu. Công cụ nào sau đây sẽ xử lý dữ liệu hiệu quả nhất?

A. Bảng tính

B. SQL

C. Bài thuyết trình

D. Tệp văn bản

Đáp án: B

2. Các nhà phân tích dữ liệu chọn SQL vì lý do nào sau đây? (có thể chọn nhiều đáp án)

A. SQL là một tiêu chuẩn nổi tiếng trong cộng đồng chuyên nghiệp

B. SQL là một chương trình phần mềm mạnh mẽ

C. SQL là một ngôn ngữ lập trình cũng có thể tạo các ứng dụng web

D. SQL có thể xử lý một lượng lớn dữ liệu

Đáp án: A,D

3. Điền vào chỗ trống: Các nhà phân tích dữ liệu thường sử dụng _____ để xử lý các tập dữ liệu rất lớn.

A. Bảng tính

B. SQL

C. Tệp văn bản

D. Trang web

Đáp án: B

4.Nhà phân tích dữ liệu sẽ sử dụng SQL thay vì bảng tính trong tình huống nào sau đây? (Có thể chọn nhiều đáp án)

A. Khi làm việc với một lượng lớn dữ liệu

B. Khi ghi lại các truy vấn và thay đổi trong suốt một dự án

C. Khi sử dụng hàm COUNTIF để tìm một thông tin cụ thể

D. Khi nhanh chóng lấy thông tin từ nhiều nguồn khác nhau trong cơ sở dữ liệu

Đáp án: A,B,D

5. Một số lợi ích của việc sử dụng SQL để phân tích là gì?

A. SQL tương tác với các chương trình cơ sở dữ liệu.

B. SQL theo dõi các thay đổi trong một nhóm.

C. SQL có thể lấy thông tin từ các nguồn cơ sở dữ liệu khác nhau

D. SQL có các hàm tích hợp sẵn.

Đáp án: A,B,C

6. Nhà phân tích dữ liệu sẽ sử dụng bảng tính thay vì SQL trong tình huống nào sau đây? (có thể chọn nhiều đáp án)

A. Khi làm việc với một tập dữ liệu nhỏ

B. Khi sử dụng một ngôn ngữ để tương tác với nhiều chương trình cơ sở dữ liệu

C. Khi làm việc với tập dữ liệu có hơn 1.000.000 dòng

D. Khi kiểm tra dữ liệu trực quan

Đáp án: Một nhà phân tích sẽ chọn sử dụng bảng tính thay vì SQL khi kiểm tra dữ liệu trực quan hoặc làm việc với một tập dữ liệu nhỏ

Truy vấn với SQL

1. Các nhà phân tích dữ liệu có thể sử dụng hàm SQL nào sau đây để làm sạch các biến chuỗi?

A. LENGTH

B. SUBSTRING

C. COUNTIF

D. TRIM

Đáp án: A,C,D

2. Bạn đang làm việc với một bảng cơ sở dữ liệu có tên là playlist chứa dữ liệu về danh sách các bài nhạc. Bảng bao gồm các cột playlist_id và name. Bạn muốn truy xuất tên bài nhạc mà không có tên nào trùng lặp và sắp xếp kết quả theo id_playlist.
Hãy viết câu truy vấn cho yêu cầu bên trên

Đáp án: SELECT DISTINCT name
FROM
playlist
ORDER BY
playlist_id

3. Bạn đang làm việc với một bảng cơ sở dữ liệu chứa dữ liệu về album nhạc. Bảng bao gồm các cột cho album_id, title, artist_id. Bạn muốn truy vấn các tiêu đề album có độ dài dưới 4 ký tự.
Hãy viết câu truy vấn cho yêu cầu bên trên

Đáp án: SELECT
title
FROM
album
WHERE
LENGTH(album.title) < 4

4. Bạn đang làm việc với một bảng cơ sở dữ liệu có chứa dữ liệu khách hàng. Bảng bao gồm các cột về vị trí của khách hàng: city, state, country. Bạn muốn lấy 3 chữ cái đầu tiên của mỗi tên quốc gia. Bạn quyết định sử dụng hàm SUBSTR để truy xuất 3 chữ cái đầu tiên của mỗi tên quốc gia và sử dụng lệnh AS để lưu trữ kết quả trong một cột mới có tên là new_country.
Hãy viết câu truy vấn cho yêu cầu bên trên

Đáp án: SELECT
SUBSTR(country, 1, 3) AS new_country
FROM
Customer

Bài tập

1. Một nhà phân tích dữ liệu chạy một truy vấn SQL để trích xuất một số dữ liệu từ cơ sở dữ liệu để phân tích thêm. Người phân tích có thể lưu kết quả truy vấn bằng cách nào? (có thể chọn nhiều đáp án)

A. Tải xuống dữ liệu dưới dạng bảng tính.

B. Chạy truy vấn SQL để tự động lưu dữ liệu.

C. Sử dụng truy vấn CẬP NHẬT để lưu dữ liệu.

D. Tạo một bảng mới cho dữ liệu.

Đáp án: A,D

2. Một nhà phân tích dữ liệu tạo nhiều bảng mới trong cơ sở dữ liệu của công ty họ. Khi dự án hoàn thành, nhà phân tích muốn loại bỏ các bảng để chúng không làm lộn xộn cơ sở dữ liệu. Họ có thể sử dụng lệnh SQL nào để xóa bảng?

A. CREATE TABLE IF NOT EXISTS

B. UPDATE

C. INSERT INTO

D. DROP TABLE IF EXISTS

Đáp án: D

3. Một nhà phân tích dữ liệu đang quản lý cơ sở dữ liệu thông tin khách hàng cho một cửa hàng bán lẻ. Người phân tích có thể sử dụng lệnh SQL nào để thêm khách hàng mới vào cơ sở dữ liệu?

A. DROP TABLE IF EXISTS

B. INSERT INTO

C. UPDATE

D. CREATE TABLE IF NOT EXISTS

Đáp án: B

Truy vấn và làm sạch với SQL nâng cao

Bài tập

1.Bạn đang làm việc với một bảng cơ sở dữ liệu có chứa dữ liệu hóa đơn. Bảng này bao gồm các cột cho id_hoa_don và trang_thai_thanh_toan. Bạn muốn xóa các mục trùng lặp cho trang_thai_thanh_toan và sắp xếp kết quả theo id_hoa_don.

Hãy viết câu lệnh truy vấn.

Đáp án: SELECT DISTINCT trang_thai_thanh_toan
FROM invoice
ORDER BY id_hoa_don

2.Bạn đang làm việc với một bảng cơ sở dữ liệu có chứa dữ liệu hóa đơn. Bảng này bao gồm các cột cho id_hoa_don và id_khach_hang. Bạn muốn xóa các mục trùng lặp cho id_khach_hang và sắp xếp kết quả theo id_hoa_don.Hãy viết câu lệnh truy vấn.

Đáp án: SELECT DISTINCT id_khach_hang
FROM invoice
ORDER BY id_hoa_don

3. Bạn đang làm việc với một bảng cơ sở dữ liệu có chứa dữ liệu hóa đơn. Bảng này bao gồm các cột cho id_hoa_don và thanh_pho. Bạn muốn xóa các mục trùng lặp cho thanh_pho và sắp xếp kết quả theo id_hoa_don.

Hãy viết câu lệnh truy vấn.

Đáp án: SELECT DISTINCT thanh_pho
FROM invoice

ORDER BY id_hoa_don

4. Bạn đang làm việc với một bảng cơ sở dữ liệu có chứa dữ liệu khách hàng (bảng customer). Bảng này bao gồm các cột về vị trí của khách hàng, chẳng hạn như city, state, country, and postal_code. Tên tiểu bang được viết tắt. Bạn muốn in ra tất cả thông tin của khách hàng mà có tên tiểu bang (state) dài hơn 2 ký tự.

Hãy viết câu truy vấn.

Đáp án: SELECT *
FROM customer
WHERE LENGTH(state) > 2

5.Bạn đang làm việc với một bảng cơ sở dữ liệu có chứa dữ liệu khách hàng (bảng customer). Bảng này bao gồm các cột về vị trí của khách hàng, chẳng hạn như city, state, country, and postal_code. Tên tiểu bang được viết tắt. Bạn muốn in ra tất cả thông tin của khách hàng mà có postal_code dài hơn 7 ký tự.

Hãy viết câu truy vấn.

Đáp án: SELECT *
FROM customer
WHERE LENGTH(postal_code) > 7

6.Bạn đang làm việc với một bảng cơ sở dữ liệu có chứa dữ liệu khách hàng (bảng customer). Bảng này bao gồm các cột về vị trí của khách hàng, chẳng hạn như city, state, country, and postal_code. Tên tiểu bang được viết tắt. Bạn muốn in ra tất cả thông tin của khách hàng mà có city dài hơn 9 ký tự.
Hãy viết câu truy vấn.

Đáp án: SELECT *
FROM customer
WHERE LENGTH(city) > 9

7.Một nhà phân tích dữ liệu đang làm sạch dữ liệu giao thông vận tải cho một

công ty đi chung xe. Nhà phân tích chuyển đổi dữ liệu về thời gian đi xe từ kiểu chuỗi thành float. Kịch bản này mô tả điều gì?

A. Processing (Xử lý dữ liệu)

B. Visualizing (Biểu diễn dữ liệu)

C. Calculating (Tính toán)

D. Typecasting (Ép kiểu dữ liệu)

Đáp án: D

8. Điền vào chỗ trống: _____ để cập đến quá trình chuyển đổi dữ liệu từ loại này sang loại khác.

A. Cleaning (làm sạch dữ liệu)

B. Typecasting (ép kiểu dữ liệu)

C. Formatting (định dạng dữ liệu)

D. Querying (truy vấn dữ liệu)

Đáp án: B

9. Trong cơ sở dữ liệu SQL, kiểu dữ liệu nào đề cập đến một số có chứa một số thập phân?

A. String

B. Integer

C. Boolean

D. Float

Đáp án: D

10. Điền vào chỗ trống: Trong cơ sở dữ liệu SQL, hàm _____ có thể được sử dụng để chuyển đổi dữ liệu từ kiểu dữ liệu này sang kiểu dữ liệu khác.

A. TRIM

B. CAST

C. LENGTH

D. SUBSTR

Đáp án: B

11. Một nhà phân tích dữ liệu đang làm việc với dữ liệu bán sản phẩm. Họ nhập dữ liệu mới vào cơ sở dữ liệu. Cơ sở dữ liệu nhận dạng dữ liệu về giá sản phẩm dưới dạng chuỗi văn bản. Người phân tích có thể sử dụng hàm SQL nào để chuyển đổi chuỗi văn bản thành float?

A. CAST

B. SUBSTR

C. TRIM

D. LENGTH

Đáp án: A

12. Hàm CAST có thể được sử dụng để chuyển đổi kiểu dữ liệu DATE thành kiểu dữ liệu DATETIME.

A. Đúng

B. Sai

Đáp án: A

13. Một nhà phân tích dữ liệu đang làm sạch dữ liệu khảo sát. Kết quả cho một câu hỏi tùy chọn chứa nhiều giá trị rỗng. Người phân tích có thể sử dụng chức năng nào để loại bỏ các giá trị rỗng khỏi kết quả?

A. CAST

B. LENGTH

C. COALESCE

D. CONCAT

Đáp án: C

14. Hàm SQL nào cho phép bạn nối các chuỗi lại với nhau để tạo chuỗi văn bản mới?

A. CONCAT

B. CAST

C. COALESCE

D. LENGTH

Đáp án: A

15. Điền vào chỗ trống: Hàm _____ có thể được sử dụng để trả về các giá trị không NULL trong danh sách.

A. TRIM

B. CAST

C. CONCAT

D. COALESCE

Đáp án: D

16. Bạn đang làm việc với một bảng cơ sở dữ liệu có chứa dữ liệu hóa đơn (bảng invoice). Bảng bao gồm các cột về vị trí thanh toán như billing_city, billing_state và billing_country. Bạn muốn lấy 4 chữ cái đầu tiên của mỗi tên thành phố. Bạn quyết định sử dụng hàm SUBSTR để lấy 4 chữ cái đầu tiên của mỗi tên thành phố (billing_city) và sắp xếp billing_city theo thứ tự tăng dần.
Hãy viết câu truy vấn

Đáp án: SELECT SUBSTR(billing_city, 1, 4)
FROM invoice
ORDER BY billing_city

17. Bạn đang làm việc với một bảng cơ sở dữ liệu có chứa dữ liệu khách hàng (bảng customer). Bảng bao gồm các cột về vị trí của khách hàng như city, state, and country. Bạn muốn lấy 2 chữ cái đầu tiên của mỗi tên tiểu bang (state). Bạn quyết định sử dụng hàm SUBSTR để truy xuất 2 chữ cái đầu tiên của mỗi tên tiểu bang (state) và sắp xếp tiểu bang (state) theo thứ tự giảm dần. Hãy viết câu truy vấn.

Đáp án: SELECT SUBSTR(state, 1, 2)
FROM customer
ORDER BY state DESC

18. Bạn đang làm việc với một bảng cơ sở dữ liệu có chứa dữ liệu nhân viên

(bảng employee). Bảng bao gồm các cột về vị trí như city, state, and country, postal_code. Bạn muốn lấy 3 chữ cái đầu tiên của mỗi postal_code. Bạn quyết định sử dụng hàm SUBSTR để truy xuất 3 chữ cái đầu tiên của mỗi postal_code và sắp xếp postal_code theo thứ tự tăng dần.
Hãy viết câu truy vấn.

Đáp án: SELECT SUBSTR(postal_code, 1, 3)
FROM employee
ORDER BY postal_code

Chương 4

Quản lý và bảo mật dữ liệu

Câu hỏi về làm thế nào tổ chức dữ liệu

1. Đảm bảo dữ liệu được xác minh đúng cách là một phần quan trọng của quá trình làm sạch dữ liệu. Nhiệm vụ nào sau đây liên quan đến việc xác minh này? Có thể chọn nhiều câu trả lời.

A. Kiểm tra lại quá trình làm sạch dữ liệu

B. Xem xét liệu dữ liệu có đáng tin cậy và phù hợp với dự án hay không

C. Sửa bất kỳ lỗi nào được tìm thấy trong dữ liệu theo cách thủ công

D. Yêu cầu các bên liên quan kiểm tra và xác nhận dữ liệu là sạch

Đáp án: A,B,C

2. Điền vào chỗ trống: Để đếm tổng số giá trị bảng tính trong một phạm vi được chỉ định, nhà phân tích dữ liệu sử dụng hàm _____.

A. COUNTA

B. TOTAL

C. SUM

D. WHOLE

Đáp án: A

3.Một nhà phân tích dữ liệu đang làm sạch một tập dữ liệu có các định dạng không nhất quán và các trường hợp lặp lại. Họ sử dụng hàm TRIM để loại bỏ các khoảng trắng thừa khỏi các biến chuỗi. Họ có thể sử dụng những công cụ nào khác để làm sạch dữ liệu?

A. Tìm kiếm và thay thế (Find and replace)

B. Xóa bản trùng lặp (Remove duplicates)

C. Nhập dữ liệu (Import data)

D. Bảo vệ trang tính (Protect data)

Đáp án: A,B

4. Để sửa lỗi đánh máy trong cột cơ sở dữ liệu, bạn nên chèn câu lệnh CASE vào một truy vấn ở đâu?

A. Bên trong câu lệnh SELECT

B. Bên trong câu lệnh FROM

C. Bên trong câu lệnh ORDER BY

D. Bên trong câu lệnh GROUP BY

Đáp án: A

Bài tập

1. Điền vào chỗ trống: Sau khi dữ liệu sạch, nhà phân tích dữ liệu chuyển sang _____ và xác minh.

A. Báo cáo (reporting)

B. Xử lý (processing)

C. Xác nhận (confirming)

D. Công khai (publishing)

Đáp án: A

2. Quá trình xác minh và báo cáo xảy ra trước quá trình làm sạch dữ liệu.

A. Đúng

B. Sai

Đáp án: B

3. Dữ liệu được thu thập cho một dự án phân tích vừa được làm sạch. Các bước tiếp theo cho một nhà phân tích dữ liệu là gì?

A. Xác minh (verification)

B. Báo cáo (reporting)

C. Kiểm thử (validation)

D. Chứng nhận (Certification)

Đáp án: A,B

4. Bước đầu tiên trong quy trình xác minh là gì?

A. So sánh dữ liệu đã được làm sạch với tập dữ liệu gốc, chưa được làm sạch và so sánh hai tập dữ liệu với nhau.

B. Tạo danh sách theo thứ tự thời gian về các sửa đổi được thực hiện đối với dữ liệu.

C. Xác định chất lượng của dữ liệu.

D. Thông báo cho người khác về quá trình làm sạch dữ liệu của bạn.

Đáp án: A

5. Một nhà phân tích dữ liệu đang trong bước xác minh. Họ xem xét vấn đề kinh doanh, mục tiêu và dữ liệu liên quan đến dự án phân tích của họ. Kịch bản này mô tả điều gì?

A. Quan sát bức tranh lớn

B. Báo cáo về dữ liệu

C. Trực quan hóa dữ liệu

D. Xem xét các bên liên quan

Đáp án: A

6. Điều gì liên quan đến việc nhìn thấy bức tranh lớn khi xác minh làm sạch dữ liệu? (có thể chọn nhiều đáp án)

A. Xem xét mục tiêu kinh doanh.

B. Xem xét vấn đề ban đầu đặt ra.

C. Xem xét dữ liệu.

D. Xem xét báo cáo.

Đáp án: A,B,C

7. Hàm nào loại bỏ khoảng trắng ở đầu, cuối và lặp lại trong dữ liệu?

A. TRIM

B. CUT

C. CROP

D. TIDY

Đáp án: A

8. Chức năng nào sau đây tự động loại bỏ khoảng trống thừa khi làm sạch dữ liệu?

A. TRIM

B. REMOVE

C. CLEAR

D. SNIP

Đáp án: A

9. Điền vào chỗ trống: TRIM là một hàm loại bỏ _____ khoảng trắng trong dữ liệu

A. Phía trước

B. Phía sau

C. Lặp lại

D. Bên trong

Đáp án: A,B,C

10. Trong khi xác minh dữ liệu đã được làm sạch, nhà phân tích dữ liệu gặp phải tên sai chính tả. Họ có thể sử dụng hàm nào để xác định xem lỗi sai chính tả đó có lặp lại trong toàn bộ tập dữ liệu hay không?

A. COUNTA

B. COUNT

C. CHECK

D. CASE

Đáp án: A.

Để xác định xem lỗi có lặp lại trong toàn bộ tập dữ liệu hay không, ta có thể sử dụng COUNTA

11. Một nhà phân tích dữ liệu sử dụng hàm COUNTA để đếm giá trị nào sau đây?

A. Tổng số giá trị trong một phạm vi được chỉ định.

B. Tổng số tiêu đề trong một phạm vi cụ thể.

C. Tổng số mục nhập trong bảng ghi thay đổi.

D. Các số cụ thể trong tập dữ liệu.

Đáp án: A

12. Nhà phân tích dữ liệu có thể sử dụng công cụ nào để tìm ra bao nhiêu lỗi giống nhau xảy ra trong một tập dữ liệu?

A. COUNTA

B. CONFIRM

C. CASE

D. COUNT

Đáp án: A

13. Công cụ SQL nào xem xét một hoặc nhiều điều kiện, sau đó trả về một giá trị ngay sau khi một điều kiện được đáp ứng?

A. CASE

B. THEN

C. WHEN

D. ELSE

Đáp án: A.

14. Điền vào chỗ trống: Một nhà phân tích dữ liệu sử dụng câu lệnh CASE để xem xét một hoặc nhiều _____, sau đó trả về một giá trị.

A. Điều kiện (conditions)

B. Nhận dạng (identifications)

C. Bổ sung (additions)

D. Sự thay đổi (changes)

Đáp án: A

15. Câu lệnh WHEN xem xét một hoặc nhiều điều kiện và trả về một giá trị ngay sau khi điều kiện đó được đáp ứng.

A. Đúng

B. Sai

Đáp án: B

Dùng câu lệnh CASE

16. Điền vào chỗ trống: Trong khi làm sạch dữ liệu, nhà phân tích dữ liệu có thể sử dụng bảng thay đổi để giữ danh sách theo thứ tự thời gian về những thay đổi mà họ thực hiện. Họ có thể tham khảo nó trong thời gian _____ nếu có sai sót hoặc thắc mắc.

A. Xác minh (verification)

B. Tài liệu (documentation)

C. Trực quan hóa (visualization)

D. Trình bày (presenting)

Đáp án: A

17. Tại thời điểm nào trong toàn bộ quá trình phân tích, nhà phân tích dữ liệu sử dụng một bảng ghi thay đổi?

A. Trong khi làm sạch dữ liệu

B. Trong khi trực quan hóa dữ liệu

C. Trong khi thu thập dữ liệu

D. Trong khi báo cáo dữ liệu

Đáp án: A

(Quá trình làm sạch dữ liệu, sẽ tạo bảng ghi thay đổi)

18. Điền vào chỗ trống: Bảng thay đổi gồm _____ danh sách các sửa đổi được thực hiện đối với một dự án.

A. Theo thời gian (chronological)

B. Ngẫu nhiên (random)

C. Xấp xỉ (approximate)

D. Đồng bộ (synchronized)

Đáp án: A

Ghi nhận kết quả của quá trình làm sạch

1. Tại sao một nhà phân tích dữ liệu lại quan trọng trong việc ghi lại sự phát triển của một tập dữ liệu? (có thể chọn nhiều đáp án)

A. Để khôi phục lỗi làm sạch dữ liệu

B. Để thông báo cho những người dùng khác về những thay đổi

C. Để xác định chất lượng của dữ liệu

D. Để xác định các phương pháp hay nhất trong việc thu thập dữ liệu

Đáp án: A,B,C

2. Điền vào chỗ trống: Trong khi làm sạch dữ liệu, tài liệu được sử dụng để theo dõi _____. (có thể chọn nhiều đáp án)

A. sự xóa (deletions)

B. sai sót (errors)

C. thay đổi (changes)

D. thiên vị (bias)

Đáp án: A,B,C

3. Việc lập tài liệu (document) làm sạch dữ liệu giúp bạn có thể đạt được những mục tiêu nào?

A. Minh bạch về quy trình làm sạch.

B. Giữ các thành viên trong nhóm đều nắm tình hình.

C. Chứng minh với các bên liên quan của dự án về lòng tin và trách nhiệm.

D. Trực quan hóa kết quả phân tích dữ liệu của bạn.

Đáp án: A,B,C

Bài tập

1. Điền vào chỗ trống: Tài liệu (document) là quá trình theo dõi _____ trong quá trình làm sạch dữ liệu. (có thể chọn nhiều đáp án).

A. Sự xóa bỏ (deletions)

B. Sự thay đổi (changes)

C. Sự thêm vào (deletions)

D. Không hoạt động (inactivity)

Đáp án: A,B,C

2. Một nhà phân tích dữ liệu sử dụng một bảng ghi thay đổi trong khi làm sạch dữ liệu. Quy trình nào được bảng ghi thay đổi?

A. Ghi tài liệu (Documentation)

B. Kiểm tra (Examination)

C. Biện minh (Illumination)

D. Tiết lộ (Disclosure)

Đáp án: A

3. Quá trình theo dõi các thay đổi, bổ sung, xóa và lỗi trong quá trình làm sạch dữ liệu là gì?

A. Ghi tài liệu (Documentation)

B. Lập danh mục (Cataloging)

C. Sự ghi lại (Recording)

D. Sự quan sát (Observation)

Đáp án: A

4. Một nhà phân tích dữ liệu thực hiện các thay đổi đối với các truy vấn SQL và sử dụng các nhận xét này để tạo một bảng ghi thay đổi. Điều này giải thích những thay đổi đã thực hiện và lý do tại sao họ thực hiện chúng.

A. Đúng

B. Sai

Đáp án: A

Nhà phân tích dùng các comment trong code, để diễn giải code và xem đó như là bảng ghi thay đổi

5. Một nhà phân tích dữ liệu đã thực hiện một truy vấn tới kho lưu trữ dưới dạng một truy vấn mới và được cải tiến. Sau đó, họ chỉ định những thay đổi họ đã thực hiện và lý do họ thực hiện chúng. Kịch bản này là một phần của quá trình nào?

A. Tạo bảng ghi thay đổi

B. Báo cáo dữ liệu

C. Trực quan hóa dữ liệu

D. Giao tiếp với các bên liên quan

Đáp án: A

6. Xem lại lịch sử phiên bản là một cách hiệu quả để xem bảng thay đổi trong SQL.

A. Đúng

B. Sai

Đáp án: B

(SQL không có lịch sử thay đổi. Lịch sử thay đổi là chức năng của bảng tính)