Google Career Certificates

fit@hcmus

Grow with Google | NIC

Chuẩn Bị Dữ Liệu để Khám phá

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

Năm 2022

1 CHUẨN BỊ DỮ LIỆU

Nội dung

Thu thập dữ liệu

Sự khác biệt giữa các định dạng và cấu trúc dữ liệu

Khám phá kiểu dữ liệu, trường dữ liệu và giá trị dữ liệu

Cách tạo ra dữ liệu

Dữ liệu tạo ra bởi người dùng

- Các bức ảnh và video upload lên các mạng xã hội.

Thu thập thông tin

- Phỏng vấn
- Quan sát
- Form được điền vào
- Bảng câu hỏi
- Khảo sát
- Cookies

Quyết định dữ liệu sẽ thu thập

Xác định loại dữ liệu nào cần thu thập và sử dụng cho mỗi dự án.

Các nhân tố cần xem xét:

- Cách thu thập dữ liệu
- Nguồn dữ liệu
- Loại dữ liệu được sử dụng
- Khối lượng dữ liệu cần thu thập
- Kiểu dữ liệu
- Khung thời gian để thu thập dữ liệu: dữ liệu lịch sử (historical data) hay tự thu thập

Nguồn dữ liệu

- Dữ liệu của bên thứ nhất
- Dữ liệu của bên thứ hai
- Dữ liệu của bên thứ ba

Nguồn dữ liệu

- Dữ liệu của bên thứ nhất: dữ liệu thu thập bởi cá nhân hay tổ chức dùng chính tài nguyên của họ
- Dữ liệu của bên thứ hai
- Dữ liệu của bên thứ ba

Nguồn dữ liệu

- Dữ liệu của bên thứ nhất
- Dữ liệu của bên thứ hai: một tổ chức sẽ thu thập dữ liệu trực tiếp từ chính khách hàng của họ rồi đem bán
- Dữ liệu của bên thứ ba

Nguồn dữ liệu

- Dữ liệu của bên thứ nhất
- Dữ liệu của bên thứ hai
- Dữ liệu của bên thứ ba: dữ liệu được thu thập bởi một bên thứ ba, bên thứ ba này cũng không thu thập trực tiếp

Tập hợp và mẫu

- **Tập hợp (population)**: tất cả các giá trị dữ liệu có thể có trong một tập dữ liệu nhất định.
- **Mẫu (sample)** là một phần của tập hợp đại diện cho tập hợp.

Chọn lựa đúng dữ liệu

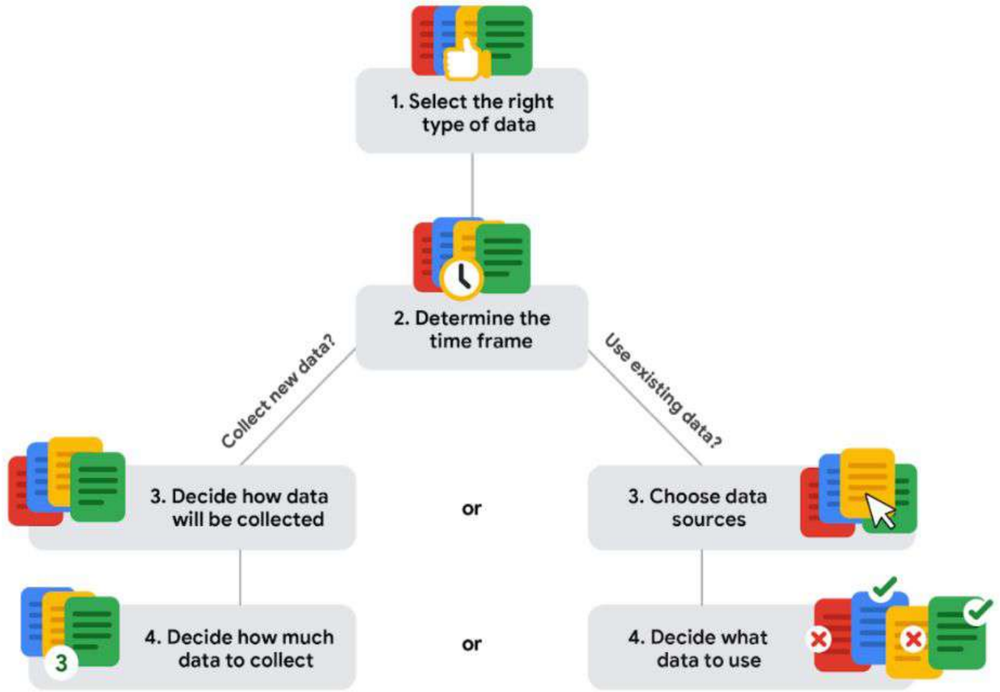
1. Chọn đúng kiểu dữ liệu
2. Chọn khung thời gian

3.1 Dữ liệu nào được thu thập

4.1 Bao nhiêu dữ liệu thu thập

3.2 Chọn nguồn dữ liệu

4.2 Quyết định loại dữ liệu nào được dùng



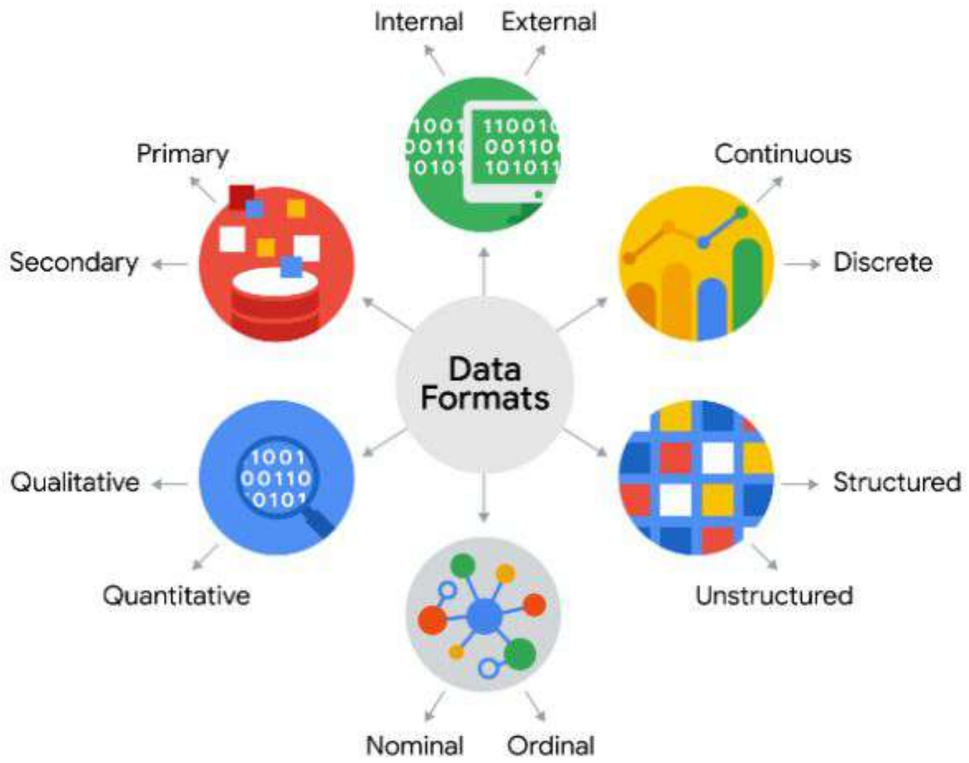
Nguồn: https://www.coursera.org/learn/data-preparation/supplement/7iFqv/selecting-the-right-data

Nội dung

- Thu thập dữ liệu
- **Sự khác biệt giữa các định dạng và cấu trúc dữ liệu**
- Khám phá kiểu dữ liệu, trường dữ liệu và giá trị dữ liệu

Định dạng dữ liệu

Chọn định dạng dữ liệu phù hợp sẽ giúp quản lý và sử dụng dữ liệu tốt nhất.



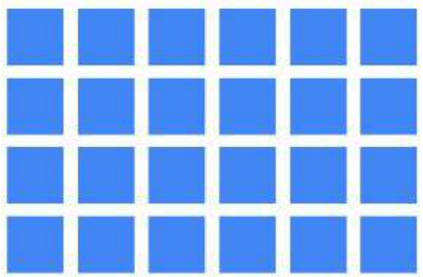
https://www.coursera.org/learn/data-preparation/supplement/mBSNa/data-formats-in-practice

Định dạng dữ liệu

- Dữ liệu chính (primary data) và Dữ liệu thứ cấp (secondary data)
- Dữ liệu nội bộ (internal data) và Dữ liệu bên ngoài (external data)
- Dữ liệu liên tục và dữ liệu rời rạc
- Dữ liệu định tính và dữ liệu định lượng
- Dữ liệu có thứ tự (nominal) và không thứ tự (ordinal)
- Có cấu trúc và phi cấu trúc

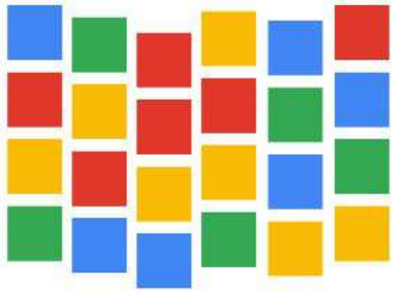
Có cấu trúc

- Kiểu dữ liệu được xác định
- Dữ liệu định lượng
- Dễ tổ chức, tìm kiếm, phân tích
- CSDL quan hệ và nhà kho dữ liệu
- Chứa dòng và cột
- Ví dụ: Excel, Google sheet, SQL, thông tin khách hàng.



Phi cấu trúc

- Nhiều kiểu dữ liệu khác nhau
- Dữ liệu định tính, khó tìm kiếm
- Tự do hơn để phân tích
- Hồ dữ liệu, kho dữ liệu, và NoSQL
- Không thể đặt trong dòng và cột
- Ví dụ: bài review sản phẩm, hình ảnh, âm thanh, video



Mô hình hóa dữ liệu

- Mô hình hóa dữ liệu (data modelling): quá trình tạo ra các sơ đồ thể hiện một cách trực quan cách dữ liệu được tổ chức và cấu trúc.
- Các biểu diễn trực quan này được gọi là mô hình dữ liệu.

Các mức độ mô hình hóa dữ liệu

Mô hình hóa dữ liệu khái niệm

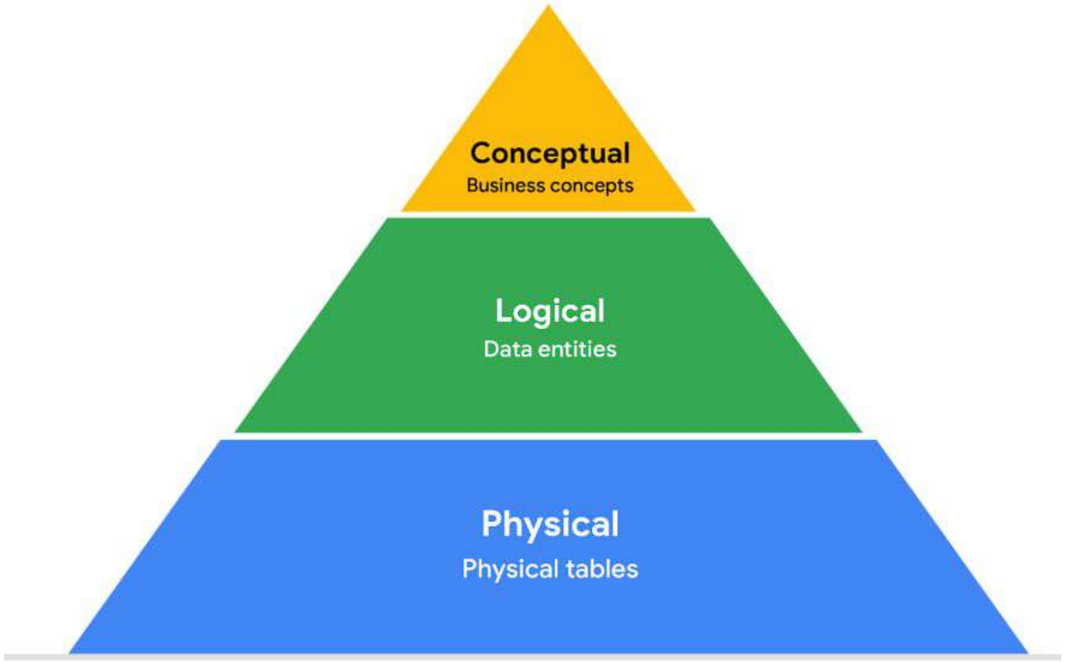
- Cái nhìn tổng quát về cấu trúc dữ liệu

Mô hình hóa dữ liệu logic

- Các chi tiết kỹ thuật về CSDL

Mô hình hóa dữ liệu vật lý

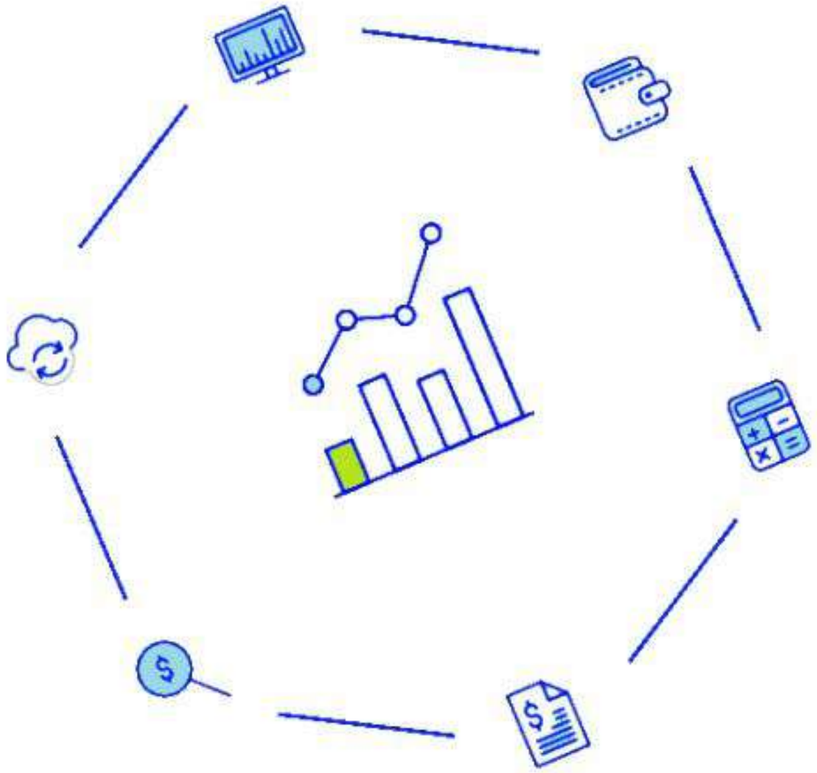
- Cách CSDL hoạt động



https://www.coursera.org/learn/data-preparation/supplement/vtp7L/data-modeling-levels-and-techniques

Nội dung

- Thu thập dữ liệu
- Sự khác biệt giữa các định dạng và cấu trúc dữ liệu
- **Khám phá kiểu dữ liệu, trường dữ liệu và giá trị dữ liệu**



Kiểu dữ liệu

Kiểu dữ liệu: một loại thuộc tính của dữ liệu cho biết loại giá trị của dữ liệu

Kiểu dữ liệu trong spreadsheet

- Kiểu số
- Kiểu chuỗi ký tự
- Kiểu Boolean (TRUE và FALSE)

Bảng dữ liệu:

- Dòng và cột, hay
- Bản ghi và trường

Tên	Tuổi	> 20?
Nam	18	FALSE
Anh	19	FALSE
Nguyên	25	TRUE
Mạnh	17	FALSE
Thắng	24	TRUE
Minh	15	FALSE
Hồng	23	TRUE

Boolean logic

Toán tử AND

- IF (Màu=”Xám”) AND (Màu=”Hồng”) thì mua

Toán tử OR

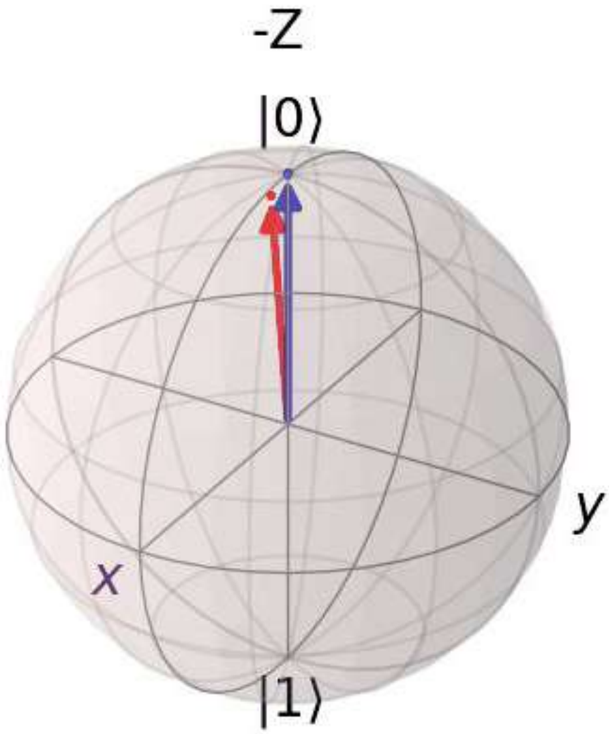
- IF (Màu=”Xám”) OR (Màu=”Hồng”) thì mua

Toán tử NOT

- IF (Màu=”Xám”) AND (Màu=NOT”Hồng”) thì mua

Kết hợp nhiều toán tử

- IF ((Màu=”Xám”) OR (Màu=”Hồng”)) AND (Chống nước = TRUE) thì mua



Dữ liệu dài và rộng

Dữ liệu rộng: một hàng với nhiều cột.

A	B	C	D	E	F	G	H	I
Series Name	**Series Code**	**Country Name**	**Country Code**	**2010 [YR2010]**	**2011 [YR2011]**	**2012 [YR2012]**	**2013 [YR2013]**	**2014 [YR**
Population, total	SP.POP.TOTL	Antigua and Bar	ATG	88028	89253	90409	91516	
Population, total	SP.POP.TOTL	Argentina	ARG	40788453	41261490	41733271	42202935	426
Population, total	SP.POP.TOTL	Aruba	ABW	101669	102046	102560	103159	1

Dữ liệu dài và rộng

Dữ liệu dài: nhiều dòng, ít cột

1	Country Name	Country	Series Name	Year	Population
2	Antigua and Bar	ATG	Population, total	2010	88028
3	Antigua and Bar	ATG	Population, total	2011	89253
4	Antigua and Bar	ATG	Population, total	2012	90409
5	Antigua and Bar	ATG	Population, total	2013	91516
6	Antigua and Bar	ATG	Population, total	2014	92562
7	Antigua and Bar	ATG	Population, total	2015	93566
8	Antigua and Bar	ATG	Population, total	2016	94527
9	Antigua and Bar	ATG	Population, total	2017	95426
10	Antigua and Bar	ATG	Population, total	2018	96286
11	Antigua and Bar	ATG	Population, total	2019	97118
12	Argentina	ARG	Population, total	2010	40788453

Biến đổi dữ liệu

Biến đổi dữ liệu: quá trình thay đổi định dạng, cấu trúc hoặc giá trị của dữ liệu.

Các hình thức biến đổi dữ liệu:

- **Thêm, sao chép** hoặc nhân bản dữ liệu
- **Xóa** trường hoặc bản ghi
- **Chuẩn hóa tên** của các biến
- **Đổi tên**, di chuyển hoặc kết hợp các cột trong cơ sở dữ liệu
- **Kết hợp** một bộ dữ liệu với một bộ dữ liệu khác
- **Lưu file** ở định dạng khác. Ví dụ: chuyển từ file dạng excel sang csv

Tại sao cần biến đổi dữ liệu

- Tổ chức dữ liệu
- Khả năng tương thích dữ liệu
- Di chuyển dữ liệu
- Hợp nhất dữ liệu
- Tăng cường dữ liệu
- So sánh dữ liệu

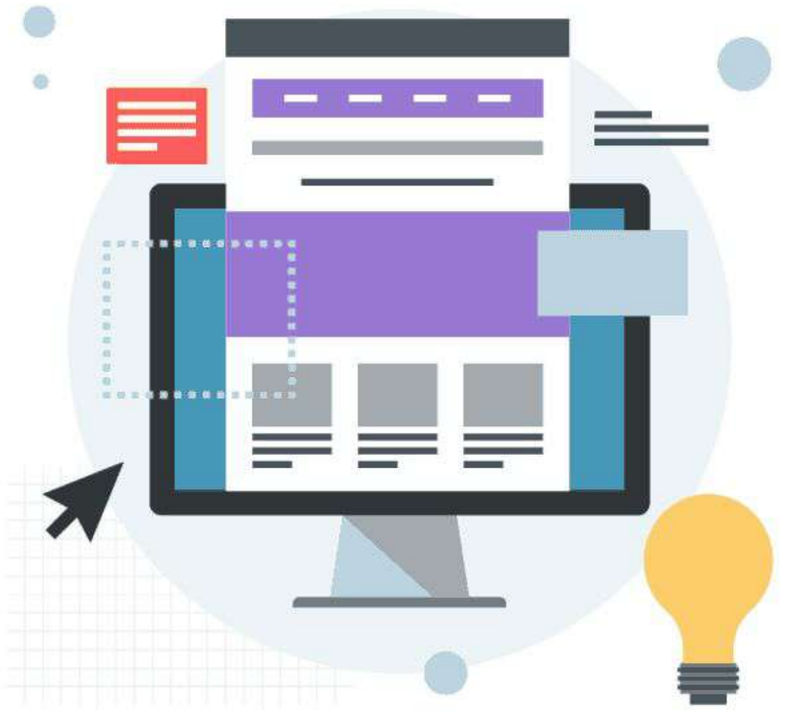
2 ĐÁNH GIÁ CHẤT LƯỢNG DỮ LIỆU

Nội dung

- Dữ liệu khách quan và không thiên kiến
- Khám phá độ tin cậy của dữ liệu
- Đạo đức dữ liệu và quyền riêng tư dữ liệu
- Hiểu về dữ liệu mở

Nội dung

- **Dữ liệu khách quan và không thiên kiến**
- Khám phá độ tin cậy của dữ liệu
- Đạo đức dữ liệu và quyền riêng tư dữ liệu
- Hiểu về dữ liệu mở



Bảo đảm tính toàn vẹn của dữ liệu

Phân tích dữ liệu để tìm ra sự thiên kiến (bias) và độ tin cậy (credibility)

Dữ liệu tốt và dữ liệu xấu

Đạo đức dữ liệu, tính riêng tư và quyền truy cập

- Ai sở hữu tất cả dữ liệu này?
- Chúng ta có bao nhiêu quyền kiểm soát đối với quyền riêng tư của dữ liệu?
- Chúng ta có thể sử dụng và tái sử dụng dữ liệu theo cách chúng ta muốn không?

Thiên kiến

Thiên kiến dữ liệu (bias data) là một loại lỗi có hệ thống làm sai lệch kết quả theo một hướng nhất định.

Ví dụ: phân tích dữ liệu bệnh nhân để xác định độ nguy hiểm của Covid

- **Thiên kiến**: chỉ xét các bệnh nhân trên 70 tuổi
- **Không thiên kiến**: xét bệnh nhân ở nhiều độ tuổi khác nhau

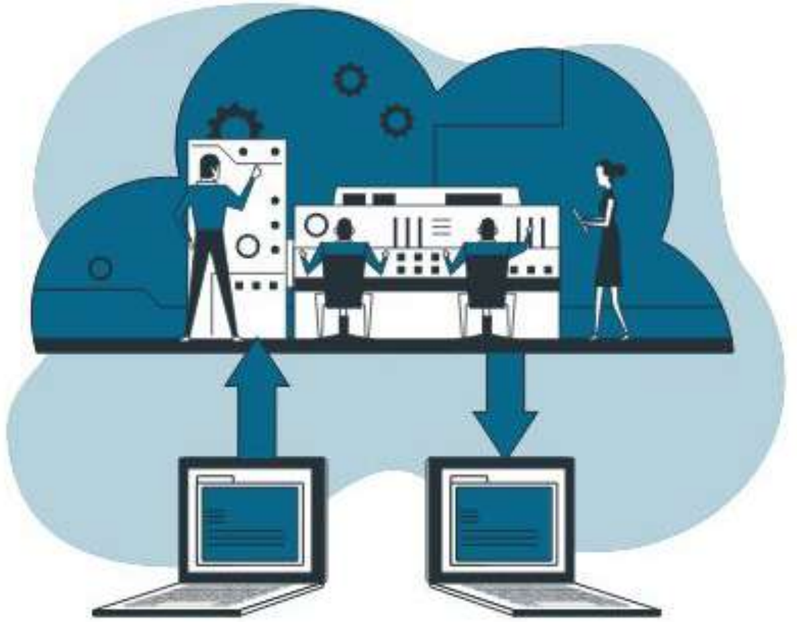
Thiên kiến khi lấy mẫu

Thiên kiến khi lấy mẫu (sampling bias): một mẫu không đại diện cho tổng thể.

- Cách giải quyết: Lấy mẫu ngẫu nhiên

Không thiên kiến khi lấy mẫu (unsampling bias): một mẫu đại diện cho tổng thể.

Trực quan hóa dữ liệu để phát hiện thiên kiến



Các loại thiên kiến khác mẫu

Thiên kiến quan sát (Observer bias)

- Còn gọi là thiên kiến quan sát (experimenter bias) hay thiên kiến nghiên cứu (research bias)
- Những người khác nhau quan sát sự kiện khác nhau

Thiên kiến lý giải (Interpretation bias)

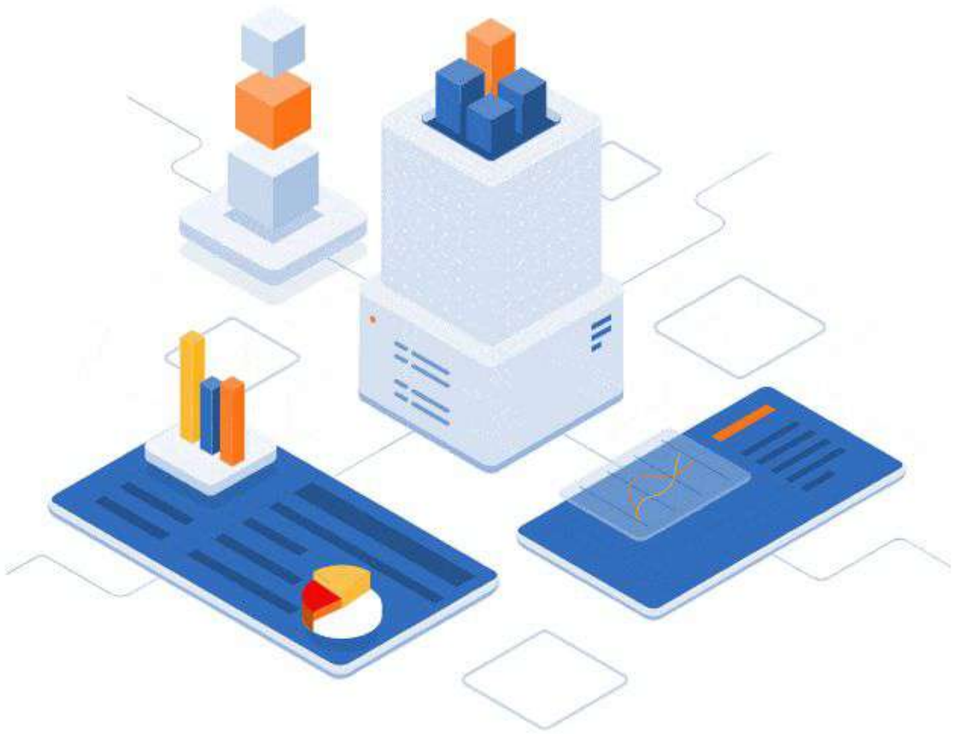
- Xu hướng lý giải một tình huống nhập nhằng theo cách tích cực hay tiêu cực

Thiên kiến xác nhận (Confirmation bias)

- Mọi người thấy những gì họ muốn thấy.

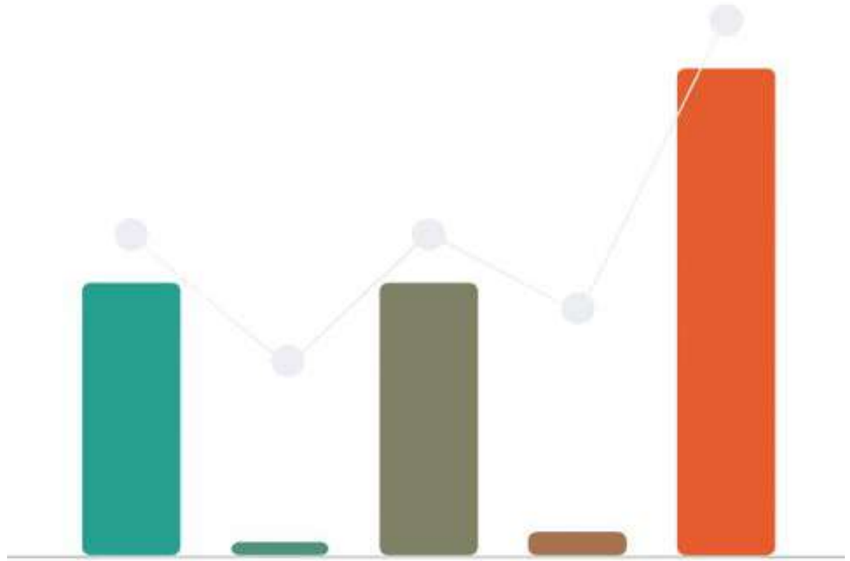
Nội dung

- Dữ liệu khách quan và không thiên kiến
- **Khám phá độ tin cậy của dữ liệu**
- Đạo đức dữ liệu và quyền riêng tư dữ liệu
- Hiểu về dữ liệu mở



Thế nào là nguồn dữ liệu tốt

- Tin cậy (Reliable)
- Nguồn gốc (Original)
- Toàn diện (Comprehensive)
- Tính mới (Current)
- Trích dẫn (Cited)



Thế nào là nguồn dữ liệu tốt

- **Tin cậy (Reliable):** nguồn dữ liệu đáng tin cậy
- Nguồn gốc (Original)
- Toàn diện (Comprehensive)
- Tính mới (Current)
- Trích dẫn (Cited)

Thế nào là nguồn dữ liệu tốt

- Tin cậy (Reliable)
- **Nguồn gốc (Original)**: xác nhận nguồn gốc của dữ liệu
- Toàn diện (Comprehensive)
- Tính mới (Current)
- Trích dẫn (Cited)

Thế nào là nguồn dữ liệu tốt

- Tin cậy (Reliable)
- Nguồn gốc (Original)
- **Toàn diện (Comprehensive)**: chứa đầy đủ các thông tin cần thiết
- Tính mới (Current)
- Trích dẫn (Cited)

Độ tin cậy của dữ liệu

- Tin cậy (Reliable)
- Nguồn gốc (Original)
- Toàn diện (Comprehensive)
- **Tính mới (Current)**: dữ liệu phải mới
- Trích dẫn (Cited)

Độ tin cậy của dữ liệu

- Tin cậy (Reliable)
- Nguồn gốc (Original)
- Toàn diện (Comprehensive)
- Tính mới (Current)
- **Trích dẫn (Cited):** Trích dẫn nguồn dữ liệu sẽ khiến thông tin của bạn đáng tin cậy hơn

Nguồn dữ liệu tốt

- Các bộ dữ liệu công khai đã được kiểm chứng
- Bài báo học thuật
- Dữ liệu tài chính
- Dữ liệu của cơ quan chính phủ

Thế nào là dữ liệu xấu

- Không Tin cậy (Not Reliable)
- Không Nguồn gốc (Not Original)
- Không Toàn diện (Not Comprehensive)
- Không Mới (Not Current)
- Không Trích dẫn (Not Cited)

Nội dung

- Dữ liệu khách quan và không thiên kiến
- Khám phá độ tin cậy của dữ liệu
- **Đạo đức dữ liệu và quyền riêng tư dữ liệu**
- Hiểu về dữ liệu mở

Đạo đức dữ liệu

Đạo đức

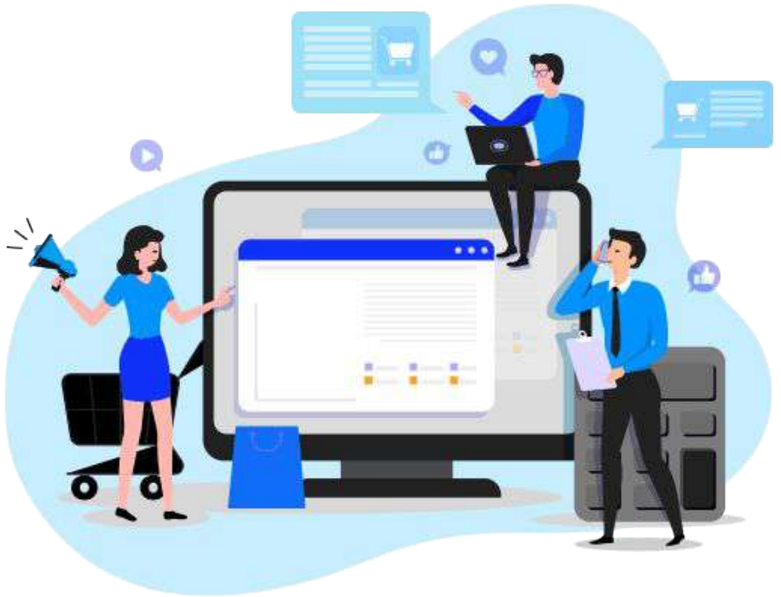
- Các tiêu chuẩn có cơ sở về đúng và sai quy định những gì con người phải làm, thường là về quyền, nghĩa vụ, lợi ích đối với xã hội, sự công bằng hoặc các đức tính cụ thể.

Đạo đức dữ liệu

- Các tiêu chuẩn có cơ sở rõ ràng về đúng và sai quy định cách dữ liệu được thu thập, chia sẻ và sử dụng

Các khía cạnh của đạo đức dữ liệu

- Quyền sở hữu (Ownership)
- Giao dịch minh bạch (Transaction transparency)
- Sự đồng thuận (Consent)
- Tiền tệ (Currency)
- Sự riêng tư (Privacy)
- Tính mở (Openness)



Các khía cạnh của đạo đức dữ liệu

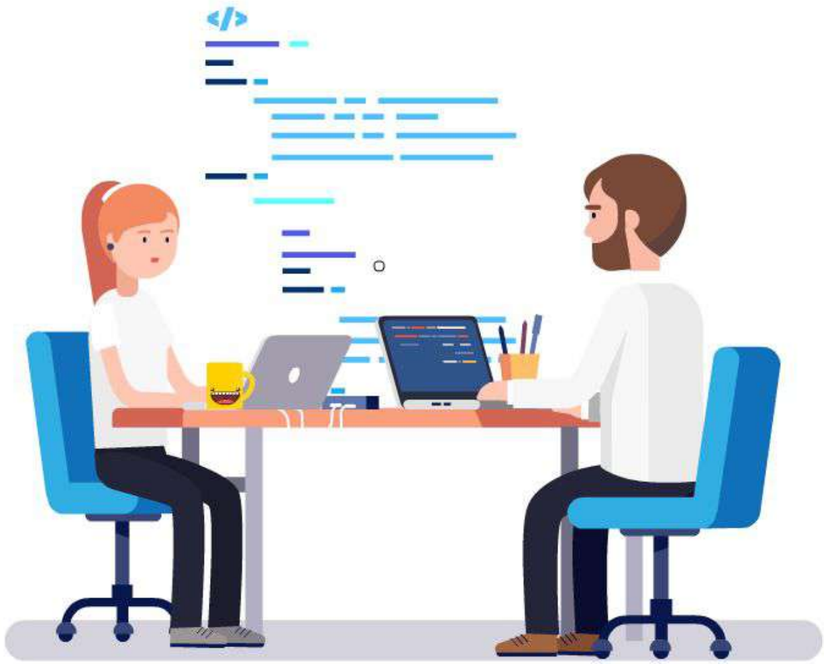
- **Quyền sở hữu (Ownership): ai sở hữu dữ liệu**
- Giao dịch minh bạch (Transaction transparency)
- Sự đồng thuận (Consent)
- Tiền tệ (Currency)
- Sự riêng tư (Privacy)
- Tính mở (Openness)

Các khía cạnh của đạo đức dữ liệu

- Quyền sở hữu (Ownership)
- **Giao dịch minh bạch (Transaction transparency):** phải cung cấp thông tin giúp người cung cấp dữ liệu hiểu được tất cả các hoạt động và thuật toán xử lý dữ liệu trên tập dữ liệu này
- Sự đồng thuận (Consent)
- Tiền tệ (Currency)
- Sự riêng tư (Privacy)
- Tính mở (Openness)

Các khía cạnh của đạo đức dữ liệu

- Quyền sở hữu (Ownership)
- Giao dịch minh bạch (Transaction transparency)
- **Sự đồng thuận (Consent)**: cá nhân có quyền được biết cách thức và lý do sử dụng dữ liệu cá nhân của họ trước khi đồng ý cung cấp
- Tiền tệ (Currency)
- Sự riêng tư (Privacy)
- Tính mở (Openness)



Các khía cạnh của đạo đức dữ liệu

- Quyền sở hữu (Ownership)
- Giao dịch minh bạch (Transaction transparency)
- Sự đồng thuận (Consent)
- **Tiền tệ (Currency)**: cá nhân nên biết về các giao dịch tài chính sinh ra do sử dụng dữ liệu cá nhân họ và quy mô của các giao dịch này
- Sự riêng tư (Privacy)
- Tính mở (Openness)

Các khía cạnh của đạo đức dữ liệu

- Quyền sở hữu (Ownership)
- Giao dịch minh bạch (Transaction transparency)
- Sự đồng thuận (Consent)
- Tiền tệ (Currency)
- **Sự riêng tư (Privacy)**
- **Tính mở (Openness)**

Quyền riêng tư dữ liệu

Quyền riêng tư dữ liệu: bảo toàn thông tin và hoạt động của chủ thể dữ liệu bất kỳ khi nào xảy ra giao dịch dữ liệu.

- Bảo vệ khỏi sự truy cập trái phép đến dữ liệu cá nhân của mình
- Tránh bị sử dụng dữ liệu của mình một cách không phù hợp
- Quyền kiểm tra, cập nhật hoặc chỉnh sửa dữ liệu của mình
- Khả năng đồng ý sử dụng dữ liệu của mình
- Quyền hợp pháp để truy cập vào dữ liệu của mình.

Ẩn danh dữ liệu

Ẩn danh dữ liệu: quá trình bảo vệ dữ liệu riêng tư hoặc nhạy cảm của mọi người bằng cách loại bỏ loại thông tin đó

Cách thực hiện:

- Làm trống
- Băm
- Che thông tin cá nhân

Bằng cách sử dụng mã có độ dài cố định để đại diện

Loại dữ liệu nào nên ẩn danh

Dữ liệu chăm sóc sức khỏe và tài chính.

Khử nhận dạng: quy trình xóa sạch những dữ liệu có thể được dùng để nhận ra cá nhân.

Những thông tin khác thường được ẩn danh:

- Số điện thoại, Tên
- Biển số xe và số xe
- Số CMND/CCCD
- Các địa chỉ IP
- Hồ sơ bệnh án
- Địa chỉ email
- Ảnh chụp
- Số tài khoản

Nội dung

- Dữ liệu khách quan và không thiên kiến
- Khám phá độ tin cậy của dữ liệu
- Đạo đức dữ liệu và quyền riêng tư dữ liệu
- **Hiểu về dữ liệu mở**

Đặc trưng của dữ liệu mở

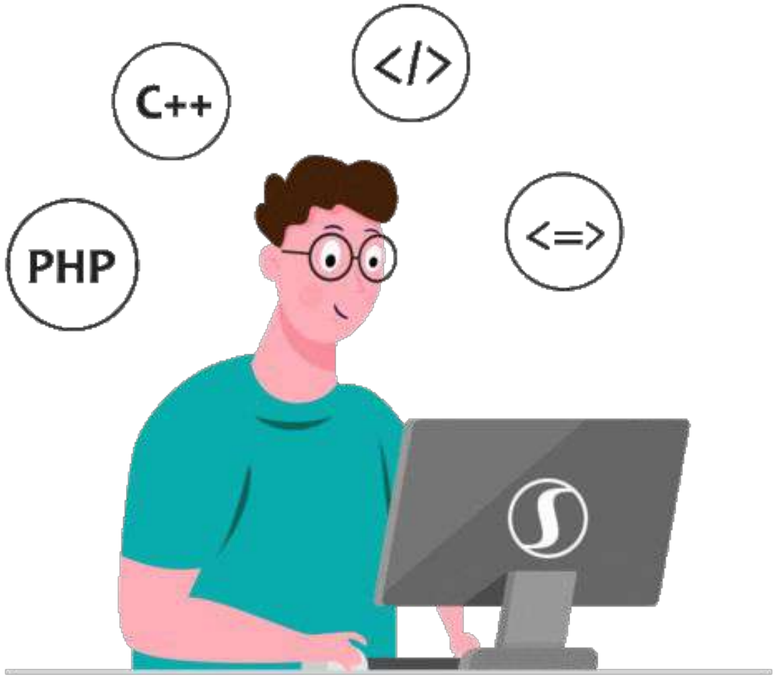
Dữ liệu mở: quyền truy cập, sử dụng và chia sẻ dữ liệu miễn phí

Tiêu chuẩn của dữ liệu mở:

- Có sẵn và được công chúng truy cập dưới dạng một tập dữ liệu hoàn chỉnh
- Được cung cấp theo các điều khoản cho phép nó được tái sử dụng và phân phối lại
- Cho phép mọi người tham gia. Ai cũng có thể sử dụng, tái sử dụng và phân phối lại dữ liệu

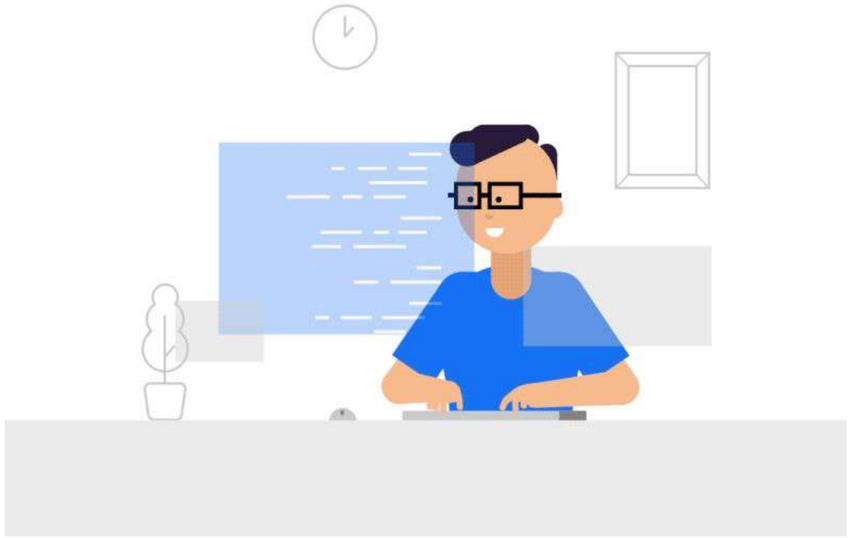
Ưu điểm của dữ liệu mở:

- Các bộ dữ liệu đáng tin cậy được sử dụng rộng rãi hơn



Đặc trưng của dữ liệu mở

- **Khả năng tương tác (Interoperability)**: khả năng của các hệ thống dữ liệu và dịch vụ kết nối và chia sẻ dữ liệu với nhau một cách công khai.
- Ví dụ: bệnh viện, phòng khám, nhà thuốc và phòng thí nghiệm cần truy cập và chia sẻ dữ liệu về bệnh nhân



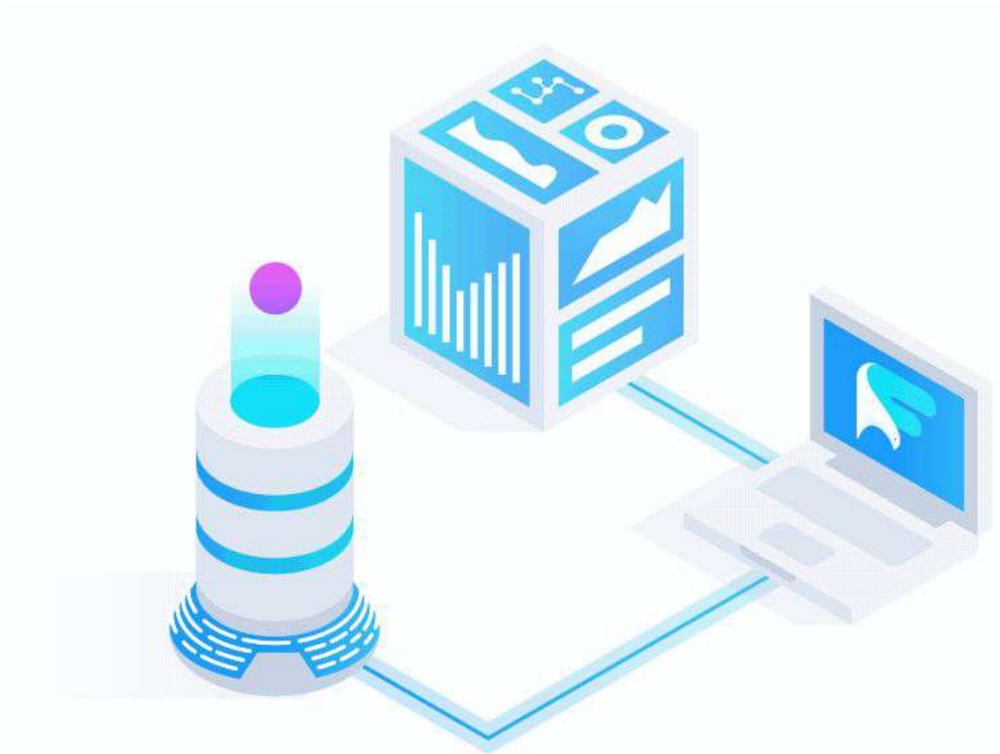
3 LÀM VIỆC VỚI CƠ SỞ DỮ LIỆU

Nội dung

- Giới thiệu về cơ sở dữ liệu
- Siêu dữ liệu
- Truy cập các nguồn dữ liệu khác nhau
- Sắp xếp và tìm kiếm
- Làm việc với tập dữ liệu lớn bằng SQL

Nội dung

- **Giới thiệu về cơ sở dữ liệu**
- Siêu dữ liệu
- Truy cập các nguồn dữ liệu khác nhau
- Sắp xếp và lọc
- Làm việc với tập dữ liệu lớn bằng SQL



Cơ sở dữ liệu

Cơ sở dữ liệu (CSDL): tập hợp dữ liệu được lưu trữ trong một hệ thống máy tính

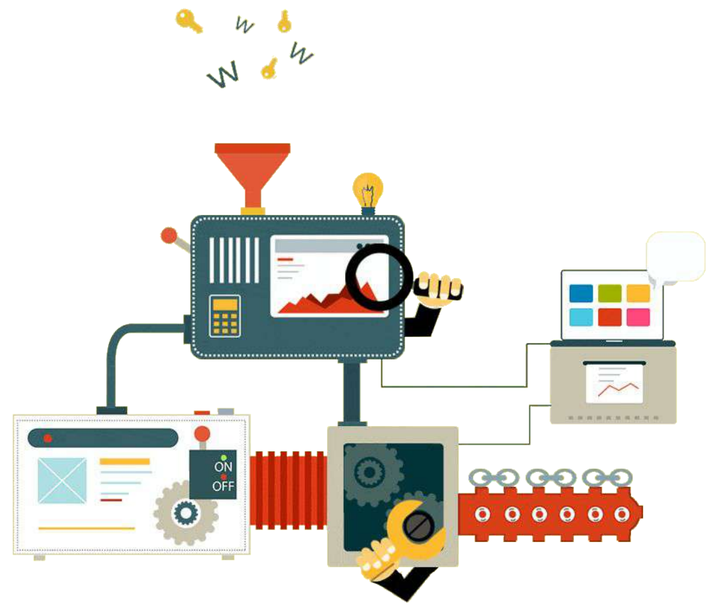
Siêu dữ liệu (metadata): dữ liệu về dữ liệu

- Dữ liệu đến từ đâu, khi nào, được tạo ra như thế nào, nội dung là gì.

CSDL lưu trữ và tổ chức dữ liệu

- Giúp các nhà phân tích dữ liệu quản lý và truy cập thông tin dễ dàng hơn.

CSDL quan hệ (relational database): CSDL có chứa các bảng liên quan có thể được kết nối thông qua các mối quan hệ của chúng



Cơ sở dữ liệu

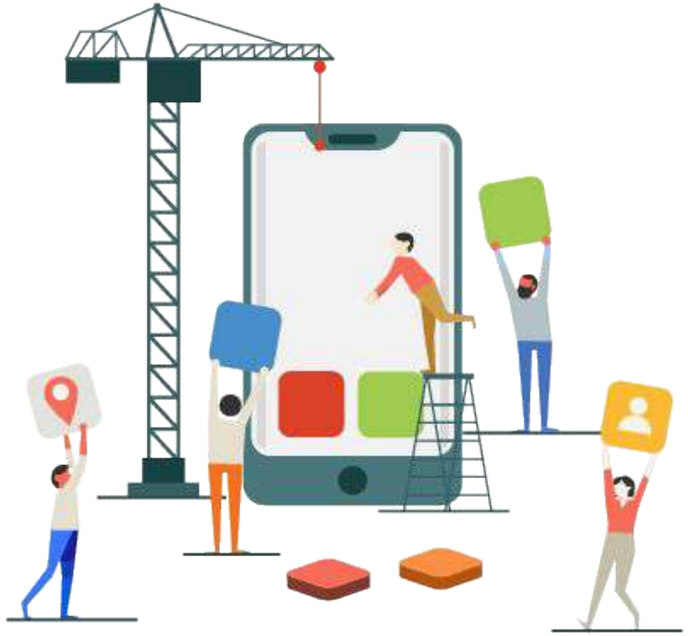
Bảng: “Car Dealerships”, “Product details”, “Repair parts”

Khóa chính: một mã định danh tham chiếu đến một cột trong đó mỗi giá trị là duy nhất. VD: *BranchID*, *VIN*, *Part ID*

- **Khóa tổ hợp:** một loại khóa chính chứa nhiều cột
- **Khóa ngoại**: một trường trong bảng chứa khóa chính của một bảng khác. VD: *Branch ID* trong Product details

SQL

- CSDL sử dụng ngôn ngữ truy vấn (query language) để giao tiếp
- **Ngôn ngữ truy vấn có cấu trúc (SQL)** là một loại ngôn ngữ truy vấn cho phép các nhà phân tích dữ liệu giao tiếp với CSDL.
- Các nhà phân tích dữ liệu có thể viết các truy vấn để lấy dữ liệu từ các bảng liên quan.



Nội dung

- Giới thiệu về cơ sở dữ liệu
- **Siêu dữ liệu**
- Truy cập các nguồn dữ liệu khác nhau
- Sắp xếp và lọc
- Làm việc với tập dữ liệu lớn bằng SQL

Siêu dữ liệu

- **Siêu dữ liệu** là dữ liệu về dữ liệu
- Trong phân tích dữ liệu: siêu dữ liệu giúp các nhà phân tích dữ liệu diễn giải nội dung của dữ liệu

Các loại siêu dữ liệu

Mô tả (descriptive)

- Mô tả và xác định dữ liệu
- ISBN, Tên tác giả của sách

Cấu trúc (structural)

- Cấu trúc của dữ liệu
- Mục lục

Quản trị (administrative)

- Thông tin kỹ thuật
- Loại file

Thông tin siêu dữ liệu cung cấp

- Tiêu đề và mô tả
- Thẻ và danh mục
- Ai đã tạo ra nó và khi nào
- Ai đã sửa đổi nó lần cuối và khi nào
- Ai có thể truy cập hoặc cập nhật nó

Ví dụ về siêu dữ liệu

- **Hình ảnh**: tên tập tin, vị trí địa lý
- **Email**: tên chủ đề, người gửi, người nhận
- **Bảng tính và tài liệu**: tiêu đề, tác giả, ngày tạo, số trang
- **Trang web**: tên người tạo trang web, tiêu đề và mô tả trang web, thời gian tạo
- **File**: tên file, kích thước file, ngày tạo và sửa, loại tập tin
- **Sách**: tên sách, tên tác giả, mục lục, thông tin nhà xuất bản, mô tả bản quyền

Siêu dữ liệu trong phân tích dữ liệu

Siêu dữ liệu: giúp dữ liệu nhất quán và thống nhất

Siêu dữ liệu làm cho dữ liệu đáng tin cậy hơn bằng cách đảm bảo dữ liệu chính xác, phù hợp và kịp thời

Kho lưu trữ siêu dữ liệu: CSDL được tạo riêng để lưu trữ siêu dữ liệu

- Mô tả trạng thái và vị trí của siêu dữ liệu
- Cấu trúc của các bảng bên trong
- Cách dữ liệu di chuyển trong kho lưu trữ
- Theo dõi ai truy cập siêu dữ liệu và khi nào

Quản trị siêu dữ liệu

Siêu dữ liệu được lưu trữ ở một vị trí trung tâm, duy nhất và nó cung cấp cho công ty thông tin được chuẩn hóa về tất cả dữ liệu của mình

- Thông tin về vị trí của mỗi hệ thống
- Cách kết nối giữa các hệ thống khác nhau

Quản trị dữ liệu (data governance): một quy trình nhằm đảm bảo việc quản lý chính thức các tài sản dữ liệu của một công ty

Nội dung

- Giới thiệu về cơ sở dữ liệu
- Siêu dữ liệu
- **Truy cập các nguồn dữ liệu khác nhau**
- Sắp xếp và lọc
- Làm việc với tập dữ liệu lớn bằng SQL

Làm việc với nhiều nguồn dữ liệu

Dữ liệu nội bộ hay dữ liệu chính: dữ liệu nằm trong hệ thống của chính công ty, được tạo ra từ bên trong công ty.

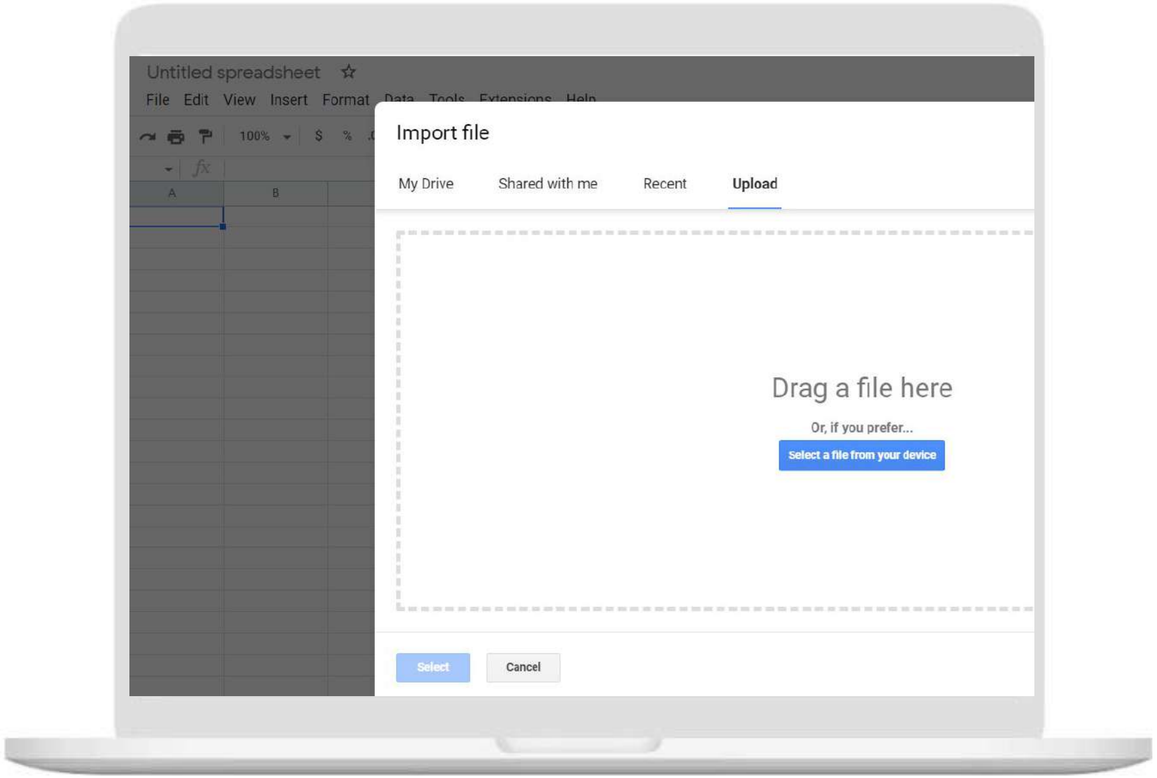
- Thu thập dữ liệu phức tạp
- Miễn phí

Dữ liệu bên ngoài hay dữ liệu thứ cấp: dữ liệu tồn tại và được tạo ra bên ngoài một tổ chức.

Các nguồn dữ liệu ngoài

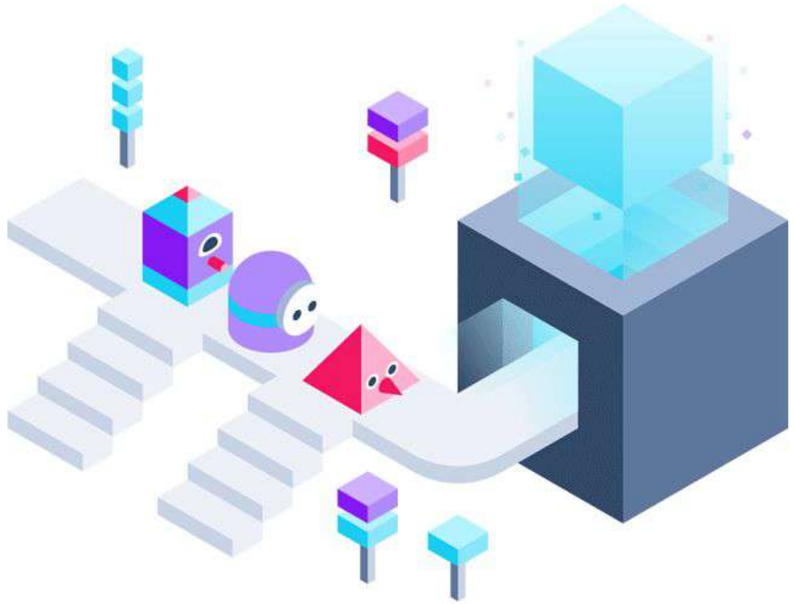
File CSV (Comma-separated values)

- CSV lưu dữ liệu ở định dạng bảng
- File -> Import -> Chọn file CSV



Nội dung

- Giới thiệu về cơ sở dữ liệu
- Siêu dữ liệu
- Truy cập các nguồn dữ liệu khác nhau
- **Sắp xếp và lọc**
- Làm việc với tập dữ liệu lớn bằng SQL



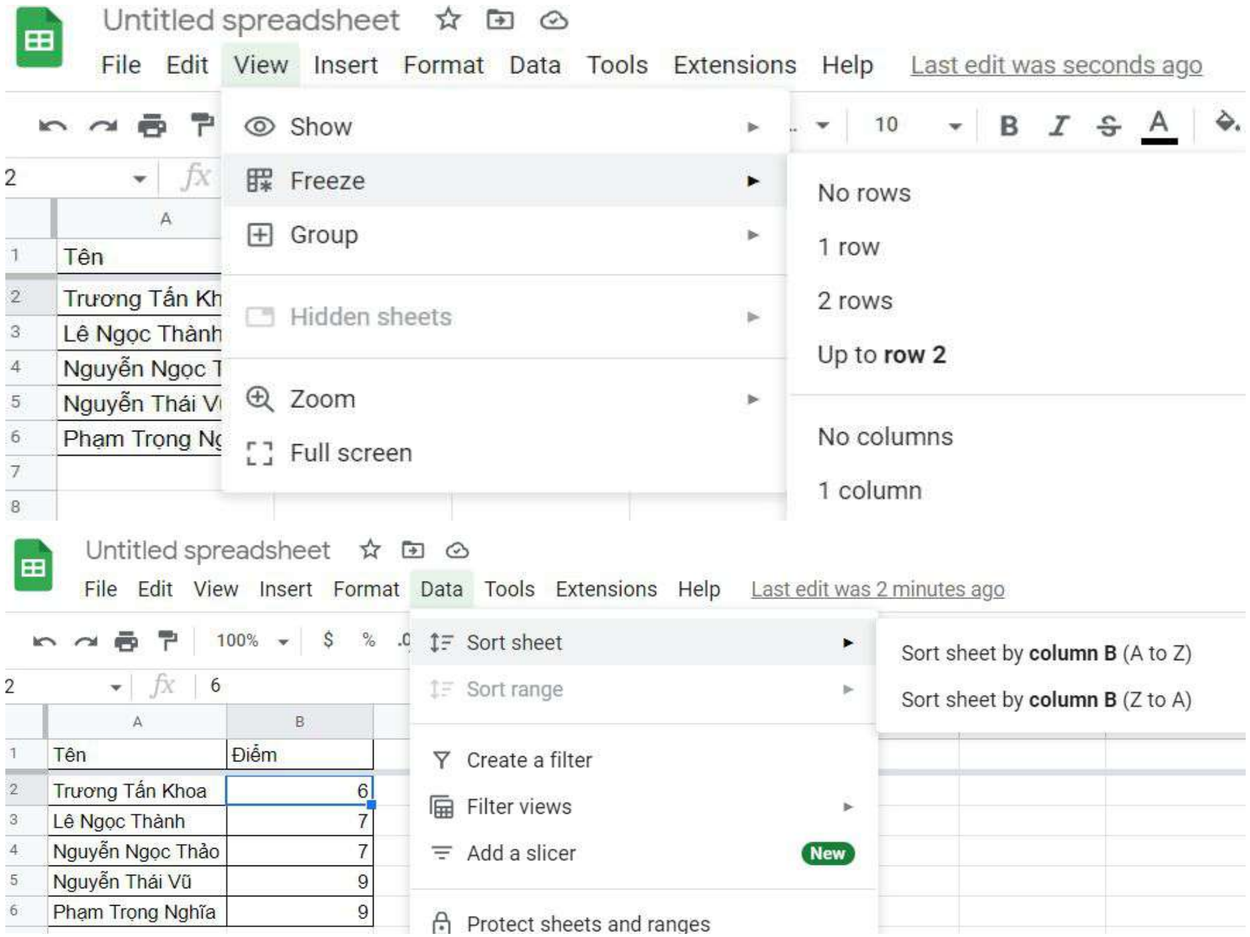
Sắp xếp

Sắp xếp là việc biến đổi dữ liệu thành một thứ tự có ý nghĩa

- Sắp tăng dần, giảm dần
- Theo thứ tự từ điển hay thứ tự số

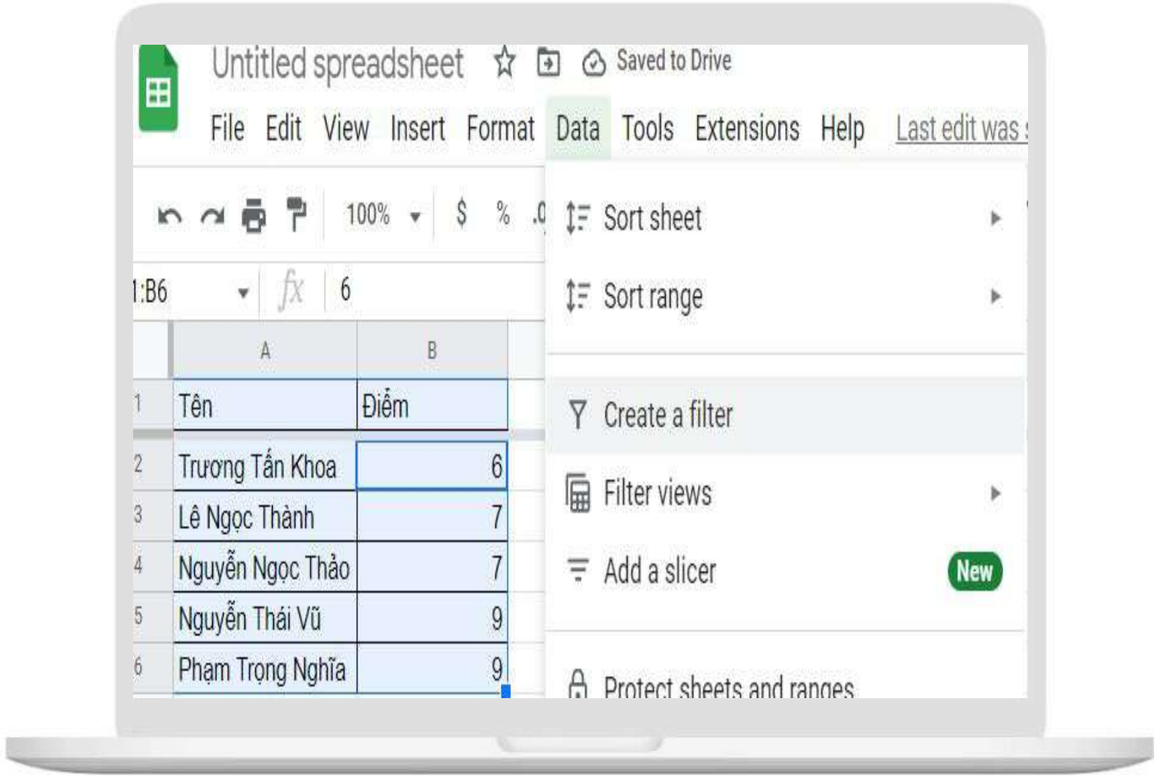
Đóng băng (Freeze) header

Sort trong google sheet



Lọc

- **Lọc (filtering)**: chỉ hiển thị dữ liệu đáp ứng một tiêu chí cụ thể trong khi ẩn phần còn lại.
- Ví dụ: chỉ xem dữ liệu về nhân viên thuộc phòng kế toán

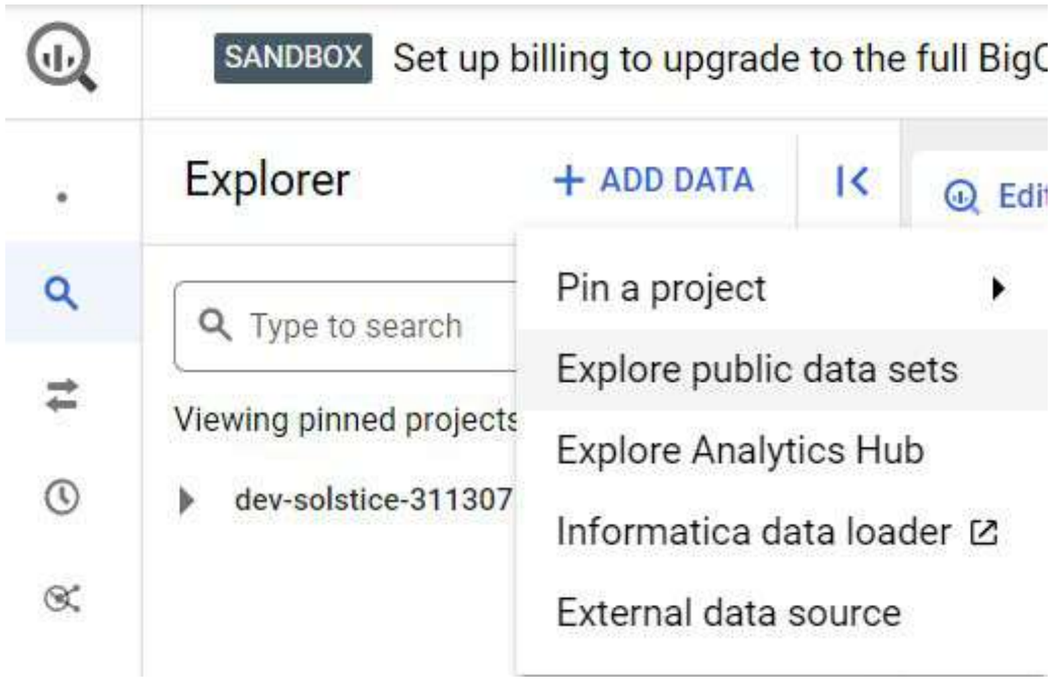


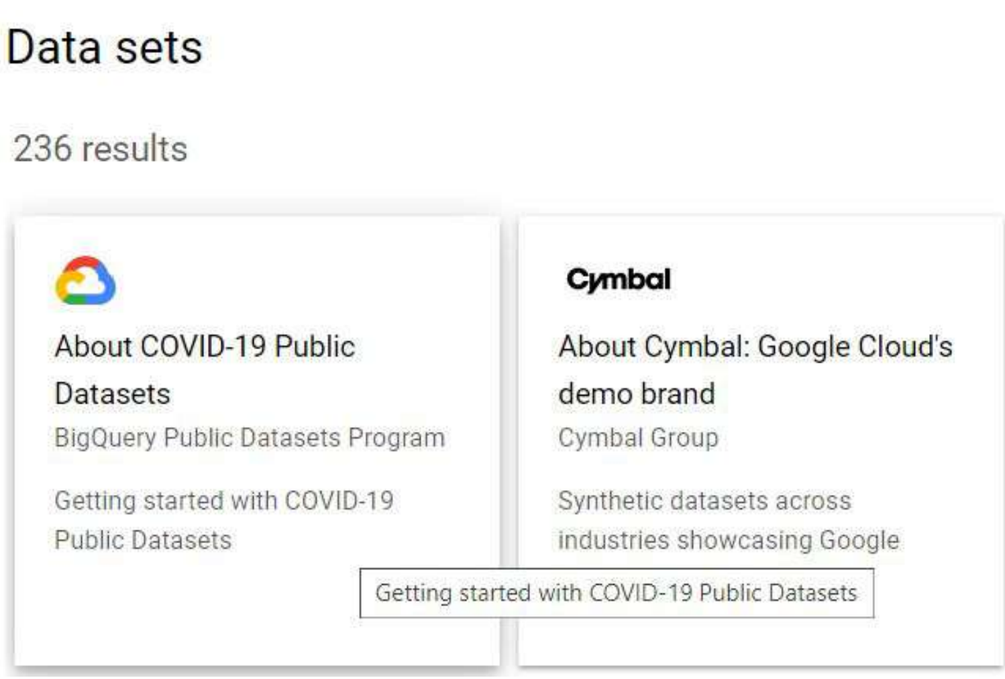
Nội dung

- Giới thiệu về cơ sở dữ liệu
- Siêu dữ liệu
- Truy cập các nguồn dữ liệu khác nhau
- Sắp xếp và lọc
- **Làm việc với tập dữ liệu lớn bằng SQL**

Sử dụng BigQuery

Thêm bộ dữ liệu Public





Sử dụng BigQuery

Truy vấn

```
SELECT * FROM `bigquery-public-data.noaa_lightning_2019` LIMIT 1000
```

Lưu ý: có thể dùng dấu nháy đơn hoặc không cho tập dữ liệu

```
SELECT * FROM bigquery-public-data.noaa_lightning_2019 LIMIT 1000
```

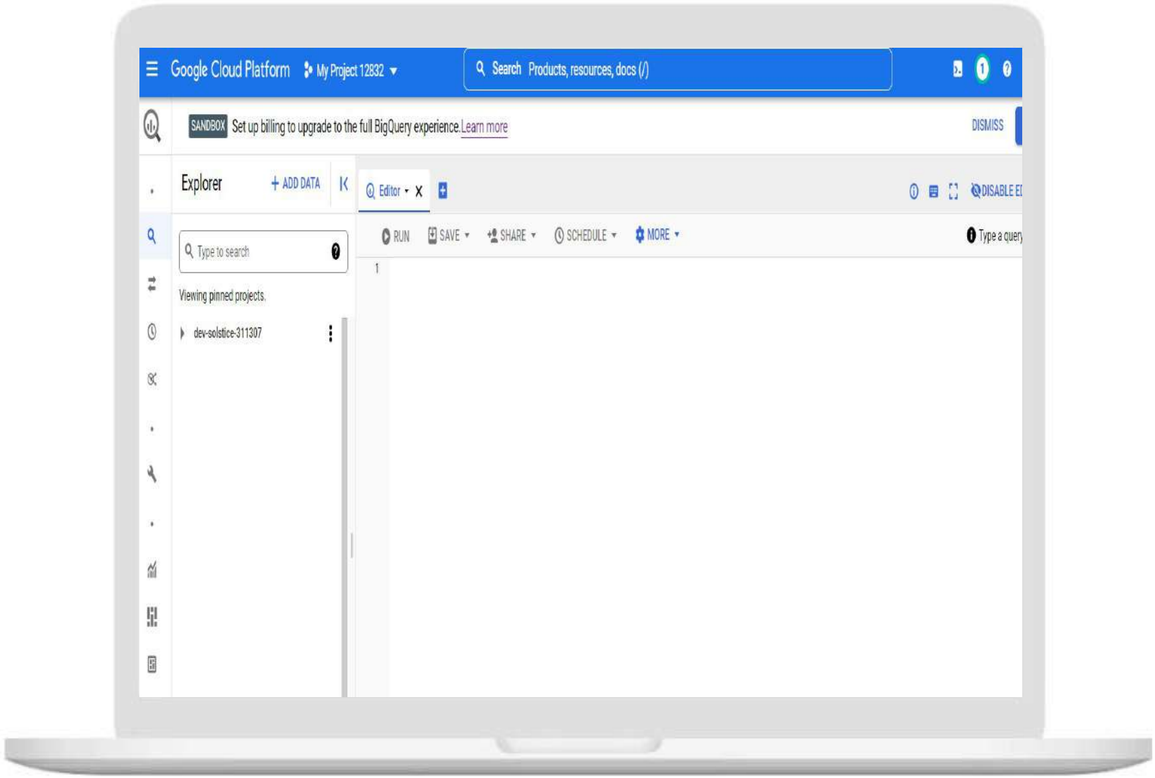
BigQuery

BigQuery: kho dữ liệu thương mại của Google.

- Chỉ cần input dữ liệu vào BigQuery và để Google xử lý phần còn lại.

Hai loại tài khoản

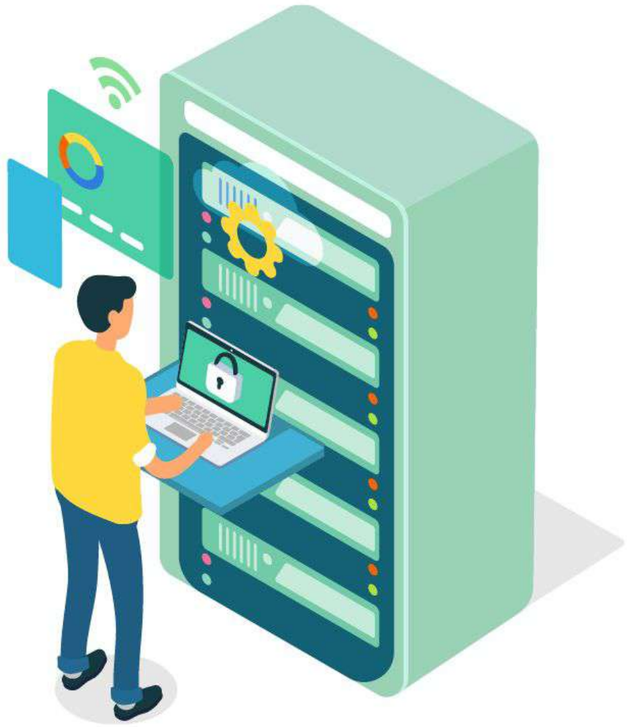
- Sandbox: miễn phí với tài khoản google.
- Free trial: trả tiền với nhiều tính năng hơn.



4 QUẢN LÝ VÀ BẢO MẬT DỮ LIỆU

Nội dung

- Tổ chức dữ liệu hiệu quả
- Bảo mật dữ liệu



Tổ chức dữ liệu hiệu quả

Giữ các file có tổ chức

Quy ước đặt tên file (naming convention)

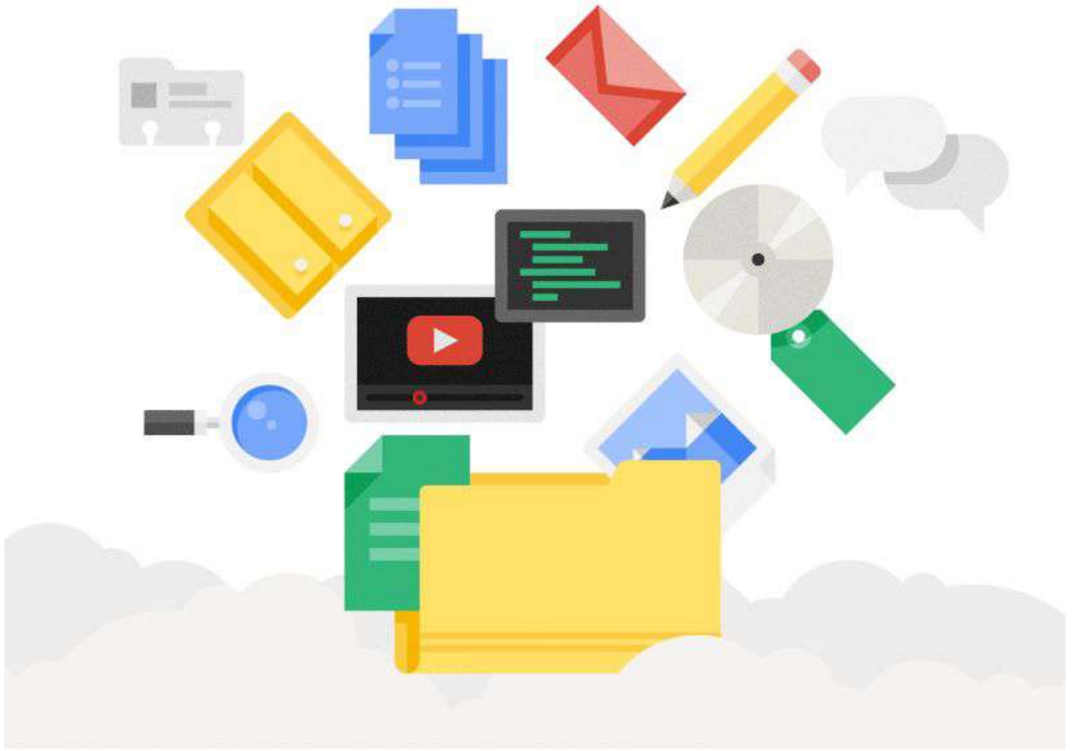
- Đặt tên file theo quy ước chung, dễ gợi nhớ

Sắp xếp thư thư mục (foldering)

- Tổ chức các file vào trong các thư mục
- File liên quan nằm trong cùng thư mục
- Chia thành thư mục con: tổng quát ở ngoài, cụ thể ở trong

Lưu trữ các file cũ (archiving older file)

- Di chuyển các file cũ sang nơi riêng



Tổ chức dữ liệu hiệu quả

Tổ chức dữ liệu

- Sớm thảo luận và thống nhất các quy ước đặt tên file
- Tên file của bạn phải phù hợp với quy ước chung của nhóm và công ty
- Đảm bảo rằng tên file có ý nghĩa
- Bao gồm ngày và số phiên bản trong tên file: YYYYMMDD hay v##

➢ Salereport20220701, Salereport20220701v02

Tổ chức dữ liệu hiệu quả

Quy ước đặt tên file

- Tránh khoảng trắng và ký tự đặc biệt, sử dụng gạch ngang, dấu gạch dưới hoặc chữ in hoa.
  VD: SaleReport_2022_0701_v02
- Tạo siêu dữ liệu hay file mẫu mô tả quy ước đặt tên file

Tên file	SaleReport_2022_07_01_v02
Chủ đề	**SaleReport**_2022_07_01_v02
Ngày tháng	SaleReport_**2022_07_01**_v02
Phiên bản	SaleReport_2022_07_01_**v02**

Bảo mật dữ liệu

Tính năng bảo mật của bảng tính:

- Bảo vệ trang tính
- Kiểm soát truy cập

Bảo mật dữ liệu: bảo vệ dữ liệu khỏi truy cập trái phép hoặc phá hoại bằng cách áp dụng các biện pháp an toàn

5 SỰ HIỆN DIỆN TRỰC TUYẾN

Nội dung

- Tạo ra và duy trì sự hiện diện trực tuyến
- Xây dựng mạng lưới phân tích dữ liệu

Tạo ra và duy trì sự hiện diện trực tuyến

Sự hiện diện của một nhà phân tích dữ liệu

Một nhà phân tích dữ liệu là một phần của **cộng đồng dữ liệu**

- Xây dựng sự hiện diện trực tuyến nhất quán và chuyên nghiệp
- Kết nối với những người khác trong cùng lĩnh vực và mở rộng mạng lưới của mình

Tiếp theo, bạn sẽ học cách bắt đầu xây dựng **sự hiện diện trực tuyến (online presence)**.

Duy trì sự hiện diện trực tuyến có thể mở ra rất nhiều cơ hội mới

Tạo ra và duy trì sự hiện diện trực tuyến

Sự hiện diện trực tuyến sẽ giúp

- Giúp các nhà tuyển dụng tiềm năng tìm thấy bạn
- Kết nối với các nhà phân tích dữ liệu khác
- Tìm hiểu và chia sẻ kết quả
- Tham gia vào các sự kiện cộng đồng

Tạo ra và duy trì sự hiện diện trực tuyến

Linked và Github

Lindked

- Tạo kết nối
- Theo dõi các xu hướng trong ngành
- Tìm kiếm cơ hội việc làm

GitHub

- Chia sẻ ý kiến và tài nguyên
- Forum và wiki
- Quản lý dự án nhóm
- Tổ chức các sự kiện cộng đồng

Xây dựng mạng lưới phân tích dữ liệu

Cố vấn

Người cố vấn (mentor) là một người chuyên nghiệp, họ chia sẻ kiến thức, kỹ năng và kinh nghiệm của họ để giúp bạn phát triển và trưởng thành.

Tìm kiếm cố vấn:

- Trong môi trường làm việc
- Các nền tảng mạng xã hội hoặc tìm kiếm cố vấn: Score.org, MicroMentor.org, Mentorship app

Nhà tài trợ (sponsor) là một người ủng hộ chuyên nghiệp, người cam kết giúp phát triển sự nghiệp của người được tài trợ

- Nhà tài trợ sẽ chọn bạn
- Cố gắng hoàn thành công việc một cách tốt nhất

TỔNG KẾT

Những ý chính cần nắm

- Cách các nhà phân tích dữ liệu xác định loại dữ liệu nào để thu thập cho quá trình phân tích.
- Dữ liệu có cấu trúc, phi cấu trúc, loại dữ liệu và định dạng dữ liệu.
- Các loại thiên kiến (bias) trong dữ liệu để bảo đảm sự tin cậy của dữ liệu .
- Cách các nhà phân tích dữ liệu sử dụng bảng tính và SQL với Cơ sở dữ liệu và tập dữ liệu.
- Dữ liệu mở, mối quan hệ và sự quan trọng giữa đạo đức dữ liệu và tính riêng tư của dữ liệu.
- Cách truy xuất cơ sở dữ liệu, rút trích, lọc, và sắp xếp dữ liệu chứa trong đó.
- Các thông lệ tốt nhất cho việc tổ chức dữ liệu và lưu trữ một cách bảo mật.

THANK YOU