

# Tài liệu đọc

## Phân Tích Dữ Liệu

### Khóa 3: Chuẩn bị dữ liệu cho Khám phá

#### Phần 1: Đề cương bài đọc

#### Bài đọc 1 Tổng quan về chuẩn bị dữ liệu

##### 1.1 Nhu cầu chuẩn bị dữ liệu và giới thiệu về các công cụ sử dụng để chuẩn bị dữ liệu.

- Dữ liệu sẽ được thu thập như thế nào
- Nguồn dữ liệu
- Giải quyết vấn đề kinh doanh của bạn
- Số lượng dữ liệu cần thu thập
- Khung thời gian

##### 1.2 Thu thập dữ liệu

- Dữ liệu chính và dữ liệu thứ cấp
- Dữ liệu nội bộ và dữ liệu bên ngoài
- Dữ liệu liên tục và dữ liệu rời rạc
- Dữ liệu định tính và dữ liệu định lượng
- Dữ liệu có thứ tự và không thứ tự
- Có cấu trúc và phi cấu trúc

##### 1.3 Định dạng và mô hình hoá dữ liệu

- Dữ liệu có cấu trúc
- Dữ liệu phi cấu trúc
- Vấn đề công bằng

##### 1.4 Khám phá các loại dữ liệu

- Mô hình dữ liệu là gì?
- Các cấp độ của mô hình hóa dữ liệu
- Các kỹ thuật của mô hình hóa dữ liệu
- Giới thiệu về mô hình Boolean logic

	1.5 Tài liệu tham khảo
<a href="#">Bài đọc 2</a>	Đánh giá chất lượng của dữ liệu
	2.1 <a href="#">Một số định nghĩa và thuật ngữ</a> 2.2 <a href="#">Dữ liệu mở</a> - Dữ liệu mở là gì? - Dữ liệu nào nên được mở 2.3 <a href="#">Các trang web và tài nguyên cho dữ liệu mở</a> 2.4 <a href="#">Tài liệu tham khảo</a>
<a href="#">Bài đọc 3</a>	Cơ sở dữ liệu (CSDL) cho phân tích dữ liệu
	3.1 <a href="#">Cơ sở dữ liệu trong phân tích dữ liệu</a> - Giới thiệu về CSDL quan hệ - Giới thiệu về SQL - Vai trò của CSDL trong phân tích dữ liệu - Khám phá tập dữ liệu 3.2 <a href="#">Kiểm tra tập dữ liệu</a> - Tình huống - Download dữ liệu - Khám phá dữ liệu - Kết luận 3.3 <a href="#">Tầm quan trọng của siêu dữ liệu</a> - Các thành phần của siêu dữ liệu - Ví dụ về siêu dữ liệu - Ngữ cảnh của dữ liệu 3.4 <a href="#">Nhập dữ liệu từ một nguồn bên ngoài vào bảng tính</a> - Nhập dữ liệu từ các bảng tính khác - Nhập dữ liệu từ file CSV - Nhập các bảng HTML từ trang web 3.5 <a href="#">Khám phá các tập dữ liệu công khai</a> 3.6 <a href="#">BigQuery</a> 3.7 <a href="#">Hướng dẫn sử dụng SQL</a> 3.8 <a href="#">Tài liệu tham khảo</a>
<a href="#">Bài đọc 4</a>	Tổ chức và bảo vệ dữ liệu của bạn
	4.1 <a href="#">Hướng dẫn tổ chức dữ liệu</a> - Hướng dẫn cho quy ước đặt tên file - Một số gợi ý về cách tổ chức file 4.2 <a href="#">Bảo mật dữ liệu</a> 4.3 <a href="#">Tài liệu tham khảo</a>
<a href="#">Bài đọc 5</a>	Tham gia vào cộng đồng dữ liệu

5.1	<a href="#">Tạo hoặc tăng cường sự hiện diện trực tuyến</a>
-	Quản lý sự hiện diện với vai trò người phân tích dữ liệu
-	Cách sử dụng LinkedIn
5.2	<a href="#">Xây dựng mạng phân tích dữ liệu</a>
-	Bí quyết kết nối mạng
5.3	<a href="#">Phát triển mạng lưới</a>
5.4	<a href="#">Tài liệu tham khảo</a>

## Phần 2: Hướng dẫn trả lời câu hỏi - Quiz

# Phần 1

## TÀI LIỆU ĐỌC BỔ TRỢ

## Bài đọc 1: Tổng quan về chuẩn bị dữ liệu

### 1. Nhu cầu chuẩn bị dữ liệu và giới thiệu về các công cụ sử dụng để chuẩn bị dữ liệu

Sau đây là một số cân nhắc cần ghi nhớ khi thu thập dữ liệu cho phân tích của bạn:

#### Dữ liệu sẽ được thu thập như thế nào

Quyết định xem bạn sẽ thu thập dữ liệu bằng cách sử dụng tài nguyên của riêng mình hay nhận (và có thể mua) từ một bên khác. Dữ liệu mà bạn tự thu thập được gọi là dữ liệu của bên thứ nhất.

#### Nguồn dữ liệu

Nếu không thu thập dữ liệu bằng tài nguyên của riêng mình, bạn có thể lấy dữ liệu từ các nhà cung cấp dữ liệu bên thứ hai hoặc bên thứ ba. **Dữ liệu của bên thứ hai** được thu thập trực tiếp bởi một nhóm khác và sau đó được bán. **Dữ liệu của bên thứ ba** được bán bởi một nhà cung cấp không tự thu thập dữ liệu. Dữ liệu của bên thứ ba có thể đến từ một số nguồn khác nhau.

#### Giải quyết vấn đề kinh doanh của bạn

Tập dữ liệu có thể có nhiều thông tin thú vị. Nhưng hãy chắc chắn chọn tập dữ liệu thực giúp giải quyết vấn đề của bạn. Ví dụ: nếu bạn đang phân tích xu hướng theo thời gian, hãy đảm bảo rằng bạn sử dụng dữ liệu chuỗi thời gian - nói cách khác là dữ liệu ngày tháng.

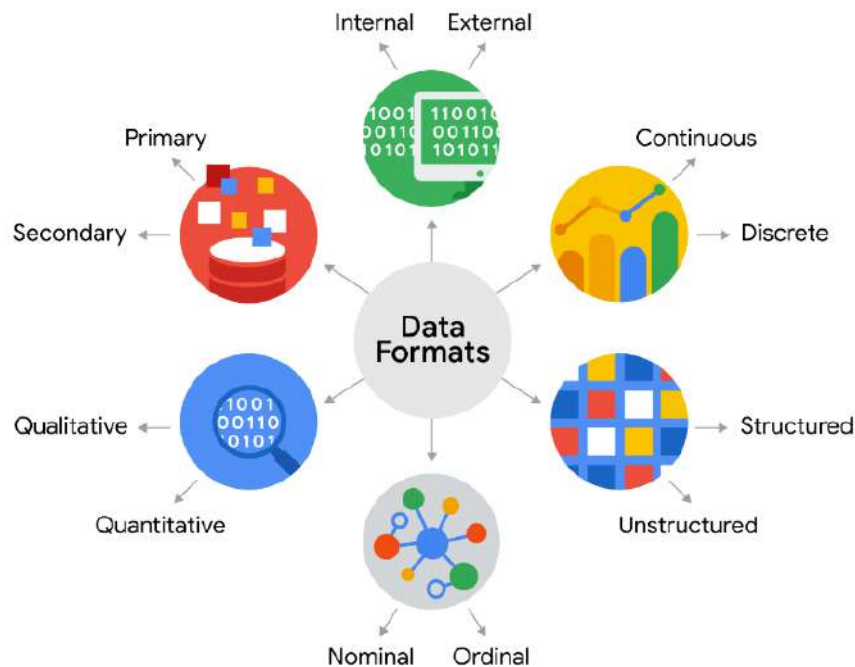
#### Số lượng dữ liệu cần thu thập

Nếu bạn đang thu thập dữ liệu của riêng mình, hãy đưa ra quyết định hợp lý về kích thước mẫu. Một mẫu ngẫu nhiên từ dữ liệu đã có sẵn có thể tốt cho một số dự án. Các dự án khác có thể cần thu thập dữ liệu chiến lược hơn để tập trung vào các tiêu chí nhất định. Mỗi dự án có nhu cầu riêng của nó.

## Khung thời gian

Nếu bạn đang thu thập dữ liệu của riêng mình, hãy quyết định xem bạn sẽ cần thu thập dữ liệu đó trong bao lâu, đặc biệt nếu bạn đang theo dõi xu hướng trong một khoảng thời gian dài. Nếu bạn cần câu trả lời ngay lập tức, bạn có thể không có thời gian để thu thập dữ liệu mới. Trong trường hợp này, bạn cần sử dụng dữ liệu lịch sử đã có sẵn.

## 2. Thu thập dữ liệu



### Dữ liệu chính (primary data) và Dữ liệu thứ cấp (secondary data)

Phân loại định dạng dữ liệu	Định nghĩa	Ví dụ
Dữ liệu chính	Thu thập trực tiếp bởi các nhà nghiên cứu	<ul style="list-style-type: none"> <li>- Dữ liệu từ một cuộc phỏng vấn bạn đã thực hiện</li> <li>- Dữ liệu từ một cuộc khảo sát được trả lại từ 20 người tham gia</li> <li>- Dữ liệu từ bảng câu hỏi bạn nhận lại từ một nhóm công nhân</li> </ul>
Dữ liệu thứ cấp	Thu thập gián tiếp bởi những người khác	<ul style="list-style-type: none"> <li>- Dữ liệu bạn đã mua từ hồ sơ khách hàng của một công ty phân tích dữ liệu địa phương</li> <li>- Dữ liệu nhân khẩu học do một trường đại học thu thập</li> <li>- Dữ liệu điều tra dân số do chính phủ liên bang thu thập</li> </ul>

### Dữ liệu nội bộ (internal data) và Dữ liệu bên ngoài (external data)

Phân loại định dạng dữ liệu	Định nghĩa	Ví dụ
Dữ liệu nội bộ	Dữ liệu tồn tại bên trong hệ thống của chính công ty	<ul style="list-style-type: none"> <li>- Tiền lương của nhân viên trong các đơn vị kinh doanh khác nhau được HR theo dõi</li> <li>- Dữ liệu bán hàng theo vị trí cửa hàng</li> <li>- Mức tồn kho sản phẩm trên các trung tâm phân phối</li> </ul>
Dữ liệu bên ngoài	Dữ liệu nằm bên ngoài công ty hoặc tổ chức	<ul style="list-style-type: none"> <li>- Mức lương trung bình trên toàn quốc cho các vị trí khác nhau trong toàn bộ tổ chức của bạn</li> <li>- Báo cáo tín dụng về khách hàng của một đại lý ô tô</li> </ul>

## Dữ liệu liên tục và dữ liệu rời rạc

Phân loại định dạng dữ liệu	Định nghĩa	Ví dụ
Dữ liệu liên tục	Dữ liệu được đo lường và có thể có hầu hết mọi giá trị số	<ul style="list-style-type: none"> <li>- Chiều cao của trẻ ở các lớp lớp ba (52.5 inch, 65.7 inch)</li> <li>- Đánh dấu thời gian chạy trong video</li> <li>- Nhiệt độ</li> </ul>
Dữ liệu rời rạc	Dữ liệu được đếm và có một số giá trị giới hạn	<ul style="list-style-type: none"> <li>- Số người đến bệnh viện hàng ngày (10, 20, 200)</li> <li>- Sức chứa tối đa của phòng cho phép</li> <li>- Vé đã bán trong tháng hiện tại</li> </ul>

## Dữ liệu định tính và dữ liệu định lượng

Phân loại định dạng dữ liệu	Định nghĩa	Ví dụ
Định tính	Đo lường một cách chủ quan và giải thích về chất lượng và đặc trưng	<ul style="list-style-type: none"> <li>- Hoạt động thể dục yêu thích nhất</li> <li>- Nhãn hiệu yêu thích của hầu hết khách hàng trung thành</li> <li>- Sở thích thời trang của thanh niên</li> </ul>
Định lượng	Đo lường cụ thể và khách quan thông qua các dữ liệu số	<ul style="list-style-type: none"> <li>- Tỷ lệ bác sĩ được hội đồng chứng nhận là phụ nữ</li> <li>- Số lượng voi ở Châu Phi</li> <li>- Khoảng cách từ Trái đất đến sao Hỏa</li> </ul>

## Dữ liệu có thứ tự (nominal) và không thứ tự (ordinal)

Phân loại định dạng dữ liệu	Định nghĩa	Ví dụ
Thứ tự	Loại dữ liệu định lượng không thể	<ul style="list-style-type: none"> <li>- Khách hàng lần đầu, khách hàng cũ, khách hàng thường xuyên</li> </ul>



Phân loại định dạng dữ liệu	Định nghĩa	Ví dụ
	phân loại dựa trên thứ tự	<ul style="list-style-type: none"> <li>- Ứng viên mới xin việc, ứng viên hiện có, ứng viên nội bộ</li> <li>- Niêm yết mới, giảm giá niêm yết, tịch thu nhà</li> </ul>
Không thứ tự	Loại dữ liệu định lượng có thứ tự hoặc độ đo	<ul style="list-style-type: none"> <li>- Xếp hạng phim (số sao: 1 sao, 2 sao, 3 sao)</li> <li>- Các lựa chọn bầu cử dựa trên thứ hạng (1, 2, 3)</li> <li>- Mức thu nhập (thu nhập thấp, thu nhập trung bình, thu nhập cao)</li> </ul>

### Có cấu trúc và phi cấu trúc

Phân loại định dạng dữ liệu	Định nghĩa	Ví dụ
Cấu trúc	Dữ liệu được tổ chức theo một định dạng nhất định, như dòng và cột	<ul style="list-style-type: none"> <li>- Báo cáo chi tiêu</li> <li>- Hoàn thuế</li> <li>- Lưu trữ hàng tồn kho</li> </ul>
Phi cấu trúc	Dữ liệu không được tổ chức theo một cách dễ nhận ra	<ul style="list-style-type: none"> <li>- Các bài đăng trên mạng xã hội</li> <li>- Email</li> <li>- Video</li> </ul>



### 3. Định dạng và mô hình hóa dữ liệu

Dữ liệu ở khắp mọi nơi và nó có thể được lưu trữ theo nhiều cách. Hai loại dữ liệu chính là:

- **Dữ liệu có cấu trúc:** được tổ chức theo một định dạng nhất định, chẳng hạn như hàng và cột.
- **Dữ liệu phi cấu trúc:** không được sắp xếp theo bất kỳ cách nào dễ nhận biết.

Ví dụ: khi bạn xếp hạng nhà hàng yêu thích của mình trực tuyến, bạn đang tạo dữ liệu có cấu trúc. Nhưng khi bạn sử dụng Google Earth để xem hình ảnh vệ tinh về vị trí nhà hàng, bạn đang sử dụng dữ liệu phi cấu trúc.

Dưới đây là phần bổ sung về các đặc điểm của dữ liệu có cấu trúc và dữ liệu phi cấu trúc:

Dữ liệu có cấu trúc	Dữ liệu phi cấu trúc
	
<ul style="list-style-type: none"> <li>• Kiểu dữ liệu được xác định</li> <li>• Dữ liệu định lượng</li> <li>• Dễ tổ chức</li> <li>• Dễ tìm kiếm</li> <li>• Dễ phân tích</li> <li>• CSDL quan hệ và nhà kho dữ liệu</li> <li>• Chứa dòng và cột</li> <li>• Ví dụ: Excel, Google sheet, SQL, thông tin khách hàng, bản ghi điện thoại, lịch sử giao dịch</li> </ul>	<ul style="list-style-type: none"> <li>• Nhiều kiểu dữ liệu khác nhau</li> <li>• Dữ liệu định tính</li> <li>• Khó tìm kiếm</li> <li>• Tự do hơn để phân tích</li> <li>• Hồ dữ liệu, kho dữ liệu, và NoSQL</li> <li>• Không thể đặt trong dòng và cột</li> <li>• Ví dụ: bài review sản phẩm, hình ảnh, âm thanh, video, tin nhắn, phụ đề điện thoại</li> </ul>

### Dữ liệu có cấu trúc

Như chúng ta đã mô tả trước đó, **dữ liệu có cấu trúc** được tổ chức theo một định dạng nhất định. Điều này giúp cho việc lưu trữ và truy vấn phục vụ nhu cầu công việc trở nên dễ dàng hơn. Nếu dữ liệu được trích xuất, cấu trúc sẽ đi cùng với dữ liệu.

## Dữ liệu phi cấu trúc

**Dữ liệu phi cấu trúc** không thể được sắp xếp theo bất kỳ cách nào dễ nhận dạng. Có nhiều dữ liệu phi cấu trúc hơn dữ liệu có cấu trúc trên thế giới. Các file video và âm thanh, file văn bản, nội dung mạng xã hội, hình ảnh vệ tinh, bản thuyết trình, tệp PDF, câu trả lời khảo sát mở và trang web đều đủ điều kiện của loại dữ liệu phi cấu trúc.

## Vấn đề công bằng

Việc thiếu cấu trúc làm cho dữ liệu phi cấu trúc khó tìm kiếm, quản lý và phân tích. Nhưng những tiến bộ gần đây trong trí tuệ nhân tạo và các thuật toán học máy đang bắt đầu thay đổi điều đó. Giờ đây, thách thức mới mà các nhà khoa học dữ liệu phải đối mặt là đảm bảo các công cụ này có tính toàn diện và không thiên vị. Nếu không, các phần tử nhất định của tập dữ liệu sẽ có trọng số và đại diện nhiều hơn các phần tử khác. Khi bạn đang học, một tập dữ liệu không công bằng không đại diện chính xác cho tập hợp, gây ra kết quả sai lệch, mức độ chính xác thấp và phân tích không đáng tin cậy.

## 4. Khám phá các loại dữ liệu

Bài đọc này giới thiệu cho bạn về mô hình hóa dữ liệu và các loại mô hình dữ liệu khác nhau. Mô hình dữ liệu giúp giữ cho dữ liệu nhất quán và cho phép mọi người chỉ ra cách dữ liệu được tổ chức. Hiểu biết cơ bản giúp các nhà phân tích và các bên liên quan khác dễ dàng hiểu được dữ liệu của họ và sử dụng nó theo những cách phù hợp.

**Lưu ý:** Là một nhà phân tích dữ liệu cơ sở, bạn sẽ không được yêu cầu thiết kế một mô hình dữ liệu. Nhưng bạn có thể bắt gặp các mô hình dữ liệu hiện có mà tổ chức của bạn đã có.

## Mô hình dữ liệu là gì?

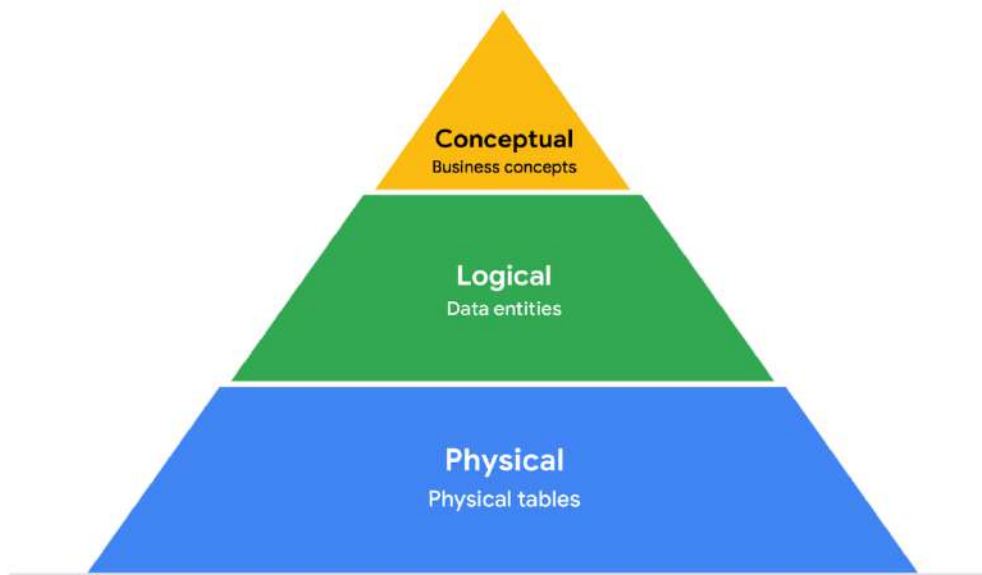
**Mô hình hóa dữ liệu** là quá trình tạo ra các sơ đồ thể hiện một cách trực quan cách dữ liệu được tổ chức và cấu trúc. Những biểu diễn trực quan này được gọi là **mô hình dữ liệu**. Bạn có thể xem mô hình dữ liệu như bản thiết kế của một ngôi nhà. Tại một thời điểm bất kỳ, sẽ có thợ điện, thợ mộc và thợ ống nước sử dụng bản thiết kế đó. Mỗi người trong số những người xây dựng này có mối quan hệ khác nhau với bản thiết kế, nhưng họ đều cần nó để hiểu cấu trúc tổng thể của ngôi nhà. Các mô hình dữ liệu cũng tương tự như

vậy; những người dùng khác nhau có các nhu cầu khác nhau đối với dữ liệu, nhưng mô hình dữ liệu cung cấp cho họ sự hiểu biết về cấu trúc tổng thể.

### Các cấp độ của mô hình hóa dữ liệu

Mỗi cấp độ của mô hình hóa dữ liệu có một mức độ chi tiết khác nhau.

#### The three most common types of data modeling



**Mô hình hóa dữ liệu khái niệm (conceptual data model)** cung cấp cái nhìn cấp cao về cấu trúc dữ liệu, chẳng hạn như cách dữ liệu tương tác trong một tổ chức. Ví dụ, một mô hình dữ liệu khái niệm có thể được sử dụng để xác định các yêu cầu nghiệp vụ cho một cơ sở dữ liệu mới. Mô hình dữ liệu khái niệm không chứa các chi tiết kỹ thuật.

**Mô hình hóa dữ liệu logic (logical data model)** tập trung vào các chi tiết kỹ thuật của cơ sở dữ liệu như các mối quan hệ, thuộc tính và thực thể. Ví dụ, một mô hình dữ liệu logic xác định cách các dòng riêng lẻ được xác định duy nhất trong cơ sở dữ liệu. Nhưng nó không nói rõ tên của các bảng cơ sở dữ liệu. Đó là công việc của một mô hình dữ liệu vật lý.

**Mô hình dữ liệu vật lý (physical data modelling)** mô tả cách một cơ sở dữ liệu hoạt động. Một mô hình dữ liệu vật lý xác định tất cả các thực thể và thuộc tính được sử dụng; ví dụ, nó bao gồm tên bảng, tên cột và kiểu dữ liệu cho cơ sở dữ liệu.

Bạn có thể tìm thấy thêm thông tin trong phần [so sánh các mô hình dữ liệu](#) này.

### Các kỹ thuật mô hình hóa dữ liệu

Có rất nhiều cách tiếp cận khi phát triển mô hình dữ liệu, nhưng hai phương pháp phổ biến là Sơ đồ mối quan hệ thực thể (**Entity Relationship Diagram - ERD**) và sơ đồ Ngôn ngữ mô hình hóa hợp nhất (**Unified Modeling Language - UML**). ERD là một cách trực quan để hiểu mối quan hệ giữa các thực thể trong mô hình dữ liệu. Biểu đồ UML là những biểu đồ rất chi tiết mô tả cấu trúc của hệ thống bằng cách chỉ ra các thực thể, thuộc tính, hoạt động của hệ thống và các mối quan hệ của chúng. Là một nhà phân tích dữ liệu học việc, bạn sẽ cần hiểu rằng có các kỹ thuật mô hình hóa dữ liệu khác nhau, nhưng trên thực tế, bạn có thể sẽ sử dụng kỹ thuật hiện có của tổ chức mình.

Bạn có thể đọc thêm về ERD, UML và từ điển dữ liệu trong bài viết [kỹ thuật mô hình dữ liệu](#) này.

### Phân tích dữ liệu và mô hình hóa dữ liệu

Mô hình hóa dữ liệu có thể giúp bạn khám phá các chi tiết cấp cao của dữ liệu của bạn và cách dữ liệu đó liên quan trên các hệ thống thông tin của tổ chức. Mô hình hóa dữ liệu đôi khi yêu cầu phân tích dữ liệu để hiểu cách dữ liệu được kết hợp với nhau; bằng cách đó, bạn biết ánh xạ dữ liệu. Cuối cùng, mô hình dữ liệu giúp mọi người trong tổ chức của bạn hiểu và cộng tác với bạn trên dữ liệu của bạn dễ dàng hơn. Điều này quan trọng đối với bạn và mọi người trong nhóm của bạn!

### Chuyển đổi dữ liệu là gì?

Trong bài đọc này, bạn sẽ khám phá cách dữ liệu được chuyển đổi và sự khác biệt giữa dữ liệu rộng và dài. **Chuyển đổi dữ liệu** là quá trình thay đổi định dạng, cấu trúc hoặc giá trị của dữ liệu. Là một nhà phân tích dữ liệu, có nhiều khả năng bạn sẽ cần phải chuyển đổi dữ liệu vào một thời điểm nào đó để giúp bạn phân tích nó dễ dàng hơn.

Việc chuyển đổi dữ liệu thường bao gồm:

- Thêm, sao chép hoặc nhân bản dữ liệu
- Xóa dòng hoặc cột
- 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: lưu bảng tính dưới dạng file các giá trị được phân tách bằng dấu phẩy (comma separated values - CSV).

### Tại sao cần chuyển đổi dữ liệu

Các mục tiêu của việc chuyển đổi dữ liệu:

- **Tổ chức dữ liệu:** dữ liệu được tổ chức tốt hơn để sử dụng hơn
- **Tương thích dữ liệu:** các ứng dụng hoặc hệ thống khác nhau sau đó có thể sử dụng cùng một dữ liệu
- **Di chuyển dữ liệu:** dữ liệu có định dạng phù hợp có thể được di chuyển từ hệ thống này sang hệ thống khác
- **Hợp nhất dữ liệu:** dữ liệu thuộc cùng tổ chức có thể được hợp nhất với nhau
- **Nâng cao dữ liệu:** dữ liệu có thể được hiển thị với các trường chi tiết hơn
- **So sánh dữ liệu:** so sánh tương tự sau đó có thể được thực hiện

### Ví dụ về chuyển đổi dữ liệu: hợp nhất dữ liệu

Mario là một thợ sửa ống nước sở hữu một công ty hệ thống ống nước. Sau nhiều năm kinh doanh, anh ấy mua một công ty ống nước khác. Mario muốn hợp nhất thông tin khách hàng từ công ty mới mua lại với thông tin của chính mình, nhưng công ty kia sử dụng cơ sở dữ liệu khác. Vì vậy, Mario cần làm cho dữ liệu tương thích. Để làm được điều này, anh ấy phải chuyển đổi định dạng dữ liệu của công ty được mua. Sau đó, anh ta phải xóa các hàng trùng lặp cho những khách hàng chung của họ. Khi dữ liệu tương thích và cùng nhau, công ty hệ thống ống nước của Mario sẽ có một cơ sở dữ liệu khách hàng hoàn chỉnh và hợp nhất.

### Ví dụ về chuyển đổi dữ liệu: tổ chức dữ liệu (dài sang rộng)

Để tạo biểu đồ dễ dàng hơn, bạn có thể cần chuyển đổi dữ liệu dài sang dữ liệu rộng. Hãy xem xét ví dụ sau về việc chuyển đổi giá cổ phiếu (được thu thập dưới dạng dữ liệu dài) thành dữ liệu rộng.

**Dữ liệu dài** là dữ liệu trong đó **mỗi hàng chứa một điểm dữ liệu** cho một mục cụ thể. Trong ví dụ dữ liệu dài bên dưới, giá cổ phiếu riêng lẻ (điểm dữ liệu) đã được thu thập cho Apple (AAPL), Amazon (AMZN) và Google (GOOGL) (các mặt hàng cụ thể) vào những ngày nhất định.

Ví dụ về dữ liệu dài: giá cổ phiếu

Symbol	Date	Open
AAPL	2018-09-18	217.79
AAPL	2018-09-17	222.15
AAPL	2018-09-14	225.75
AAPL	2018-09-13	223.52
AMZN	2018-09-18	1918.65
AMZN	2018-09-17	1954.73
AMZN	2018-09-14	1992.93
AMZN	2018-09-13	2000
GOOGL	2018-09-18	1162.66
GOOGL	2018-09-17	1177.77
GOOGL	2018-09-14	1188
GOOGL	2018-09-13	1179.7

**Dữ liệu rộng** là dữ liệu trong đó **mỗi hàng chứa nhiều điểm dữ liệu** cho các mục cụ thể được xác định trong các cột.

Ví dụ dữ liệu rộng: giá cổ phiếu

Symbol	AAPL	AMZN	GOOGL
Date			
2018-09-13	223.52	2000	1179.7
2018-09-14	225.75	1992.93	1188
2018-09-17	222.15	1954.73	1177.77
2018-09-18	217.79	1918.65	1162.66

Với dữ liệu được chuyển đổi thành dữ liệu rộng, bạn có thể tạo biểu đồ so sánh mức độ thay đổi cổ phiếu của từng công ty trong cùng một khoảng thời gian.

Bạn có thể nhận thấy rằng tất cả dữ liệu có trong định dạng dài cũng ở định dạng rộng. Nhưng dữ liệu rộng sẽ dễ đọc và dễ hiểu hơn. Đó là lý do tại sao các nhà phân tích dữ liệu thường chuyển đổi dữ liệu dài sang dữ liệu rộng nhiều hơn chuyển đổi dữ liệu rộng thành dữ liệu dài. Bảng sau đây tóm tắt khi nào dùng mỗi định dạng:

Dữ liệu rộng được ưu tiên khi	Dữ liệu dài được ưu tiên khi
Tạo bảng và biểu đồ với một vài biến số về mỗi chủ đề	Lưu trữ nhiều biến về mỗi chủ đề. Ví dụ, lãi suất trong vòng 60 năm của mỗi ngân hàng
So sánh cụ thể các biểu đồ đường	Thực hiện phân tích hoặc vẽ đồ thị thống kê nâng cao

### \*Tài liệu tham khảo\*

- [1]<https://www.coursera.org/learn/data-preparation/supplement/7iFqv/selecting-the-right-data>
- [2]<https://www.coursera.org/learn/data-preparation/supplement/mBSNa/data-formats-in-practice>
- [3]<https://www.coursera.org/learn/data-preparation/supplement/tkt9D/the-structure-of-data>
- [4]<https://www.coursera.org/learn/data-preparation/supplement/vtp7L/data-modeling-levels-and-techniques>
- [5]<https://www.coursera.org/learn/data-preparation/supplement/EOCT4/transforming-data>



## Bài đọc 2: Đánh giá chất lượng của dữ liệu

### 1. Một số định nghĩa và thuật ngữ

#### Ẩn danh dữ liệu là gì?

Bạn đã học về tầm quan trọng của quyền riêng tư trong phân tích dữ liệu. Tiếp theo, chúng ta sẽ nói về **ẩn danh dữ liệu** và những loại dữ liệu nào nên được ẩn danh. **Thông tin nhận dạng cá nhân, Personally identifiable information (PII)**, là thông tin có thể được sử dụng một mình hoặc với dữ liệu khác để theo dõi danh tính của một người.

Ẩn danh dữ liệu là quá trình bảo vệ dữ liệu riêng tư hoặc nhạy cảm của mọi người bằng cách loại bỏ loại thông tin đó. Thông thường, ẩn danh dữ liệu bao gồm việc làm trống, băm hoặc che thông tin cá nhân, thường bằng cách sử dụng mã có độ dài cố định để đại diện cho các cột dữ liệu hoặc ẩn dữ liệu với các giá trị đã thay đổi.

#### Vai trò của bạn trong việc ẩn danh dữ liệu

Các tổ chức có trách nhiệm bảo vệ dữ liệu của họ và thông tin cá nhân mà dữ liệu có thể chứa. Là một nhà phân tích dữ liệu, bạn phải hiểu dữ liệu nào cần được ẩn danh, nhưng nhìn chung bạn không phải là người chịu trách nhiệm chính việc ẩn danh dữ liệu đó. Một ngoại lệ hiếm hoi nếu bạn sao chép dữ liệu với mục đích thử nghiệm hoặc phát triển. Trong trường hợp này, bạn có thể được yêu cầu ẩn danh dữ liệu trước khi làm việc với nó.

#### Loại dữ liệu nên được ẩn danh

Dữ liệu chăm sóc sức khỏe và tài chính là hai trong số những loại dữ liệu nhạy cảm nhất. Các ngành công nghiệp này phụ thuộc rất nhiều vào kỹ thuật ẩn danh dữ liệu. Nếu các thông tin này rò rỉ ra ngoài, hậu quả sẽ rất lớn. Đó là lý do tại sao dữ liệu trong hai ngành này thường trải qua quá trình **khử nhận dạng**, đây là một quy trình được sử dụng để **xóa sạch tất cả thông tin nhận dạng cá nhân khỏi dữ liệu**.

Ẩn danh dữ liệu được sử dụng trong hầu hết các ngành. Do đó các nhà phân tích dữ liệu cần phải hiểu những điều cơ bản. Dưới đây là danh sách dữ liệu thường được ẩn danh:

Số điện thoại

Tên

- Biển số xe
- Số an sinh xã hội (Căn cước công dân)
- Địa chỉ IP
- Hồ sơ bệnh án

- Địa chỉ email
- Ảnh chụp
- Số tài khoản

Đối với một số người, họ hiểu rằng loại dữ liệu này nên được ẩn danh. Đối với những người khác, chúng ta phải rất cụ thể về những gì cần được ẩn danh. Hãy tưởng tượng một thế giới mà tất cả chúng ta đều có quyền truy cập vào địa chỉ, số tài khoản và thông tin nhận dạng khác của nhau. Điều đó sẽ xâm phạm quyền riêng tư của nhiều người và khiến thế giới kém an toàn hơn. Ẩn danh dữ liệu là một trong những cách chúng ta có thể giữ dữ liệu riêng tư và an toàn!

## 2. Dữ liệu mở

Cũng giống như quyền riêng tư của dữ liệu, dữ liệu mở là một chủ đề được tranh luận rộng rãi ngày nay. Các nhà phân tích dữ liệu nghĩ nhiều về dữ liệu mở và là một nhà phân tích dữ liệu trong tương lai, bạn cần hiểu những điều cơ bản để thành công trong vai trò mới của mình.

### Dữ liệu mở là gì?

Trong phân tích dữ liệu, **dữ liệu mở** là một phần của **đạo đức dữ liệu**, liên quan đến việc sử dụng dữ liệu một cách có đạo đức. **Tính mở** đề cập đến quyền truy cập, sử dụng và chia sẻ dữ liệu miễn phí. Nhưng để dữ liệu được coi là mở, nó phải:

- Có sẵn và được công khai 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 để sử dụng, tái sử dụng và phân phối lại dữ liệu

Dữ liệu chỉ có thể được coi là mở khi nó đáp ứng cả ba tiêu chuẩn này.

### Dữ liệu nào nên được mở

Một trong những lợi ích lớn nhất của dữ liệu mở là những cơ sở dữ liệu đáng tin cậy có thể được sử dụng rộng rãi hơn. Tất cả những dữ liệu tốt đó có thể được tận dụng, chia sẻ và kết hợp với các dữ liệu khác. Điều này ảnh hưởng rất lớn đến sự hợp tác khoa học, những tiến bộ trong nghiên cứu, năng lực phân tích và ra quyết định. Nhưng điều quan trọng là phải nghĩ về các cá nhân được thể diện công khai, dữ liệu mở.

**Dữ liệu của bên thứ ba** được thu thập bởi một thực thể không có mối quan hệ trực tiếp với dữ liệu. Chúng ta đã nói về loại dữ liệu này trước đó. Ví dụ: các bên thứ ba có thể thu thập thông tin về khách truy cập vào một trang web nhất định. Làm điều này cho phép

bên thứ ba này tạo hồ sơ khách hàng, giúp họ hiểu rõ hơn về hành vi của người dùng và quảng cáo hiệu quả hơn với nhiều sự cá nhân hóa.

**Thông tin nhận dạng cá nhân (PII)** là dữ liệu có khả năng nhận dạng một người và làm cho thông tin về họ được biết đến. Điều quan trọng là phải giữ cho dữ liệu này an toàn. PII có thể bao gồm địa chỉ của một người, thông tin thẻ tín dụng, số an sinh xã hội, hồ sơ y tế, v.v.

Mọi người đều muốn giữ thông tin cá nhân về mình ở chế độ riêng tư. Bởi vì dữ liệu của bên thứ ba luôn sẵn có, điều quan trọng là phải cân bằng giữa tính mở của dữ liệu với quyền riêng tư của các cá nhân.

### 3. Các trang web và tài nguyên cho dữ liệu mở

May mắn cho các nhà phân tích dữ liệu, có rất nhiều trang web và tài nguyên đáng tin cậy có sẵn cho dữ liệu mở. Điều quan trọng cần nhớ là ngay cả những dữ liệu có uy tín cũng cần được đánh giá liên tục, nhưng những trang web này là điểm khởi đầu hữu ích:

- [Trang dữ liệu của chính phủ Hoa Kỳ](#): Data.gov là một trong những nguồn dữ liệu toàn diện nhất ở Hoa Kỳ. Tài nguyên này cung cấp cho người dùng dữ liệu và công cụ mà họ cần để nghiên cứu, thậm chí giúp họ phát triển các ứng dụng web và thiết bị di động cũng như thiết kế trực quan hóa dữ liệu.
- [Cục điều tra dân số Hoa Kỳ](#): Nguồn dữ liệu mở này cung cấp thông tin nhân khẩu học từ các chính quyền liên bang, tiểu bang và địa phương cũng như các tổ chức thương mại ở Hoa Kỳ.
- [Mạng dữ liệu mở](#): Nguồn dữ liệu này có một cỗ máy tìm kiếm thực sự mạnh mẽ và các bộ lọc nâng cao. Tại đây, bạn có thể tìm thấy dữ liệu về các chủ đề như tài chính, an toàn cộng đồng, cơ sở hạ tầng, nhà ở và phát triển.
- [Tập dữ liệu công khai trên đám mây của Google](#): Có một loạt các tập dữ liệu công khai có sẵn thông qua Chương trình tập dữ liệu công khai của Google Cloud mà bạn có thể tìm thấy đã được tải vào BigQuery.

- Tìm kiếm tập dữ liệu: Tìm kiếm tập dữ liệu là một cỗ máy tìm kiếm được thiết kế đặc biệt cho các tập dữ liệu; bạn có thể sử dụng công cụ này để tìm kiếm các tập dữ liệu cụ thể.

#### 4. Tài liệu tham khảo

[1]<https://www.coursera.org/learn/data-preparation/supplement/rtTel/data-anonymization>

[2]<https://www.coursera.org/learn/data-preparation/supplement/dj6K0/the-open-data-debate>

[3]<https://www.coursera.org/learn/data-preparation/supplement/3hAmz/sites-and-resources-for-open-data>

## Bài đọc 3: Cơ sở dữ liệu (CSDL) cho phân tích dữ liệu

### 1. Cơ sở dữ liệu trong phân tích dữ liệu?

Cơ sở dữ liệu (CSDL) cho phép các nhà phân tích thao tác, lưu trữ và xử lý dữ liệu. Điều này giúp họ tìm kiếm thông tin trên dữ liệu hiệu quả hơn rất nhiều.

#### Giới thiệu về cơ sở dữ liệu quan hệ

**Cơ sở dữ liệu quan hệ** là cơ sở dữ liệu chứa một loạt các bảng có thể được kết nối với nhau để hiển thị các mối quan hệ. Về cơ bản, chúng cho phép các nhà phân tích dữ liệu tổ chức và liên kết dữ liệu dựa trên những điểm chung của dữ liệu.

Trong một bảng thông thường (không phải bảng của cơ sở dữ liệu quan hệ), bạn sẽ thấy tất cả các biến mình quan tâm khi phân tích có mối quan hệ với nhau. Nếu không nhóm chúng lại quá trình xử lý sẽ khá khó khăn. Đây là một lý do tại sao cơ sở dữ liệu quan hệ rất phổ biến trong phân tích dữ liệu: chúng đơn giản hóa rất nhiều quy trình phân tích và làm cho dữ liệu dễ tìm và sử dụng hơn trên toàn bộ cơ sở dữ liệu.

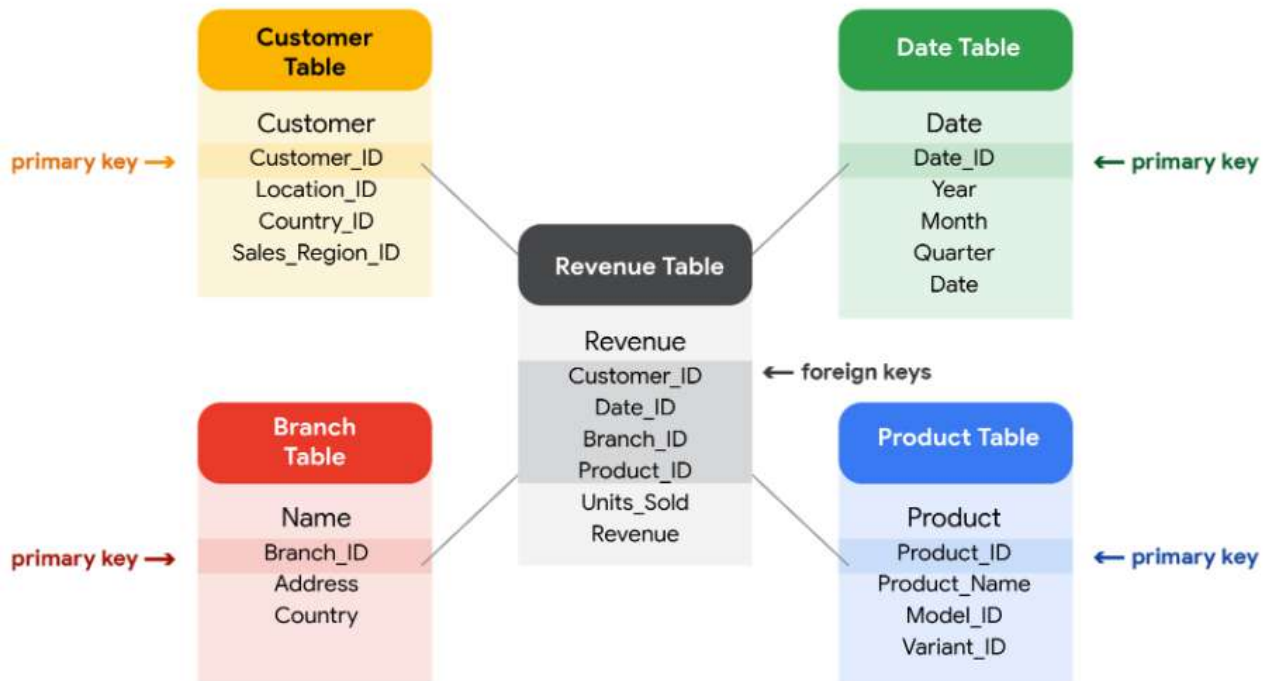
#### Các khái niệm chính trong CSDL quan hệ

Các bảng trong cơ sở dữ liệu quan hệ được kết nối với nhau bằng các trường chung mà chúng có. Bạn có thể đã nghe về khóa chính và khóa ngoại trước đây. Một cách ngắn gọn, **khóa chính** là một mã định danh tham chiếu đến một cột trong đó mỗi giá trị là duy nhất. Nói cách khác, đó là một cột của bảng được sử dụng để xác định duy nhất từng dòng trong bảng đó. Giá trị được gán cho khóa chính trong một hàng cụ thể phải là duy nhất trong toàn bộ bảng. Ví dụ: nếu `customer_id` là khóa chính cho bảng *Customer*, thì sẽ không có hai khách hàng nào có cùng một `customer_id`.

Ngược lại, **khóa ngoại** là một trường trong bảng, đồng thời là khóa chính trong bảng khác. Một bảng có thể chỉ có một khóa chính, nhưng nó có thể có nhiều khóa ngoại. Các khóa này chính là thứ tạo ra mối quan hệ giữa các bảng trong cơ sở dữ liệu quan hệ, giúp tổ chức và kết nối dữ liệu trên nhiều bảng trong cơ sở dữ liệu.

Một số bảng không yêu cầu khóa chính. Ví dụ: bảng *Revenue* có thể có nhiều khóa ngoại và không có khóa chính. Một khóa chính cũng có thể được tạo bằng cách sử dụng nhiều cột của bảng. Loại khóa chính này được gọi là **khóa tổng hợp**. Ví dụ: nếu `customer_id` và

location\_id là hai cột của khóa tổng hợp cho một bảng *Customer*, thì các giá trị được gán cho các trường đó trong bất kỳ hàng nào phải là duy nhất trong toàn bộ bảng.



## Giới thiệu về SQL

Cơ sở dữ liệu sử dụng một ngôn ngữ đặc biệt để giao tiếp được gọi là ngôn ngữ truy vấn. 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 cơ sở dữ liệu. Vì vậy, một nhà phân tích dữ liệu sẽ sử dụng SQL để tạo một truy vấn để xem dữ liệu cụ thể mà họ muốn từ bên trong tập hợp lớn hơn. Trong cơ sở dữ liệu quan hệ, các nhà phân tích dữ liệu có thể viết các truy vấn để lấy dữ liệu từ các bảng quan hệ. SQL là một công cụ mạnh mẽ để làm việc với cơ sở dữ liệu – chúng ta sẽ tìm hiểu thêm về nó ở các phần sau!

## 2. Kiểm tra tập dữ liệu

Là một nhà phân tích dữ liệu, bạn sẽ sử dụng dữ liệu để trả lời các câu hỏi và giải quyết vấn đề. Khi bạn phân tích dữ liệu và đưa ra kết luận, bạn đang tạo ra những hiểu biết sâu sắc có thể ảnh hưởng đến các quyết định kinh doanh, thúc đẩy thay đổi tích cực và giúp các bên liên quan của bạn đạt được mục tiêu của họ.

Trước khi bạn bắt đầu phân tích, phải kiểm tra dữ liệu của bạn để xác định xem dữ liệu đó có chứa thông tin cụ thể mà bạn cần để trả lời câu hỏi của các bên liên quan hay không. Trong bất kỳ tập dữ liệu nhất định nào, có thể xảy ra trường hợp:

- Dữ liệu không có (bạn có dữ liệu bánh sandwich, nhưng bạn cần dữ liệu bánh pizza)
- Dữ liệu không đủ (bạn có dữ liệu bánh pizza cho ngày 1-7 tháng 6, nhưng bạn cần dữ liệu cho cả tháng 6)
- Dữ liệu không chính xác (dữ liệu bánh pizza của bạn liệt kê giá một miếng là 250 đô la, điều này khiến bạn đặt câu hỏi về tính hợp lệ của tập dữ liệu)

Kiểm tra tập dữ liệu của bạn sẽ giúp bạn xác định những câu hỏi nào có thể trả lời được và những dữ liệu nào vẫn còn thiếu. Bạn có thể khôi phục dữ liệu này từ một nguồn bên ngoài hoặc ít nhất là giới thiệu cho các bên liên quan của bạn rằng nên sử dụng một nguồn dữ liệu khác.

Trong bài đọc này, hãy tưởng tượng bạn là một nhà phân tích dữ liệu đang kiểm tra dữ liệu bảng tính để xác định xem nó có thể trả lời câu hỏi của các bên liên quan hay không.

## Tình huống

Bạn là một nhà phân tích dữ liệu làm việc cho một công ty kem. Ban lãnh đạo quan tâm đến việc cải thiện doanh số bán kem của công ty.

Công ty đã thu thập dữ liệu về doanh số bán hàng của mình - nhưng không nhiều. Dữ liệu có sẵn là từ nguồn dữ liệu nội bộ và dựa trên doanh số bán hàng cho năm 2019. Bạn được yêu cầu xem lại dữ liệu và cung cấp một số thông tin chi tiết về doanh số bán kem của công ty. Ban giám đốc muốn có câu trả lời cho những câu hỏi sau:

1. Hương vị phổ biến nhất của kem là gì?
2. Nhiệt độ ảnh hưởng đến doanh số bán hàng như thế nào?
3. Cuối tuần và ngày lễ ảnh hưởng đến doanh số bán hàng như thế nào?
4. Lợi nhuận của khách hàng mới và khách hàng cũ khác nhau như thế nào?

## Download dữ liệu

Bạn có thể tải xuống dữ liệu dùng trong bài đọc này ở [đây](#).

## Khám phá dữ liệu

### Câu hỏi 1: Hương vị phổ biến nhất của kem là gì?



Để khám phá hương vị phổ biến nhất, trước tiên bạn cần xác định "phổ biến" có nghĩa là gì. Hương vị phổ biến nhất có phải là hương vị tạo ra doanh thu nhiều nhất trong năm 2019 không? Hay đó là hương vị có số lượng đơn vị bán ra nhiều nhất trong năm 2019? Đôi khi các lựa chọn đo lường của bạn bị giới hạn bởi dữ liệu bạn có — bạn có thể xem lại bảng tính của mình để tìm hiểu xem một trong hai định nghĩa “phổ biến” này có hợp lý hay không dựa trên dữ liệu có sẵn.

Nhấp vào tab **hương vị** (*flavor*) trên bảng tính của bạn để xem dữ liệu có liên quan. Trang tính **flavor** có ba cột và 209 hàng dữ liệu. Các tiêu đề cột là **tuần** (week), **đơn vị đã bán** (*unit sold*) và **hương vị** (*flavor*). Tập dữ liệu này không đi kèm với mô tả dữ liệu, vì vậy bạn phải tự mình tìm ra ý nghĩa của các cột. Dựa trên dữ liệu, bạn suy ra rằng các cột này cung cấp thông tin về số lượng đơn vị đã bán cho mỗi hương vị kem, theo tuần, trong năm 2019.

week	units sold	flavor
1	6	chocolate
1	15	lemon
1	12	strawberry
1	6	vanilla
2	16	chocolate
2	7	lemon
2	7	strawberry

Trong trường hợp này, bạn có thể tìm ra hương vị phổ biến nhất bằng cách sử dụng **đơn vị đã bán** làm thước đo cho mình. Đặc biệt, bạn có thể sử dụng cột đơn vị đã bán để tính tổng số đơn vị bán được trong năm cho từng hương vị. Thật không may, tập dữ liệu không cung cấp số lượng bán hàng hàng năm theo hương vị. Trong trường hợp này, bước tiếp theo của bạn sẽ là hỏi các bên liên quan xem dữ liệu về doanh số hàng năm trên mỗi hương vị có sẵn từ một nguồn khác hay không. Nếu không, bạn có thể thêm tuyên bố về các giới hạn của dữ liệu hiện tại vào phân tích của mình.

## Câu hỏi 2: Nhiệt độ ảnh hưởng đến doanh số bán hàng như thế nào?

Để khám phá câu hỏi thứ hai của bạn, bạn nhấp vào tab **nhiệt độ** (*temperatures*) và xem dữ liệu. Bảng nhiệt độ có hai cột và 366 hàng dữ liệu. Các tiêu đề cột là **nhiệt độ** (*temperature*) và **doanh số bán hàng** (*sales*). Vì không có bảng mô tả dữ liệu đi kèm, chúng ta có thể hiểu bảng này theo hai cách. Cách 1: Dữ liệu có thể hiển thị tổng doanh



thu năm 2019 theo nhiệt độ, ví dụ: hàng đầu tiên là tổng doanh thu \$39,69 trong ba ngày riêng biệt mà mỗi ngày có nhiệt độ cao nhất 60 độ. Cách 2: dữ liệu có thể hiển thị doanh số bán hàng và nhiệt độ cho từng ngày trong năm 2019, hàng đầu tiên để cập đến một ngày duy nhất có nhiệt độ cao nhất là 60 độ và doanh thu \$ 39,69).

temperature	sales
60	39.69
80	61.59
58	33.44
96	80.02
95	80.75
54	31.94
74	53.62
58	36.79
54	28.37
46	19.81
71	53.77
74	58.61
61	38.21

Vậy, cách hiểu nào là đúng? Đây có thể là doanh số bán hàng hàng ngày vì có 365 dòng cho nhiệt độ và nhiều hàng có cùng nhiệt độ và giá trị bán hàng khác nhau. Điều này ngụ ý rằng mỗi hàng dành cho một ngày duy nhất và không phải là bản tóm tắt của nhiều ngày. Tuy nhiên, nếu không có thêm thông tin, bạn không thể chắc chắn về điều này. Ngoài ra, bạn không biết liệu dữ liệu hiện tại được liệt kê theo thứ tự liên tiếp theo ngày hay theo thứ tự khác. Bước tiếp theo của bạn là liên hệ với chủ sở hữu của tập dữ liệu để làm rõ.

Nếu nhiệt độ thực sự ảnh hưởng đến doanh số bán hàng, bạn sẽ có thể cung cấp cho các bên liên quan của mình thông tin chi tiết, chẳng hạn như sau: “Khi nhiệt độ cao nhất hàng ngày trên X độ, doanh số bán kem trung bình tăng thêm Y. Vì vậy, doanh nghiệp nên có kế hoạch tăng hàng tồn kho trong những thời điểm này để tối đa hóa doanh số bán hàng”.

### Câu hỏi 3: Cuối tuần và ngày lễ ảnh hưởng đến doanh số bán hàng như thế nào?

Tiếp theo, bạn bấm vào tab **bán hàng (sales)** để xem dữ liệu về các ngày bán. Bảng bán hàng có hai cột và 366 hàng dữ liệu. Các tiêu đề cột là **ngày (dates)** và **doanh số bán hàng (sales)**. Dữ liệu này rất có thể là tổng doanh số hàng ngày trong năm 2019, vì doanh số bán hàng được ghi lại cho từng ngày trong năm 2019.

date	sales
1/1/2019	59.96
1/2/2019	67.06
1/3/2019	74.24
1/4/2019	78.11
1/5/2019	84.76
1/6/2019	100.60
1/7/2019	100.13
1/8/2019	96.36
1/9/2019	85.80
1/10/2019	70.39
1/11/2019	60.81
1/12/2019	58.66
1/13/2019	61.10

Bạn có thể sử dụng dữ liệu này để xác định xem một ngày cụ thể rơi vào cuối tuần hay ngày lễ và thêm cột phản ánh thông tin này vào trang tính của bạn. Từ đó, bạn có thể tìm hiểu xem liệu doanh số bán hàng vào cuối tuần và ngày lễ có lớn hơn doanh số bán hàng vào những ngày khác hay không. Điều này sẽ rất hữu ích cho việc lập kế hoạch lưu kho và tiếp thị.

#### **Câu hỏi 4: Lợi nhuận của khách hàng mới và khách hàng cũ khác nhau như thế nào?**

Tập dữ liệu của bạn không chứa dữ liệu bán hàng liên quan đến khách hàng mới. Nếu không có dữ liệu này, bạn sẽ không thể trả lời câu hỏi cuối cùng của mình. Tuy nhiên, có thể xảy ra trường hợp công ty thu thập dữ liệu khách hàng và lưu trữ trong một bảng dữ liệu khác.

Nếu vậy, bước tiếp theo của bạn là tìm hiểu cách truy cập vào dữ liệu khách hàng của công ty. Sau đó, bạn có thể kết hợp dữ liệu doanh thu với bảng dữ liệu khách hàng để phân loại từng lần bán hàng từ khách hàng mới hoặc khách hàng cũ và phân tích sự khác biệt về lợi nhuận giữa hai nhóm khách hàng. Thông tin này sẽ giúp các bên liên quan phát triển các chiến dịch tiếp thị cho các loại khách hàng cụ thể hơn để tăng lòng trung thành với thương hiệu và lợi nhuận tổng thể.

#### **Kết luận**

Khi làm việc trên các dự án phân tích, không phải lúc nào bạn cũng có tất cả dữ liệu cần thiết hoặc có liên quan. Trong nhiều trường hợp như vậy, bạn có thể chuyển sang các nguồn dữ liệu khác để lấp đầy khoảng trống.

Bất chấp những hạn chế của tập dữ liệu, bạn vẫn có thể cung cấp cho các bên liên quan một số thông tin chi tiết có giá trị. Tiếp theo, kế hoạch hành động tốt nhất cho bạn sẽ là chủ động đặt câu hỏi, xác định các bộ dữ liệu có liên quan khác hoặc tự mình thực hiện một số nghiên cứu. Bất kể bạn đang làm việc với dữ liệu nào, việc kiểm tra cẩn thận dữ liệu sẽ có tác động lớn đến chất lượng tổng thể của phân tích.

### 3. Tầm quan trọng của siêu dữ liệu

Phân tích dữ liệu là một lĩnh vực phát triển mạnh về thu thập và tổ chức dữ liệu. Trong bài đọc này, bạn sẽ tìm hiểu về cách phân tích và hiểu thấu đáo mọi khía cạnh về dữ liệu của bạn.

Nhìn vào bất kỳ dữ liệu nào bạn tìm thấy. Một số câu hỏi đặt ra. Dữ liệu này là gì? Nó từ đâu đến? Nó hữu ích hay không? Làm sao bạn biết? Đây là nơi siêu dữ liệu xuất hiện để cung cấp sự hiểu biết sâu hơn về dữ liệu. Nói một cách đơn giản, siêu dữ liệu là dữ liệu về dữ liệu. Trong quản trị cơ sở dữ liệu, nó cung cấp thông tin về các dữ liệu khác và giúp các nhà phân tích dữ liệu diễn giải nội dung của dữ liệu trong cơ sở dữ liệu.

Bất kể bạn đang làm việc với số lượng dữ liệu lớn hay nhỏ, siêu dữ liệu cho thấy sự hiểu biết của nhóm phân tích, giúp giao tiếp về dữ liệu trong toàn doanh nghiệp và giúp việc sử dụng lại dữ liệu dễ dàng hơn. Về bản chất, siêu dữ liệu cho biết ai, thế nào, khi nào, ở đâu, cái nào, như thế nào và tại sao có dữ liệu.

#### Các thành phần của siêu dữ liệu

Trước khi đi vào các ví dụ về siêu dữ liệu, chúng ta sẽ tìm hiểu các loại siêu dữ liệu thường dùng.

##### **Tiêu đề và mô tả**

Tên của file hoặc trang web bạn đang kiểm tra là gì? Nó chứa loại nội dung gì?

##### **Thẻ và danh mục**

Tổng quan chung về dữ liệu mà bạn có là gì? Dữ liệu có được lập chỉ mục hoặc mô tả một cách cụ thể không?

##### **Ai đã tạo ra nó và khi nào**

Dữ liệu đến từ đâu và được tạo khi nào? Được tạo ra gần đây, hay nó đã tồn tại trong một thời gian dài?

##### **Ai đã sửa đổi nó lần cuối và khi nào**

Có bất kỳ thay đổi nào được thực hiện đối với dữ liệu không? Nếu có, có phải các sửa đổi xảy ra gần đây hay không?

## **Ai có thể truy cập hoặc cập nhật nó**

Tập dữ liệu này có công khai không? Các quyền đặc biệt có cần thiết để tùy chỉnh hoặc sửa đổi tập dữ liệu không?

### **Ví dụ về siêu dữ liệu**

Trong thế giới kỹ thuật số ngày nay, siêu dữ liệu có ở khắp mọi nơi và việc cung cấp siêu dữ liệu trên nhiều phương tiện và thông tin mà bạn tương tác ngày càng trở nên phổ biến hơn. Dưới đây là một số ví dụ thực tế về siêu dữ liệu:

#### **Hình ảnh**

Bất cứ khi nào ảnh được chụp bằng máy ảnh, siêu dữ liệu như tên file, ngày, giờ và vị trí địa lý sẽ được thu thập và lưu cùng với ảnh đó.

#### **Email**

Khi một email được gửi hoặc nhận, có rất nhiều siêu dữ liệu hiển thị như dòng tiêu đề, người gửi, người nhận và ngày giờ gửi. Ngoài ra còn có siêu dữ liệu ẩn bao gồm tên máy chủ, địa chỉ IP, định dạng HTML và chi tiết phần mềm.

#### **Bảng tính và tài liệu**

Các bảng tính và tài liệu đã chứa đầy một lượng dữ liệu đáng kể nên không có gì ngạc nhiên khi siêu dữ liệu cũng đi kèm với chúng. Tiêu đề, tác giả, ngày tạo, số trang, nhận xét của người dùng cũng như tên của các tab, bảng và cột đều là siêu dữ liệu mà người ta có thể tìm thấy trong bảng tính và tài liệu.

#### **Trang web**

Mỗi trang web đều có một số trường siêu dữ liệu chuẩn, chẳng hạn như thẻ và danh mục, tên của người tạo trang web, tiêu đề và mô tả trang web, thời gian tạo và bất kỳ icon này.

#### **File kỹ thuật số**

Thông thường, nếu bạn nhấp chuột phải vào bất kỳ file máy tính nào, bạn sẽ thấy siêu dữ liệu của nó. Điều này có thể bao gồm tên file, kích thước file, ngày tạo và sửa đổi cũng như loại file.

#### **Sách**

Siêu dữ liệu không chỉ là kỹ thuật số. Mỗi cuốn sách đều có một số siêu dữ liệu chuẩn trên bìa và bên trong sẽ cho bạn biết tên sách, tên tác giả, mục lục, thông tin nhà xuất bản, mô tả bản quyền, chỉ mục và mô tả ngắn gọn về nội dung của cuốn sách.

### **Ngữ cảnh của dữ liệu**

Biết được nội dung và bối cảnh của dữ liệu cũng như cách cấu trúc dữ liệu của bạn, rất có giá trị trong sự nghiệp của bạn với tư cách là một nhà phân tích dữ liệu. Khi phân tích dữ liệu, điều quan trọng là phải luôn hiểu được bức tranh toàn cảnh. Nó không chỉ là về dữ

liệu bạn đang xem mà còn là cách các dữ liệu đó kết hợp với nhau. Siêu dữ liệu đảm bảo rằng bạn có thể tìm, sử dụng, bảo quản và tái sử dụng dữ liệu trong tương lai. Hãy nhớ rằng bạn có trách nhiệm quản lý và sử dụng toàn bộ dữ liệu; siêu dữ liệu cũng quan trọng như chính dữ liệu..

## 4. Nhập dữ liệu từ một nguồn bên ngoài vào bảng tính

Khi bạn làm việc với bảng tính, có một số cách khác nhau để nhập dữ liệu. Bài đọc này trình bày cách bạn có thể nhập dữ liệu từ các nguồn bên ngoài, cụ thể là:

- Các bảng tính khác
- Tập CSV
- Bảng HTML (trong các trang web)

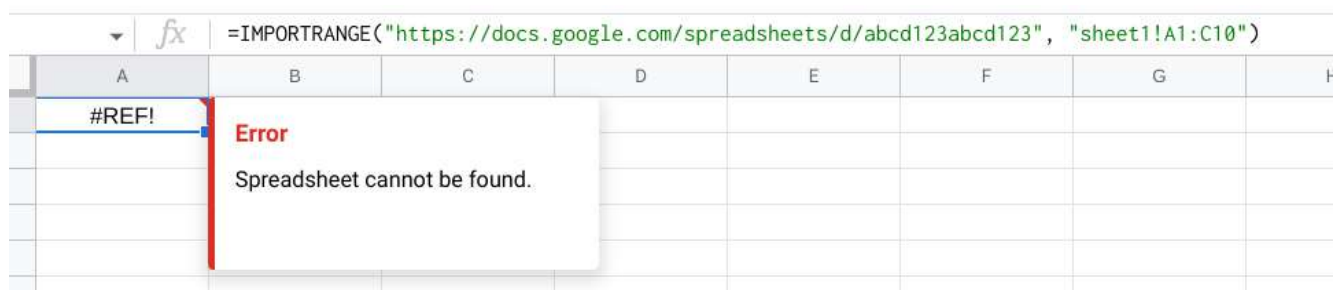
### Nhập dữ liệu từ các bảng tính khác

Trong nhiều trường hợp, bạn có thể mở một bảng tính hiện có và cần thêm dữ liệu bổ sung từ một bảng tính khác.

#### Google sheet

Trong Google sheet, bạn có thể sử dụng hàm IMPORTRANGE. Nó cho phép bạn xác định một dải các ô trong bảng tính khác để sao chép vào bảng tính mà bạn đang làm việc. Bạn phải được cho phép truy cập vào bảng tính chứa dữ liệu trong lần đầu tiên bạn nhập dữ liệu.

**URL hiển thị bên dưới chỉ là ví dụ về cú pháp. Đừng nhập nó vào bảng tính của bạn. Thay thế nó bằng một URL của một bảng tính bạn đã tạo để bạn có thể kiểm soát quyền truy cập vào nó bằng cách nhấp vào nút Cho phép truy cập.**



Tham khảo trang [IMPORTRANGE](#) của Trung tâm trợ giúp của Google để biết thêm thông tin về cú pháp. Ngoài ra còn có ví dụ được sử dụng sau trong chương trình ở [Chức năng nâng cao để làm sạch dữ liệu nhanh chóng](#).

#### Microsoft Excel

Để nhập dữ liệu từ một bảng tính khác, hãy làm như sau:

**Bước 1:** Chọn **Data** từ menu chính.

**Bước 2:** Nhấp vào **Get Data**, chọn **From File**, sau đó chọn **From Workbook**

**Bước 3:** Duyệt và chọn file bảng tính, sau đó nhấp vào **Import**.

**Bước 4:** Trong **Navigator**, chọn trang tính cần nhập.

**Bước 5:** Nhấp vào **Load** để nhập tất cả dữ liệu trong trang tính; hoặc nhấp vào **Transform Data** để mở Power Query Editor từ đó điều chỉnh các cột và hàng dữ liệu bạn muốn nhập.

**Bước 6:** Nếu bạn đã nhấp vào Transform Data, hãy nhấp vào **Close & Load** rồi chọn một trong hai tùy chọn:

**Close & Load** - nhập dữ liệu vào một trang tính mới

**Close & Load to ...** - nhập dữ liệu vào trang tính hiện có

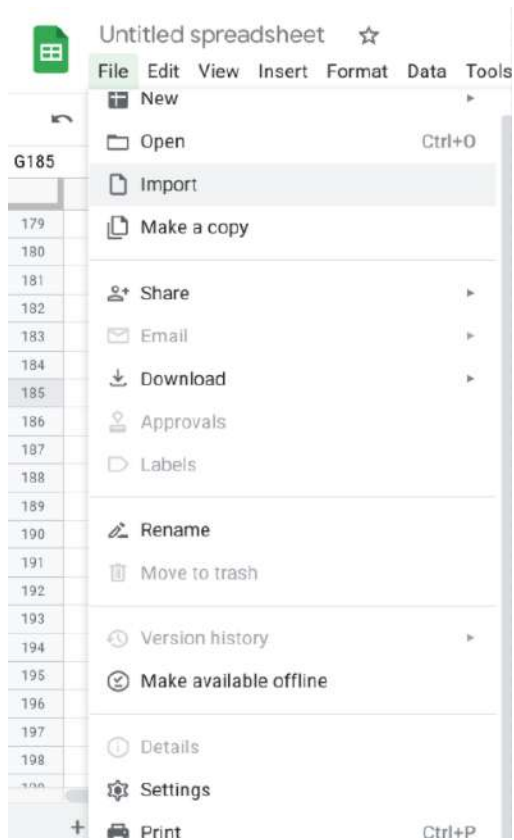
Nếu các hướng dẫn này không hoạt động đối với phiên bản Excel mà bạn có. Ghé thăm trung tâm đào tạo trực tuyến miễn phí, [Microsoft Excel for Windows Training](#), bạn sẽ tìm thấy mọi thứ bạn cần biết

Nếu bạn đang sử dụng Numbers, hãy tìm kiếm [Hướng dẫn sử dụng Numbers](#) để biết thêm thông tin.

## Nhập dữ liệu từ file CSV

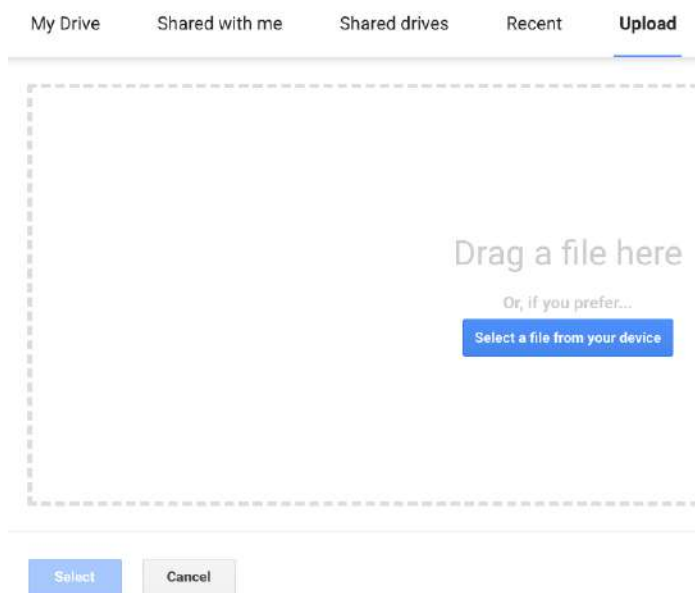
### Google sheet

**Bước 1:** Mở menu **File** trong bảng tính của bạn và chọn **Import** để mở cửa sổ Import

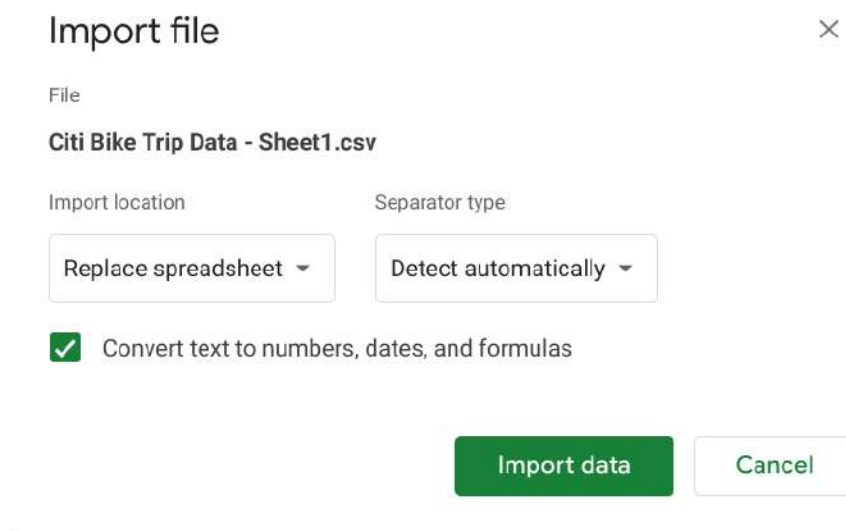


**Bước 2:** Chọn **Upload** rồi chọn file CSV bạn muốn nhập.

### Import file



**Bước 3:** Từ đây, bạn sẽ có một vài tùy chọn. Đối với **Import Location**, bạn có thể chọn thay thế bảng tính hiện tại, tạo bảng tính mới, chèn dữ liệu CSV dưới dạng trang tính mới, thêm dữ liệu vào bảng tính hiện tại hoặc thay thế dữ liệu trong một ô cụ thể. Dữ liệu sẽ chỉ được chèn dưới dạng văn bản thuần nếu bạn bỏ chọn “Convert text to numbers, dates, and formulas”, đây là cài đặt mặc định. Đôi khi file CSV sử dụng dấu phân tách như dấu chấm phẩy hoặc thậm chí là khoảng trắng thay vì dấu phẩy. Đối với **Separator type**, bạn có thể chọn Tab hoặc Comma hoặc chọn Custom để nhập một ký tự khác đang được sử dụng làm dấu phân tách.



**Bước 4:** Chọn **Import data**. Dữ liệu trong file CSV sẽ được tải vào trang tính của bạn và bạn có thể bắt đầu sử dụng nó!

**Lưu ý:** Bạn cũng có thể sử dụng hàm **IMPORTDATA** trong ô bảng tính để nhập dữ liệu bằng cách sử dụng URL vào file CSV. Tham khảo trang [IMPORTDATA](#) của Trung tâm trợ giúp của Google để biết thêm thông tin và cú pháp.

## Microsoft Excel

**Bước 1:** Mở bảng tính mới hoặc bảng tính hiện có

**Bước 2:** Nhấp vào **Data** trong menu chính và chọn tùy chọn **From Text / CSV**.

**Bước 3:** Duyệt và chọn file CSV, sau đó nhấp vào **Import**.

**Bước 4:** Từ đây, bạn sẽ có một vài tùy chọn. Bạn có thể thay đổi dấu phân cách từ dấu phẩy sang một ký tự khác, chẳng hạn như dấu chấm phẩy. Bạn cũng có thể bật hoặc tắt



tính năng phát hiện kiểu dữ liệu tự động. Và cuối cùng, bạn có thể chuyển đổi dữ liệu của mình bằng cách nhấp vào **Transform Data** để mở Power Query Editor.

**Bước 5:** Trong hầu hết các trường hợp, hãy chấp nhận cài đặt mặc định ở bước trước và nhấp vào **Load** để tải dữ liệu trong file CSV vào bảng tính. Dữ liệu trong file CSV sẽ được tải vào bảng tính và bạn có thể bắt đầu làm việc với dữ liệu.

Nếu các hướng dẫn này không hoạt động đối với phiên bản Excel mà bạn có. Ghé thăm trung tâm đào tạo trực tuyến miễn phí, [Microsoft Excel for Windows Training](#), bạn sẽ tìm thấy mọi thứ bạn cần biết

Nếu bạn đang sử dụng Numbers, hãy tìm kiếm [Hướng dẫn sử dụng Numbers](#) để biết thêm thông tin.

## Nhập các bảng HTML từ trang web

Nhập bảng HTML là một phương pháp rất cơ bản để trích xuất dữ liệu từ các trang web công cộng. [Rút trích nội dung trang Web](#) giới thiệu cách thực hiện việc này với Google Sheet hoặc Microsoft Excel.

### Google sheets

Trong Google Sheet, bạn có thể sử dụng hàm **IMPORTHTML**. Nó cho phép bạn nhập dữ liệu từ bảng HTML (hoặc danh sách) trên một trang web.

A1     =IMPORTHTML("http://en.wikipedia.org/wiki/Demographics\_of\_India", "table", 4)

	A	B	C	D	E	F	G	H	I	J	K
1	Years	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930[38]
2	Total Fertility Rate in India	5.761	5.77	5.78	5.79	5.8	5.81	5.82	5.83	5.85	5.86

Tham khảo trang [IMPORTHTML](#) của Trung tâm trợ giúp của Google để biết thêm thông tin về cú pháp. Nếu bạn đang nhập một danh sách, hãy thay thế "table" bằng "list" trong ví dụ trên. Số 4 là chỉ mục để cập đến thứ tự của các bảng trên một trang web. Nó giống như một con trỏ chỉ ra bảng nào trên trang mà bạn muốn nhập dữ liệu từ đó.

**Bạn có thể tự thử điều!** Trong trang tính trống, hãy sao chép và dán từng hàm IMPORTHTML sau vào ô A1 và xem điều gì sẽ xảy ra. Bạn sẽ thực sự nhập dữ liệu từ bốn bảng HTML khác nhau trong một bài viết trên Wikipedia: [Nhân khẩu học của Ấn Độ](#). Bạn có thể so sánh dữ liệu đã nhập của mình với các bảng trong bài viết.

= IMPORTHTML ("http://en.wikipedia.org/wiki/Demographics\_of\_India", "table", 1)

= IMPORTHTML ("http://en.wikipedia.org/wiki/Demographics\_of\_India", " table", 2)

= IMPORTHTML ("http://en.wikipedia.org/wiki/Demographics\_of\_India", "table", 3)  
= IMPORTHTML ("http://en.wikipedia.org/wiki/Demographics\_of\_India", "table", 4)

### Microsoft Excel

Bạn có thể nhập dữ liệu từ các trang web bằng tùy chọn **From Web**:

**Bước 1:** Mở bảng tính mới hoặc bảng tính hiện có.

**Bước 2:** Nhấp vào Data trong menu chính và chọn tùy chọn **From Web**.

**Bước 3:** Nhập URL và nhấp vào OK.

**Bước 4:** Trong Navigator, chọn bảng để nhập.

**Bước 5:** Nhấp vào **Load** để tải dữ liệu từ bảng vào bảng tính của bạn.

Nếu các hướng dẫn này không hoạt động đối với phiên bản Excel mà bạn có. Ghé thăm trung tâm đào tạo trực tuyến miễn phí, [Microsoft Excel for Windows Training](#), bạn sẽ tìm thấy mọi thứ bạn cần biết

Nếu bạn đang sử dụng Numbers, hãy tìm kiếm [Hướng dẫn sử dụng Numbers](#) để biết thêm thông tin.

## 5. Khám phá các tập dữ liệu công khai

**Dữ liệu mở** giúp tạo nhiều **tập dữ liệu công khai** mà bạn có thể truy cập để đưa ra quyết định hướng dữ liệu. Dưới đây là một số tài nguyên bạn có thể sử dụng để bắt đầu tìm kiếm tập dữ liệu công khai:

- [Bộ dữ liệu công khai trên đám mây của Google](#) cho phép các nhà phân tích dữ liệu truy cập vào bộ dữ liệu công khai có nhu cầu cao và giúp dễ dàng khám phá các hiểu biết trên đám mây.
- [Tìm kiếm tập dữ liệu](#) có thể giúp bạn tìm trực tuyến các tập dữ liệu có sẵn bằng các tìm kiếm từ khóa.
- [Kaggle](#) có chức năng tìm kiếm Open Data có thể giúp bạn tìm các tập dữ liệu để thực hành.
- Cuối cùng, [BigQuery](#) lưu trữ hơn 150 bộ dữ liệu công khai mà bạn có thể truy cập và sử dụng.

### Bộ dữ liệu sức khỏe cộng đồng

- Dữ liệu [Global Health Observatory](#): Bạn có thể tìm kiếm bộ dữ liệu từ trang này hoặc khám phá các bộ sưu tập dữ liệu nổi bật từ Tổ chức Y tế Thế giới.

- [Tập dữ liệu The Cancer Imaging Archive \(TCIA\)](#): Cũng giống như tập dữ liệu trước đó, dữ liệu này được lưu trữ bởi Google Cloud Public Datasets và có thể được tải lên BigQuery.
- [1000 Genome](#): Đây là một tập dữ liệu khác từ các tài nguyên Google Cloud Public có thể được tải lên BigQuery.

### Bộ dữ liệu khí hậu công khai

- [National Climatic Data Center](#): Trang Liên kết Nhanh của NCDC có tuyển chọn các tập dữ liệu mà bạn có thể khám phá.
- [NOAA Public Dataset Gallery](#): NOAA Public Dataset Gallery chứa một bộ sưu tập có thể tìm kiếm các tập dữ liệu công khai.

### Bộ dữ liệu chính trị - xã hội công khai

- [UNICEF State of the World's Children](#): Bộ dữ liệu này của UNICEF bao gồm một bộ sưu tập các bảng có thể được tải xuống.
- [CPS Labor Force Statistics](#): Trang này chứa các liên kết đến một số bộ dữ liệu có sẵn mà bạn có thể khám phá.
- [The Stanford Open Policing Project](#): Tập dữ liệu này có thể được tải xuống dưới dạng tệp .CSV để bạn sử dụng.

## 6. BigQuery

[BigQuery](#) là một nhà kho dữ liệu trên Google Cloud mà các nhà phân tích dữ liệu có thể sử dụng để truy vấn, lọc các tập dữ liệu lớn, tổng hợp kết quả và thực hiện các hoạt động phức tạp.

Bài đọc này cung cấp hướng dẫn để tạo tài khoản BigQuery của riêng bạn, chọn tập dữ liệu công khai và tải lên file CSV. Sau khi đọc xong bài đọc này, bạn có thể xác nhận quyền truy cập của mình vào bảng điều khiển BigQuery trước khi chuyển sang các hoạt động.

**Lưu ý:** Các tài nguyên bổ sung để làm việc trên một số nền tảng cơ sở dữ liệu SQL khác cũng được cung cấp ở cuối bài đọc này nếu bạn chọn làm việc với chúng thay vì BigQuery.

## Loại tài khoản BigQuery

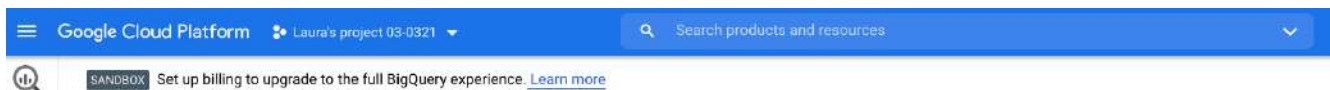
Có hai loại tài khoản khác nhau: sandbox và free trial. Tài khoản sandbox cho phép bạn thực hiện các truy vấn và khám phá các tập dữ liệu công khai miễn phí, nhưng có các hạn chế khác ngoài các *hạn ngạch* và *giới hạn tiêu chuẩn*. Nếu muốn sử dụng BigQuery với các giới hạn tiêu chuẩn, bạn có thể thiết lập tài khoản dùng thử miễn phí. Chi tiết hơn:

- **Tài khoản sandbox miễn phí:** không yêu cầu phương thức thanh toán. Tuy nhiên, giới hạn ở 12 dự án. Nó cũng không cho phép bạn chèn các bản ghi mới vào cơ sở dữ liệu hoặc cập nhật các giá trị trường của các bản ghi hiện có. Các hoạt động của ngôn ngữ thao tác dữ liệu (DML) này không được hỗ trợ trong sandbox.
- **Tài khoản free trial:** yêu cầu phương thức thanh toán để thiết lập tài khoản có thể lập hóa đơn (billable account), nhưng cung cấp đầy đủ chức năng trong thời gian dùng thử.

Dù dùng loại tài khoản nào, bạn có thể nâng cấp lên tài khoản trả phí bất kỳ lúc nào và giữ lại tất cả các dự án hiện có của mình. Nếu bạn thiết lập tài khoản dùng thử miễn phí nhưng chọn không nâng cấp lên tài khoản trả phí khi thời gian dùng thử kết thúc, bạn vẫn có thể thiết lập tài khoản sandbox miễn phí tại thời điểm đó. Tuy nhiên, các dự án từ tài khoản dùng thử của bạn sẽ không được chuyển sang tài khoản sandbox. Bạn sẽ cần phải bắt đầu lại từ đầu.

## Thiết lập tài khoản sandbox miễn phí để sử dụng trong bài này.

- Làm theo [hướng dẫn từng bước](#) sau hoặc xem video [Thiết lập BigQuery](#), bao gồm sandbox và các tùy chọn thanh toán.
- Để biết thêm thông tin chi tiết về cách sử dụng sandbox, hãy bắt đầu với tài liệu, [Sử dụng sandbox BigQuery](#).
- Sau khi thiết lập tài khoản, bạn sẽ thấy tên dự án bạn đã tạo cho tài khoản trong biểu ngữ và **SANDBOX** ở đầu bảng điều khiển BigQuery của bạn.



## Thiết lập một tài khoản dùng thử miễn phí (nếu bạn muốn)

Nếu không muốn gặp các giới hạn về sandbox trong BigQuery, bạn có thể thiết lập tài khoản dùng thử miễn phí để sử dụng trong chương trình này.

- Làm theo [hướng dẫn từng bước](#) sau hoặc xem video [Thiết lập BigQuery](#), bao gồm sandbox và các tùy chọn thanh toán. Bản dùng thử miễn phí cung cấp khoản tín dụng \$300 trong 90 ngày. Bạn sẽ không đạt đến gần giới hạn chi tiêu đó nếu chỉ sử dụng bảng điều khiển BigQuery để thực hành các truy vấn SQL. Sau khi bạn dùng hết khoản tín dụng \$300 (hoặc sau 90 ngày), thời gian dùng thử miễn phí của bạn sẽ hết hạn và bạn cần phải tự mình chọn nâng cấp lên tài khoản trả phí để tiếp tục sử dụng các dịch vụ của Google Cloud Platform, bao gồm cả BigQuery. **Phương thức thanh toán của bạn sẽ không bao giờ tự động bị tính phí sau khi thời gian dùng thử miễn phí của bạn kết thúc.** Nếu bạn chọn nâng cấp tài khoản của mình, bạn sẽ bắt đầu bị tính phí.

- Sau khi thiết lập tài khoản, bạn sẽ thấy **My First Project** trong biểu ngữ và trạng thái tài khoản của bạn phía trên biểu ngữ - số dư tín dụng và số ngày còn lại trong thời gian dùng thử của bạn.

 Free trial status: \$300.00 credit and 91 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.



## Cách truy cập bảng điều khiển BigQuery

Trong trình duyệt của bạn, truy cập đến [console.cloud.google.com/bigquery](https://console.cloud.google.com/bigquery).

**Lưu ý:** Truy cập [console.cloud.google.com](https://console.cloud.google.com) trong trình duyệt của bạn sẽ đưa bạn đến trang tổng quan chính của Google Cloud Platform. Để điều hướng đến BigQuery từ trang tổng quan, hãy làm như sau:

- Nhấp vào biểu tượng menu Navigator (biểu tượng Hamburger) trong biểu ngữ.
- Cuộn xuống phần **BIG DATA**.
- Nhấp vào **BigQuery** và chọn **SQL workspace**.

Xem video [Cách sử dụng BigQuery](#) để biết giới thiệu về từng phần của không gian làm việc BigQuery SQL.

## (Tùy chọn) Khám phá tập dữ liệu công khai BigQuery

Bạn sẽ khám phá tập dữ liệu công khai trong hoạt động sắp tới, vì vậy bạn có thể thực hiện các bước này sau nếu muốn.

- Tham khảo [hướng dẫn từng bước](#) này.

## (Tùy chọn) Tải file CSV lên BigQuery

Các bước này được cung cấp để bạn có thể tự làm việc với tập dữ liệu tại thời điểm này. Bạn sẽ tải file CSV lên BigQuery ở phần sau trong chương trình.

- Tham khảo [hướng dẫn từng bước](#) này.

### Bắt đầu với các cơ sở dữ liệu khác (nếu không sử dụng BigQuery)

Việc theo dõi các hoạt động của khóa học sẽ dễ dàng hơn nếu bạn sử dụng BigQuery, nhưng nếu bạn đang kết nối và thực hiện truy vấn SQL trên các nền tảng cơ sở dữ liệu khác thay vì BigQuery, có thể tham khảo các tài nguyên:

- [Bắt đầu với MySQL](#): Đây là hướng dẫn thiết lập và sử dụng MySQL.
- [Bắt đầu với Microsoft SQL Server](#): Đây là hướng dẫn để bắt đầu sử dụng SQL Server.
- [Bắt đầu với PostgreSQL](#): Đây là hướng dẫn để bắt đầu sử dụng PostgreSQL.
- [Bắt đầu với SQLite](#): Đây là hướng dẫn bắt đầu nhanh để sử dụng SQLite.

## 7. Hướng dẫn sử dụng SQL

Hướng dẫn này bao gồm: cách viết các truy vấn SQL, tạo nên các tài liệu, các ví dụ để minh họa. Đây là một nguồn tài nguyên hữu ích khi bạn đang sử dụng SQL; bạn có thể nhảy thẳng đến phần liên quan để xem lại các phần mình cần.

### Viết hoa và phân biệt chữ hoa chữ thường

Với SQL, viết hoa thường không quan trọng. Bạn có thể viết *SELECT* hoặc *select* hoặc *SeLeCT*. Tất cả đều đúng! Nhưng bạn nên sử dụng cách viết hoa thống nhất để các truy vấn trông chuyên nghiệp hơn.

Để viết các truy vấn SQL chuyên nghiệp, bạn nên viết hoa tất cả các chữ cho các phần mở đầu mệnh đề (ví dụ: *SELECT*, *FROM*, *WHERE*, v.v.). Các hàm cũng phải viết hoa toàn bộ (ví dụ: *SUM ()*). Tất cả các tên cột phải là chữ thường (tham khảo phần về *snake\_case* ở phần sau của hướng dẫn này). Tên bảng phải ở dạng CamelCase (tham khảo phần về CamelCase ở phần sau của hướng dẫn này). Điều này làm cho các truy vấn của bạn nhất quán và dễ đọc hơn trong khi không ảnh hưởng đến dữ liệu sẽ được lấy khi bạn chạy chúng. Trường hợp duy nhất viết hoa quan trọng là khi nó nằm trong dấu ngoặc kép (xem thêm về dấu ngoặc kép bên dưới).

Các nhà cung cấp cơ sở dữ liệu SQL có thể sử dụng các biến thể hơi khác của SQL. Các biến thể này được gọi là **phương ngữ SQL**. Một số phương ngữ SQL phân biệt chữ hoa chữ thường. BigQuery là một trong số đó. Vertica là một loại khác. Nhưng hầu hết, như



MySQL, PostgreSQL và SQL Server, không phân biệt chữ hoa chữ thường. Điều này có nghĩa nếu bạn tìm kiếm `country_code = 'us'`, nó sẽ trả về tất cả các dòng có "us", "uS", "Us" và "US". BigQuery không như vậy. BigQuery phân biệt chữ hoa chữ thường, vì vậy tìm kiếm tương tự sẽ chỉ trả về các mục nhập có `country_code` chính xác là 'us'. Nếu `country_code` là 'US', BigQuery sẽ không trả lại các dòng như một phần trong kết quả của bạn.

### Sử dụng nháy đơn hay nháy kép: " hay ""

Trong hầu hết các trường hợp, khi tham chiếu đến chuỗi, việc bạn sử dụng dấu nháy đơn " hay dấu nháy kép "" sẽ không ảnh hưởng. Ví dụ, `SELECT` dùng như bắt đầu mệnh đề. Nếu bạn đặt `SELECT` trong các dấu ngoặc kép như 'SELECT' hoặc "SELECT", thì SQL sẽ coi nó như một chuỗi văn bản. Truy vấn của bạn sẽ trả về lỗi vì truy vấn của bạn cần mệnh đề `SELECT`.

Nhưng có hai tình huống mà bạn sử dụng loại dấu ngoặc kép nào quan trọng:

- Khi bạn muốn các chuỗi được nhận dạng bằng bất kỳ phương ngữ SQL nào
- Khi chuỗi của bạn chứa dấu nháy đơn hoặc dấu ngoặc kép

Trong mỗi phương ngữ SQL có các quy tắc cho những gì được chấp nhận và những gì không. Một quy tắc chung trên hầu hết các phương ngữ SQL là sử dụng dấu nháy đơn cho chuỗi. Điều này loại bỏ rất nhiều nhầm lẫn. Vì vậy, nếu chúng ta muốn tham chiếu quốc gia US trong mệnh đề `WHERE` (ví dụ: `country_code = 'US'`), thì hãy sử dụng các dấu ngoặc kép xung quanh chuỗi 'US'.

Tình huống thứ hai là khi chuỗi của bạn có dấu nháy bên trong nó. Giả sử bạn có một cột món ăn yêu thích trong bảng có tên là Món ăn yêu thích và cột còn lại tương ứng với mỗi người bạn.

Friend	Favorite_food
Rachel DeSantos	Shepherd's pie
Sujin Lee	Tacos
Najil Okoro	Spanish paella

Bạn thấy rằng món ăn yêu thích của Rachel có dấu nháy '. Nếu bạn sử dụng dấu nháy đơn trong mệnh đề `WHERE` để tìm người bạn có món ăn yêu thích này, câu truy vấn như sau:

```
SELECT
    Friend
FROM
    FavoriteFoods
WHERE
    Favorite_food = 'Shepherd's pie'
```

**Cách viết này không hoạt động.** Nếu bạn chạy truy vấn này, bạn sẽ gặp lỗi. Điều này là do SQL nhận ra một chuỗi văn bản là một thứ gì đó bắt đầu bằng một nháy đơn ' và kết thúc bằng một nháy đơn khác '. Vì vậy, trong truy vấn sai ở trên, SQL nghĩ rằng Favorite\_food bạn đang tìm là 'Shepherd'. Chỉ là 'Shepherd' vì dấu nháy đơn trong Shepherd's kết thúc chuỗi.

Nói chung, đây sẽ trường hợp duy nhất bạn sử dụng dấu ngoặc kép thay vì dấu nháy đơn. Vì vậy, truy vấn của bạn sẽ trông giống như sau:

```
SELECT
    Friend
FROM
    FavoriteFoods
WHERE
    Favorite_food = "Shepherd's pie"
```

SQL hiểu chuỗi văn bản bắt đầu bằng dấu nháy đơn ' hoặc dấu nháy kép ". Vì chuỗi này bắt đầu bằng dấu ngoặc kép, SQL sẽ mong đợi một dấu nháy kép khác báo hiệu kết thúc chuỗi. Điều này giữ cho dấu nháy đơn được an toàn, vì vậy nó sẽ trả về "Shepherd's pie" chứ không phải "Shepherd".

### Viết comment trong SQL

Khi bạn quen thuộc với SQL hơn, bạn sẽ có thể đọc và hiểu các truy vấn nhanh chóng. Nhưng bạn vẫn nên có các comment trong câu truy vấn để nhắc nhở bản thân về những gì bạn đang thực hiện. Nếu bạn chia sẻ truy vấn của mình, nó cũng giúp người khác hiểu được ý định của mình.



Ví dụ:

```
--This is an important query used later to join with the accounts table
SELECT
    rowkey, --key used to join with account_id
    Info.date, --date is in string format YYYY-MM-DD HH:MM:SS
    Info.code --e.g., 'pub-###'
FROM
    Publishers
```

Bạn có thể sử dụng # thay cho hai dấu gạch ngang -- trong truy vấn trên nhưng hãy nhớ rằng # không được nhận ra trong tất cả các phương ngữ SQL (MySQL không nhận ra #). Vì vậy, tốt nhất là sử dụng -- và nhất quán với nó. Khi bạn thêm nhận xét vào truy vấn bằng cách sử dụng -- công cụ truy vấn cơ sở dữ liệu sẽ bỏ qua mọi thứ trong cùng một dòng với -- Nó sẽ tiếp tục xử lý truy vấn bắt đầu từ dòng tiếp theo.

### Kiểu Snake\_case cho tên cột

Điều quan trọng là luôn đảm bảo rằng đầu ra của truy vấn có tên dễ hiểu. Nếu bạn tạo một cột mới (giả sử từ một phép tính hoặc từ việc nối các trường mới), cột mới sẽ nhận được một tên mặc định chung (ví dụ: f0). Ví dụ:

```
SELECT
    SUM(tickets),
    COUNT(tickets),
    SUM(tickets) AS total_tickets,
    COUNT(tickets) AS number_of_purchases
FROM
    purchases
```

Kết quả sẽ là:

f0	f1	total_tickets	number_of_purchases
8	4	8	4

Hai cột đầu tiên được đặt tên là f0 và f1 vì chúng không được đặt tên trong truy vấn trên. SQL mặc định là f0, f1, f2, f3, v.v. Chúng ta đặt tên cho hai cột cuối cùng là total\_tickets và number\_of\_purchases để các tên cột này hiển thị trong kết quả truy vấn. Đây là lý do tại sao bạn nên đặt tên dễ gợi nhớ cho các cột của mình, đặc biệt là khi sử dụng các hàm.

Sau khi chạy truy vấn, bạn muốn có thể hiểu nhanh kết quả của mình, như hai cột cuối cùng mà chúng ta đã mô tả trong ví dụ.

Bạn có thể nhận thấy rằng tên cột có gạch dưới giữa các từ. Không bao giờ có khoảng trắng trong tên. Nếu 'total\_tickets' có khoảng trắng và nó sẽ trông như ' total tickets' thì SQL sẽ đổi SUM (tickets) thành 'total'. Do có khoảng trắng, SQL sẽ sử dụng 'total' làm tên và sẽ không hiểu ý bạn là 'tickets'. Vì vậy, khoảng trắng là xấu trong tên SQL. Không bao giờ sử dụng dấu cách trong tên.

Cách tốt nhất là sử dụng snake\_case. Điều này có nghĩa là 'total tickets', có khoảng cách giữa hai từ, phải được viết là 'total\_ticket' với dấu gạch dưới thay vì dấu cách.

### Kiểu CamelCase cho tên bảng

Bạn cũng có thể sử dụng cách viết hoa CamelCase khi đặt tên cho bảng của mình. Viết hoa CamelCase có nghĩa là bạn viết hoa đầu mỗi từ, giống như lạc đà hai bước. Vì vậy, bảng TicketsByOccasion sử dụng cách viết hoa CamelCase. Lưu ý rằng viết hoa từ đầu tiên trong CamelCase là tùy chọn; camelCase cũng được sử dụng. Một số người phân biệt giữa hai kiểu bằng cách gọi CamelCase, PascalCase và dành riêng camelCase khi từ đầu tiên không được viết hoa, giống như lạc đà một bước; ví dụ: ticketByOccasion.

Cuối cùng thì việc sử dụng CamelCase hay không là một sự lựa chọn phong cách. Có những cách khác để bạn có thể đặt tên cho bảng của mình, bao gồm:

- Tất cả chữ thường hoặc chữ hoa, chẳng hạn như ticketbyoccasion hoặc TICKETSBYOCCASION
- Với snake\_case, chẳng hạn như ticket\_by\_occasion

Lưu ý rằng cách dùng tất cả các chữ cái viết thường hoặc viết hoa có thể gây khó khăn cho việc đọc tên bảng của bạn, vì vậy, tùy chọn này không được khuyến khích sử dụng.

Tùy chọn thứ hai, solid\_case, về mặt kỹ thuật là ổn. Với các từ được phân tách bằng dấu gạch dưới, tên bảng của bạn rất dễ đọc, nhưng nó có thể rất dài vì bạn đang thêm dấu gạch dưới. Nó cũng cần nhiều thời gian hơn để viết. Nếu bạn sử dụng tên bảng này nhiều, nó có thể trở thành một việc phiền phức.

Tóm lại, việc sử dụng `solid_case` hoặc `CamelCase` cho tên bảng là tùy thuộc vào bạn. Chỉ cần đảm bảo tên bảng của bạn dễ đọc và nhất quán. Ngoài ra, hãy nhớ tìm hiểu xem công ty của bạn có cách đặt tên bảng của họ hay không. Nếu có, hãy luôn tuân theo quy ước đặt tên của họ để có sự nhất quán.

### Thụt vào đầu dòng

Theo nguyên tắc chung, bạn muốn giữ độ dài của mỗi dòng trong truy vấn  $\leq 100$  ký tự. Điều này làm cho các truy vấn của bạn dễ đọc. Ví dụ: kiểm tra truy vấn này với một dòng có  $>100$  ký tự:

```
SELECT
CASE WHEN genre = 'horror' THEN 'Will not watch' WHEN genre = 'documentary'
THEN 'Will watch alone' ELSE 'Watch with others' END AS
Watch_category, COUNT(movie_title) AS number_of_movies
FROM
    MovieTheater
GROUP BY
    1
```

Truy vấn này khó đọc và khó khắc phục sự cố hoặc chỉnh sửa. Sau đây, là một truy vấn mà chúng ta dùng quy tắc  $\leq 100$  ký tự:

```
SELECT
    CASE
        WHEN genre = 'horror' THEN 'Will not watch'
        WHEN genre = 'documentary' THEN 'Will watch alone'
        ELSE 'Watch with others'
        END AS watch_category, COUNT(movie_title) AS number_of_movies
FROM
    MovieTheater
GROUP BY
    1
```

Với cách viết này, mệnh đề `SELECT` sẽ dễ hiểu hơn nhiều. Dù cho cả hai cách viết trên sẽ chạy mà không có vấn đề gì vì thụt lề không quan trọng trong SQL. Nhưng thụt lề thích hợp vẫn quan trọng để giữ cho các dòng ngắn gọn. Và nó sẽ được đánh giá cao bởi bất kỳ ai đọc truy vấn của bạn, bao gồm cả chính bạn!

## Comment nhiều dòng

Nếu bạn đưa ra nhận xét chiếm nhiều dòng, bạn có thể sử dụng `--` cho mỗi dòng. Hoặc, nếu bạn có nhiều hơn hai dòng nhận xét, để rõ ràng và dễ dàng hơn, bạn có thể sử dụng `/*` để bắt đầu nhận xét và `*/` để đóng nhận xét. Ví dụ, bạn có thể sử dụng phương pháp -- như bên dưới:

```
-- Date: September 15, 2020
-- Analyst: Jazmin Cisneros
-- Goal: Count the number of rows in the table
SELECT
    COUNT(*) number_of_rows -- the * stands for all so count all
FROM
    table
```

Hay dùng phương pháp `/* */`

```
/*
Date: September 15, 2020
Analyst: Jazmin Cisneros
Goal: Count the number of rows in the table
*/
SELECT
    COUNT(*) AS number_of_rows -- the * stands for all so count all
FROM
    table
```

Trong SQL, bạn sử dụng phương pháp nào không quan trọng. SQL bỏ qua các nhận xét bất kể bạn sử dụng gì: `#`, `--`, hoặc `/* */`. Vì vậy, nó phụ thuộc vào sở thích cá nhân của bạn. Phương thức `/*` và `*/` cho nhận xét nhiều dòng thường trông gọn gàng hơn và giúp tách các nhận xét khỏi truy vấn. Nhưng không có một phương pháp đúng hay sai.

## Trình xử lý văn bản cho SQL

Khi bạn tham gia vào một công ty, bạn có thể mong đợi mỗi công ty sử dụng nền tảng SQL và phương ngữ SQL của riêng họ. Nền tảng SQL mà họ sử dụng (ví dụ: BigQuery, MySQL hoặc SQL Server) là nơi bạn sẽ viết và chạy các truy vấn SQL của mình. Nhưng hãy nhớ rằng không phải tất cả các nền tảng SQL đều cung cấp trình soạn thảo văn bản để viết mã SQL. Các trình soạn thảo văn bản SQL cung cấp cho bạn một giao diện nơi bạn có thể viết các truy vấn SQL của mình theo cách dễ dàng hơn và được mã hóa bằng màu

sắc. Trên thực tế, tất cả mã mà chúng ta đang làm việc cho đến nay đều được viết bằng trình soạn thảo văn bản SQL!

## Ví dụ với Sublime Text

Nếu nền tảng SQL của bạn không có mã hóa màu, bạn có thể muốn nghĩ đến việc sử dụng trình soạn thảo văn bản như [Sublime Text](#) hoặc [Atom](#). Phần này cho biết cách SQL được hiển thị trong Sublime Text. Đây là một truy vấn trong Sublime Text:

```
1 SELECT
2     column_name
3 FROM
4     table
5 WHERE
6     condition = 'match'
```

Với Sublime Text, bạn cũng có thể thực hiện chỉnh sửa nâng cao như xóa thụt lề trên nhiều dòng cùng một lúc. Ví dụ: giả sử truy vấn của bạn bằng cách nào đó có thụt lề sai vị trí và trông như thế này:

```
1      SELECT
2      column_name
3      FROM
4      table
5      WHERE
6      condition = 'match'
```

Điều này thực sự khó đọc, vì vậy bạn muốn loại bỏ những thực thể đó và bắt đầu lại. Trong nền tảng SQL thông thường, bạn sẽ phải đi vào từng dòng và nhấn BACKSPACE để xóa từng thực thể trên mỗi dòng. Nhưng trong Sublime, bạn có thể loại bỏ tất cả các thực thể cùng một lúc bằng cách chọn tất cả các dòng và nhấn **Command** (hoặc **CTRL** trong Windows) + **[**. Điều này giúp loại bỏ thực thể từ mọi dòng. Sau đó, bạn có thể chọn các dòng mà bạn muốn thực thể (tức là các dòng 2, 4 và 6) bằng cách nhấn phím Command (hoặc phím CTRL trong Windows) và chọn các dòng đó. Sau đó, trong khi vẫn giữ phím

Command (hoặc phím CTRL trong Windows), hãy nhấn ] để thực hiện dòng 2, 4 và 6 cùng một lúc. Thao tác này sẽ xóa truy vấn của bạn và làm cho nó trông giống như sau:

```
1  SELECT
2  |   column_name
3  FROM
4  |   table
5  WHERE
6  |   condition = 'match'
```

Sublime Text cũng hỗ trợ các biểu thức chính quy. **Biểu thức chính quy** (hoặc **regex**) có thể được sử dụng để tìm kiếm và thay thế các mẫu chuỗi trong truy vấn. Chúng ta sẽ không đề cập đến biểu thức chính quy ở đây, nhưng bạn có thể muốn tự mình tìm hiểu thêm về chúng vì chúng là một công cụ rất mạnh mẽ.

Bạn có thể bắt đầu với các tài nguyên sau:

- [Tìm kiếm và thay thế trong Sublime Text](#)
- [Hướng dẫn về Regex](#) (nếu bạn không biết cụm từ thông dụng là gì)
- [Regex cheat sheet](#)

## 8. Tài liệu tham khảo

[1]<https://www.coursera.org/learn/data-preparation/supplement/uXqEX/databases-in-data-analytics>

[2]<https://www.coursera.org/learn/data-preparation/supplement/OFIHG/inspecting-a-data-set-a-guided-hands-on-tour>

[3]<https://www.coursera.org/learn/data-preparation/supplement/mdF9p/metadata-is-as-important-as-the-data-itself>

[4]<https://www.coursera.org/learn/data-preparation/supplement/esVz6/from-external-source-to-a-spreadsheet>

[5]<https://www.coursera.org/learn/data-preparation/supplement/8yrhM/exploring-public-datasets>

[6]<https://www.coursera.org/learn/data-preparation/supplement/DYOQK/using-bigquery>

[7]<https://www.coursera.org/learn/data-preparation/supplement/gLnvK/in-depth-guide-sql-best-practices>



## Bài đọc 4: Tổ chức và bảo vệ dữ liệu của bạn

### 1. Hướng dẫn tổ chức dữ liệu

#### Hướng dẫn cho quy ước đặt tên file

Hãy sớm thảo luận và thống nhất các quy ước đặt tên file trong một dự án để tránh đổi tên file nhiều lần.

Thống nhất việc đặt tên file của bạn với các quy ước đặt tên file hiện có của nhóm hoặc công ty bạn.

Đảm bảo rằng tên file của bạn có ý nghĩa; cân nhắc bao gồm thông tin như tên dự án và bất kỳ thông tin nào khác sẽ giúp bạn nhanh chóng xác định (và sử dụng) file cho đúng mục đích.

Bao gồm ngày và số phiên bản trong tên file; các định dạng phổ biến là YYYYMMDD cho ngày tháng và v ## cho các phiên bản (hoặc bản sửa đổi).

Tạo file văn bản dưới dạng file mẫu có nội dung mô tả quy ước đặt tên file và tên file áp dụng quy ước đó.

Tránh khoảng trắng và ký tự đặc biệt trong tên file. Thay vào đó, hãy sử dụng dấu gạch ngang, dấu gạch dưới hoặc chữ in hoa. Dấu cách và các ký tự đặc biệt có thể gây ra lỗi trong một số ứng dụng.

#### Một số gợi ý về cách tổ chức file

Hãy nhớ những mẹo sau để luôn tổ chức file cho tốt khi bạn làm việc với chúng:

- Tạo các thư mục và thư mục con trong một hệ thống phân cấp hợp lý để các file liên quan được lưu trữ cùng nhau.
- Tách biệt công việc đang diễn ra với công việc đã hoàn thành để các file dự án hiện tại của bạn dễ tìm hơn. Lưu trữ (archive) các file cũ hơn trong một thư mục riêng biệt hoặc ở vị trí lưu trữ bên ngoài.
- Nếu các file của bạn không được sao lưu tự động, hãy thường xuyên sao lưu chúng theo cách thủ công để tránh mất công việc quan trọng.

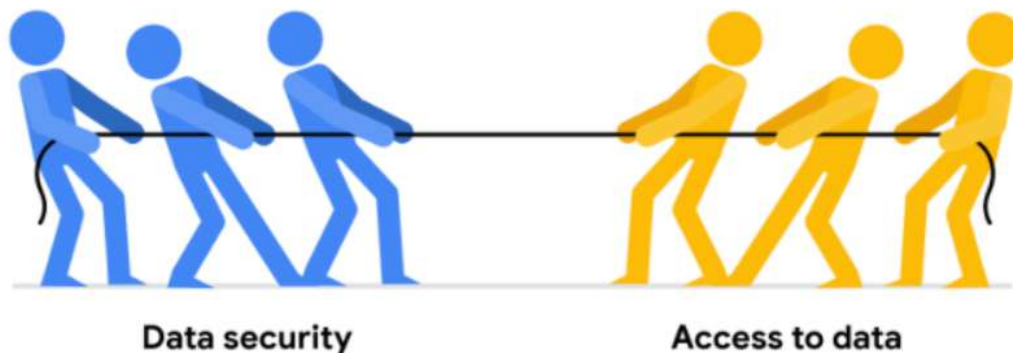
### 2. Bảo mật dữ liệu

#### Cuộc chiến giữa bảo mật và phân tích dữ liệu

**Bảo mật dữ liệu** nghĩa là bảo vệ dữ liệu khỏi bị truy cập trái phép hoặc phá hủy bằng cách áp dụng các biện pháp an toàn. Thông thường, mục đích của bảo mật dữ liệu là ngăn người dùng truy cập trái phép hoặc xem dữ liệu nhạy cảm. Các nhà phân tích dữ liệu phải tìm cách cân bằng giữa bảo mật dữ liệu với nhu cầu phân tích thực tế của họ. Điều này có

thể phức tạp - chúng ta muốn giữ cho dữ liệu của mình an toàn và bảo mật, nhưng chúng ta cũng muốn sử dụng nó càng sớm càng tốt để có thể thực hiện các quan sát kịp thời và có ý nghĩa.

Để làm được điều này, các công ty cần tìm cách cân bằng các biện pháp bảo mật dữ liệu với nhu cầu truy cập dữ liệu của họ.



May mắn thay, có một số biện pháp bảo mật có thể giúp các công ty thực hiện điều đó. Hai điều chúng ta sẽ nói ở đây là mã hóa (encryption) và token hóa (tokenization).

**Mã hóa** sử dụng một thuật toán đặc biệt để thay đổi dữ liệu và khiến dữ liệu đó không thể sử dụng được bởi những người dùng và ứng dụng không biết thuật toán. Thuật toán này được lưu dưới dạng “khóa” có thể được sử dụng để đảo ngược mã hóa; vì vậy nếu bạn có khóa, bạn vẫn có thể sử dụng dữ liệu ở dạng ban đầu.

**Token hóa** thay thế các phần tử dữ liệu bạn muốn bảo vệ bằng dữ liệu được tạo ngẫu nhiên được gọi là “token”. Dữ liệu gốc được lưu trữ ở một vị trí riêng biệt và được ánh xạ tới các token. Để truy cập vào dữ liệu gốc hoàn chỉnh, người dùng hoặc ứng dụng cần có quyền sử dụng dữ liệu được mã hóa và ánh xạ token. Điều này có nghĩa là ngay cả khi dữ liệu được mã hóa bị tấn công, dữ liệu gốc vẫn an toàn và bảo mật ở một vị trí riêng biệt.

Mã hóa và token hóa chỉ là một số lựa chọn bảo mật dữ liệu hiện có. Có rất nhiều cách khác, như sử dụng các thiết bị xác thực với công nghệ AI.

Là một nhà phân tích dữ liệu cơ sở, bạn có thể không chịu trách nhiệm xây dựng các hệ thống này. Rất nhiều công ty có nhóm chuyên về bảo mật dữ liệu hoặc thuê các công ty



thứ ba chuyên về bảo mật dữ liệu để tạo ra các hệ thống này. Nhưng điều quan trọng cần biết là tất cả các công ty có trách nhiệm bảo mật dữ liệu của họ và hiểu một số hệ thống tiềm năng mà công ty tương lai của bạn có thể sử dụng.

### 3. Tài liệu tham khảo

[1]<https://www.coursera.org/learn/data-preparation/supplement/fLKJl/organization-guidelines>

[2]<https://www.coursera.org/learn/data-preparation/supplement/CmRzN/balancing-security-and-analytics>

## Bài đọc 5: Tham gia vào cộng đồng dữ liệu

### 1. Tạo hoặc tăng cường sự hiện diện trực tuyến

#### Đăng ký

Đăng ký với LinkedIn rất đơn giản. Chỉ cần làm theo các bước sau:

1. Truy cập đến [linkedin.com](https://www.linkedin.com)
2. Nhấp vào **Join Now** hay **Join with resume**

Nếu bạn nhấp vào **Join Now**:

1. Nhập địa chỉ email và mật khẩu của bạn và nhấp vào **Agree & Join** (hoặc nhấp vào **Join with google** để liên kết với tài khoản Google).
2. Nhập họ và tên của bạn và nhấp vào **Continue**.
3. Nhập quốc gia/vùng của bạn, mã bưu điện và vị trí của bạn cùng với khu vực (điều này giúp LinkedIn tìm thấy các cơ hội việc làm gần bạn).
4. Nhập chức danh gần đây nhất của bạn hoặc chọn **I'm a student**.
5. Nếu bạn đã nhập chức danh công việc gần đây nhất của mình, hãy chọn loại công việc và nhập tên công ty gần đây nhất của bạn.
6. Nếu bạn chọn làm việc tự do, LinkedIn sẽ hỏi về ngành của bạn.
7. Bấm xác nhận địa chỉ email của bạn. Bạn sẽ nhận được một email từ LinkedIn.
8. Để xác nhận địa chỉ email của bạn, hãy nhấp vào **Agree & Confirm** trong email của bạn.
9. LinkedIn sau đó sẽ hỏi bạn có đang tìm việc hay không. Nhấp vào câu trả lời phù hợp. Nếu bạn chọn Có, LinkedIn sẽ giúp bạn bắt đầu tìm kiếm cơ hội việc làm.

Nếu bạn nhấp vào **Join with resume**:

1. Nhấp vào **Upload your resume** và chọn file để tải lên.
2. Làm theo bất kỳ bước nào trong **Join now** nếu cảm thấy phù hợp.

Tùy chọn **Join with resume** giúp bạn tiết kiệm thời gian vì nó tự động điền hầu hết thông tin từ sơ yếu lý lịch của bạn. Và chỉ như vậy, hồ sơ ban đầu của bạn đã sẵn sàng!

## Thêm thông tin cơ bản trong hồ sơ của bạn

Bạn nên dành thời gian điền vào mọi phần trong hồ sơ của mình. Điều này giúp các nhà tuyển dụng tìm thấy hồ sơ của bạn và giúp những người bạn kết nối hiểu rõ hơn về bạn. Bắt đầu với ảnh của bạn. Dưới đây là một số mẹo để giúp bạn chọn một bức ảnh tuyệt vời cho hồ sơ mới của mình:

- Chọn một hình ảnh giống bạn: Bạn muốn đảm bảo rằng hồ sơ của bạn là hình ảnh đại diện tốt nhất về bạn và bao gồm cả ảnh của bạn. Bạn muốn một kết nối tiềm năng hoặc nhà tuyển dụng tiềm năng có thể nhận ra bạn từ ảnh hồ sơ nếu bạn đến gặp.
- Sử dụng ngành của bạn làm ví dụ: Nếu bạn gặp khó khăn trong việc quyết định chọn ảnh đại diện như thế nào cho phù hợp, hãy xem các hồ sơ khác trong cùng ngành hoặc từ các công ty bạn quan tâm để hiểu rõ hơn về những gì bạn nên làm.
- Chọn hình ảnh có độ phân giải cao: Độ phân giải càng tốt thì càng tạo ấn tượng tốt hơn, vì vậy hãy đảm bảo hình ảnh bạn chọn không bị mờ. Kích thước hình ảnh lý tưởng cho ảnh hồ sơ LinkedIn là 400 x 400 pixel. Sử dụng ảnh mà khuôn mặt của bạn chiếm ít nhất 60% không gian trong khung.
- Nhớ mỉm cười: Ảnh hồ sơ của bạn là một bức ảnh chụp nhanh con người của bạn, vì vậy bạn có thể nghiêm túc trong bức ảnh của mình. Nhưng mỉm cười giúp tạo sự thoải mái cho các mối quan hệ tiềm năng và các nhà tuyển dụng tiềm năng.



## Thêm kết nối

Kết nối là một cách tuyệt vời để cập nhật thông tin với đồng nghiệp cũ, đồng nghiệp hiện tại, bạn học hoặc thậm chí các công ty bạn muốn làm việc cùng. Thế giới là một nơi rộng lớn với rất nhiều người. Vì vậy, đây là một số mẹo để giúp bạn bắt đầu.

1. Kết nối với những người bạn biết ngoài đời thật.
2. Thêm một liên lạc cá nhân vào tin nhắn mời của bạn. Thay vì chỉ cho họ biết bạn muốn kết nối, hãy cho họ biết lý do.
3. Đảm bảo rằng ảnh hồ sơ của bạn là hiện tại để mọi người có thể nhận ra bạn.
4. Thêm giá trị. Cung cấp cho họ tài nguyên, liên kết trang web hoặc thậm chí một số nội dung mà họ có thể thấy thú vị trong lời mời kết nối của bạn.

## Tìm kiếm nhà lãnh đạo và người có ảnh hưởng

LinkedIn là một nơi tuyệt vời để tìm những người tuyệt vời và những ý tưởng tuyệt vời. Từ công nghệ đến tiếp thị và mọi thứ liên quan, có đủ loại người có ảnh hưởng và nhà lãnh đạo tư tưởng đang hoạt động trên LinkedIn. Nếu bạn muốn biết suy nghĩ của một số bộ óc có ảnh hưởng và được kính trọng nhất trong một lĩnh vực nào đó, thì LinkedIn là một nơi tuyệt vời để bắt đầu. Theo dõi những người bạn yêu thích chỉ mất vài phút. Bạn có thể tìm kiếm người hoặc công ty riêng lẻ hoặc bạn có thể sử dụng các danh sách này làm điểm xuất phát.

[Những người có ảnh hưởng hàng đầu trên LinkedIn](#)  
[LinkedIn Top Voices 2020: Data Science & AI](#)

## Tìm kiếm một vị trí mới

Trên LinkedIn, việc cho nhà tuyển dụng tiềm năng biết rằng bạn đang sẵn sàng cho một công việc mới thật đơn giản. Chỉ cần làm theo các bước sau:

1. Nhấp vào biểu tượng **Me** ở đầu trang chủ LinkedIn của bạn.
2. Nhấp vào **View profile**.
3. Nhấp vào menu thả xuống của phần **Add profile** và trong Intro, hãy chọn **Looking for a new job**.

Chọn các bộ lọc thích hợp cho các vị trí mới mà bạn đang tìm kiếm và cập nhật hồ sơ của để phù hợp hơn với vai trò mà bạn đang ứng tuyển.

### Luôn cập nhật hồ sơ của bạn

Thêm thông tin vào hồ sơ của bạn để giữ cho hồ sơ luôn hoàn chỉnh, cập nhật và thú vị. Ví dụ: hãy nhớ thêm Chứng chỉ Google Data Analytics vào hồ sơ của bạn sau khi bạn hoàn thành chương trình!

## 2. Xây dựng mạng phân tích dữ liệu

### Sử dụng LinkedIn để kết nối

Một **kết nối** là người mà bạn biết và tin tưởng trên cơ sở cá nhân hoặc nghề nghiệp. Kết nối của bạn là những người tạo nên mạng lưới của bạn. Khi nói đến mạng lưới của bạn, điều quan trọng phải nhớ là chất lượng hơn số lượng. Vì vậy, đừng tập trung vào việc bạn có bao nhiêu kết nối. Thay vào đó, hãy đảm bảo rằng mọi người mà bạn kết nối đều tăng thêm giá trị cho mạng của bạn và ngược lại.

### Mời những người bạn biết chứ không phải chỉ đưa ra những yêu cầu lạnh lùng

Thêm kết nối trên LinkedIn thật dễ dàng. Bạn mời mọi người tham gia mạng của mình và họ chấp nhận lời mời của bạn. Khi bạn gửi lời mời, bạn có thể đính kèm ghi chú cá nhân. Ghi chú cá nhân rất được khuyến khích.

Một cách tuyệt vời để tăng số lượng kết nối của bạn là mời bạn cùng lớp, bạn bè, giáo viên hoặc thậm chí là thành viên của câu lạc bộ hoặc tổ chức mà bạn đang tham gia. LinkedIn cũng đưa ra các đề xuất về kết nối dựa trên thông tin hồ sơ của bạn. Đây là một ví dụ (mẫu) mà bạn có thể sử dụng để kết nối với đồng nghiệp cũ:

**Dan Tyre**

You worked with Dan in the same group

**Hi <fill in name here>,**

Please accept my invitation to connect. It has been a while since we were at <fill in company name here> and I look forward to catching up with you. I'm looking for job opportunities and would love to hear about what you're doing and who is hiring in your organization.

**Best regards,**

**<fill in your name here>**

**Yêu cầu lạnh lùng** trên LinkedIn là những lời mời kết nối với những người mà bạn không biết về mặt cá nhân hay nghề nghiệp. Khi bạn bắt đầu xây dựng mạng lưới của mình, cách tốt nhất là kết nối với những người bạn đã biết. Tuy nhiên, những yêu cầu lạnh lùng có thể là cách duy nhất để kết nối với những người làm việc tại các công ty mà bạn quan tâm. Bạn có thể tìm hiểu nhiều về văn hóa của công ty và cơ hội việc làm từ những nhân viên hiện tại. Cách tốt nhất, hãy hạn chế việc gửi các yêu cầu lạnh lùng và chỉ khi không còn cách nào khác để kết nối.

### Yêu cầu thư giới thiệu (tham khảo)

**Thư giới thiệu** trên LinkedIn là một cách tuyệt vời để người khác xác nhận cho bạn. Yêu cầu mọi người nhận xét về hiệu suất làm việc trước đây của bạn, cách bạn xử lý một dự án đầy thách thức hoặc điểm mạnh của bạn với tư cách là một nhà phân tích dữ liệu. Bạn có thể chọn chấp nhận, từ chối, hiển thị hoặc ẩn các đề xuất trong hồ sơ của mình.

Dưới đây là một số mẹo để giới thiệu:

- Tiếp cận với nhiều người để có cái nhìn 360 độ: người giám sát, đồng nghiệp, báo cáo trực tiếp, đối tác và khách hàng

- Cá nhân hóa yêu cầu giới thiệu bằng một thông báo tùy chỉnh
- Đề xuất những điểm mạnh và khả năng mà họ có thể làm nổi bật như một phần của yêu cầu
- Sẵn sàng viết thư giới thiệu để đáp lại
- Đọc kỹ thư giới thiệu trước khi bạn chấp nhận nó vào hồ sơ của mình

Đôi khi, phần khó nhất của việc nhận được thư giới thiệu là tạo ra thông điệp yêu cầu phù hợp. Dưới đây là một ví dụ (mẫu) mà bạn có thể sử dụng để yêu cầu thư giới thiệu:

### Ask Dan to recommend you

Include a personalized message with your request



**Dan Tyre**

You worked with Dan in the same group

**Hi <fill in name here>,**

How are you? I hope you are well. I'm preparing for a new job search and would appreciate it if you could write a recommendation that highlights my *<insert your specific skill here>*. Our experience working on *<insert project here>* is a great example and I would be happy to provide other examples if you need them. Please let me know if I can write a recommendation for you. I would be very glad to return the favor.

**Thanks in advance for your support!**

**<fill in your name here>**

Hỏi một vài mối quan hệ để giới thiệu bạn và nêu rõ lý do tại sao bạn nên được thuê. Các đề xuất giúp các nhà tuyển dụng tiềm năng hiểu rõ hơn về bạn là ai và chất lượng công việc của bạn.

## Tổng kết

Khi bạn viết những bài đăng có ý nghĩa và trả lời người khác một cách chân thành, những người trong và thậm chí bên ngoài mạng lưới của bạn sẽ cởi mở và sẵn sàng giúp đỡ bạn trong quá trình tìm kiếm việc làm.

## 3. Phát triển mạng lưới

Trong bài đọc này, bạn sẽ được giới thiệu các cơ hội trực tuyến và trực tiếp để kết nối với các nhà phân tích dữ liệu khác. Đây là một phần trong cách bạn phát triển các mối quan hệ nghề nghiệp, điều này rất quan trọng khi bạn mới bắt đầu sự nghiệp của mình.

### Kết nối trực tuyến

Nếu bạn dành vài giờ trên mạng xã hội mỗi ngày, bạn có thể hoàn toàn thoải mái khi kết nối trực tuyến với các nhà phân tích dữ liệu khác. Tuy nhiên, bạn nên xem ở đâu nếu bạn không biết bất kỳ nhà phân tích dữ liệu nào?

Ngay cả khi bạn không sử dụng mạng xã hội và mới tạo hồ sơ LinkedIn ngày hôm qua, bạn vẫn có thể sử dụng sự hiện diện trực tuyến của mình để tìm và kết nối với các nhà phân tích dữ liệu khác.

Biết nơi để tìm là chìa khóa. Dưới đây là một số gợi ý về nơi bắt đầu trực tuyến:

- **Đăng ký** các bản tin như [Data Elixir](#). Điều này không chỉ cung cấp cho bạn kho tàng thông tin hữu ích một cách thường xuyên mà bạn còn học được tên của các chuyên gia khoa học dữ liệu có thể theo dõi, hoặc thậm chí có thể kết nối nếu bạn có lý do chính đáng.
- **Hackathons** (các cuộc thi) như cuộc thi được tài trợ bởi [Kaggle](#), một trong những cộng đồng khoa học dữ liệu và máy học lớn nhất trên thế giới. Tham gia hackathon có thể không dành cho tất cả mọi người. Nhưng sau khi tham gia cộng đồng, bạn thường có quyền truy cập vào các diễn đàn nơi bạn có thể trò chuyện và kết nối với các nhà phân tích dữ liệu khác.



- **Các buổi gặp mặt**, hoặc các cuộc họp trực tuyến thường mang tính địa phương đối với khu vực địa lý của bạn. Nhập tìm kiếm cho gặp gỡ khoa học dữ liệu gần tôi' để xem bạn nhận được kết quả nào. Thường có một lịch trình được đăng cho các cuộc họp sắp tới để bạn có thể tham dự và gặp gỡ các nhà phân tích dữ liệu khác. Tìm hiểu thêm thông tin về [các cuộc gặp gỡ diễn ra trên khắp thế giới](#).
- **Các nền tảng** như LinkedIn và Twitter. Sử dụng tìm kiếm trên một trong hai nền tảng để tìm các thẻ “data science” hay “data analysis” để theo dõi. Bạn cũng có thể đăng các câu hỏi hoặc bài viết của riêng mình để tạo phản hồi và xây dựng kết nối theo cách đó. Tại thời điểm viết bài này, hashtag #dataanalyst trên LinkedIn có 11.842 người theo dõi, hashtag #dataanalytics có 98.412 người theo dõi và hashtag #datascience có 746.945 người theo dõi. Nhiều hashtag giống nhau hoạt động trên Twitter và thậm chí trên Instagram.
- **Hội thảo trên web** có thể giới thiệu một loạt các diễn giả và thường được ghi lại để thuận tiện cho việc truy cập và phát lại. Bạn có thể xem ai có mặt trong bảng hội thảo trên web và theo dõi họ. Thêm vào đó, rất nhiều hội thảo trên web miễn phí. Một lựa chọn thú vị là loạt [hội thảo trên web Tableau](#). Tìm hiểu cách Tableau đã sử dụng Tableau trong các bộ phận nội bộ của nó.

### Các cuộc họp mặt trực tiếp (ngoại tuyến)

Các cuộc họp mặt trực tiếp rất có giá trị trong một thế giới số hóa. Chúng là một cách tuyệt vời để gặp gỡ mọi người. Rất nhiều mối quan hệ trực tuyến bắt đầu từ những cuộc gặp gỡ trực tiếp và được tiếp tục sau khi mọi người trở về nhà. Nhiều tổ chức tài trợ cho các cuộc gặp gỡ thường niên cũng cung cấp tài nguyên cho các cuộc họp ảo trong suốt thời gian còn lại của năm.

Dưới đây là một số gợi ý để tìm các cuộc họp mặt trực tiếp trong khu vực của bạn:

- **Các hội nghị** thường trình bày các ý tưởng và chủ đề mới lạ. Chi phí của các hội nghị khác nhau, và một số rất đắt. Nhưng rất nhiều hội nghị giảm giá cho sinh viên và một số hội nghị như [Women in Analytics](#) nhằm mục đích tăng số lượng các nhóm ít người trong lĩnh vực này. Các công ty tư vấn và nghiên cứu hàng đầu như [Gartner](#) cũng tài trợ cho các hội nghị về dữ liệu và phân tích. Danh sách các cuộc họp và [sự kiện trực tuyến](#)

của [KDNuggets](#) dành cho AI, phân tích, dữ liệu lớn, khoa học dữ liệu và học máy rất hữu ích.

- **Các hiệp hội** tập hợp các thành viên để thúc đẩy một lĩnh vực như khoa học dữ liệu. [Hiệp hội phân tích kỹ thuật số](#). [Danh sách KDNuggets](#) gồm các xã hội và nhóm để phân tích, khai thác dữ liệu, khoa học dữ liệu và khám phá kiến thức rất hữu ích.
- **Các cộng đồng người dùng** và **hội nghị thượng đỉnh** cung cấp các sự kiện cho người dùng các công cụ phân tích dữ liệu; đây là cơ hội để học hỏi từ những người giỏi nhất. Bạn đã thấy [cộng đồng Tableau](#) chưa?
- **Các tổ chức phi lợi nhuận** thúc đẩy việc sử dụng khoa học dữ liệu một cách có đạo đức và có thể tổ chức các sự kiện vì sự phát triển nghề nghiệp của các thành viên của họ. [Hiệp hội Khoa học Dữ liệu](#) là một ví dụ.

### Những điều quan trọng

Kết nối của bạn sẽ giúp bạn nâng cao kiến thức và kỹ năng của mình. Tạo và giữ kết nối cũng rất quan trọng đối với những người đã làm việc trong lĩnh vực phân tích dữ liệu. Vì vậy, hãy tìm kiếm các cộng đồng trực tuyến quảng bá các công cụ phân tích dữ liệu hoặc khoa học dữ liệu tiên tiến. Và nếu có tại nơi bạn sống, hãy tìm kiếm các buổi gặp mặt để kết nối trực tiếp với nhiều người hơn. Tận dụng lợi thế của cả hai con đường để có được điều tốt nhất của cả hai! Trò chuyện và trao đổi thông tin trực tiếp sẽ dễ dàng hơn, nhưng lợi thế chính của các kết nối trực tuyến là chúng không giới hạn ở nơi bạn sống. Các cộng đồng trực tuyến thậm chí có thể kết nối bạn với một cộng đồng quốc tế.

## 4. Tài liệu tham khảo

[1]<https://www.coursera.org/learn/data-preparation/supplement/3QDa4/getting-started-with-linkedin>

[2]<https://www.coursera.org/learn/data-preparation/supplement/N3NOB/building-connections-on-linkedin>

[3]<https://www.coursera.org/learn/data-preparation/supplement/bn5w2/developing-a-net-work>

# Phần 2

## HƯỚNG DẪN

## TRẢ LỜI CÂU HỎI

## Câu hỏi về thu thập dữ liệu

1. Phương pháp thu thập dữ liệu nào là phổ biến nhất sử dụng bởi các nhà khoa học.

- A. Phỏng vấn (Interview)
- B. Khảo sát (survey)
- C. Bảng câu hỏi (questionnaire)
- D. Quan sát (observation)

Đáp án: D

(Quan sát là phương pháp thu thập dữ liệu thường được các nhà khoa học sử dụng nhất)

2. Các tổ chức như Trung tâm Kiểm soát Dịch bệnh Hoa Kỳ (CDC) thường sử dụng dữ liệu thu thập được từ các bệnh viện. Loại dữ liệu nào mà CDC sử dụng nếu nó được các bệnh viện thu thập, sau đó bán cho CDC để phân tích? Điều khiển dữ liệu

- A. Dữ liệu của bên thứ nhất
- B. Dữ liệu của bên thứ hai
- C. Dữ liệu của bên thứ ba
- D. Dữ liệu của nhiều bên

Đáp án: B

(Dữ liệu do các bệnh viện thu thập, sau đó được CDC sử dụng, là một ví dụ về dữ liệu của bên thứ hai)

3. Điền vào chỗ trống: Trong phân tích dữ liệu, \_\_\_\_\_ để cập đến tất cả các giá trị dữ liệu có thể có trong một tập dữ liệu nhất định.

- A. Tập hợp (population)
- B. Vật mẫu (sample)
- C. Đại diện (representation)
- D. Nguồn (source)

Đáp án: A

(Trong phân tích dữ liệu, một tập hợp đề cập đến tất cả các giá trị dữ liệu có thể có trong một tập dữ liệu nhất định)

## Câu hỏi về định dạng và cấu trúc dữ liệu

1. Điền vào chỗ trống: Thời lượng phim là ví dụ về dữ liệu \_\_\_\_\_.

- A. Rời rạc
- B. Định tính
- C. Danh nghĩa
- D. Liên tục

Đáp án: D

2. Đặc điểm của dữ liệu phi cấu trúc là gì? Chọn tất cả những câu phù hợp.

- A. Không có trật tự
- B. Có thể có cấu trúc bên trong
- C. Vừa khít với các hàng và cột
- D. Có một cấu trúc để nhận biết

Đáp án: A, B

3. Dữ liệu có cấu trúc cho phép dữ liệu được nhóm lại với nhau để tạo thành các quan hệ. Điều này giúp các nhà phân tích dễ dàng làm gì với dữ liệu? Chọn tất cả những câu phù hợp.

- A. Viết lại
- B. Lưu trữ
- C. Tìm kiếm
- D. Phân tích

Đáp án: B, C, D

4. Ví dụ nào sau đây là dữ liệu phi cấu trúc?

A.Email

B.Vị trí GPS

C.Liên hệ được lưu trên điện thoại

D.Điểm đánh giá về một nhà hàng địa phương được yêu thích

Đáp án: A

### Câu hỏi về kiểu dữ liệu, trường và giá trị

1. Điền vào chỗ trống: Các công cụ tìm kiếm trên Internet là một ví dụ về cách các toán tử Boolean được sử dụng. Toán tử Boolean \_\_\_\_\_ mở rộng số lượng kết quả khi được sử dụng trong tìm kiếm từ khóa.

A.AND

B.OR

C.NOT

D.WITH

Đáp án: B

2. Câu nào sau đây mô tả sự khác biệt chính giữa dữ liệu rộng và dài?

A.Chủ thể dữ liệu rộng có thể có dữ liệu trong nhiều cột. Chủ thể dữ liệu dài có thể có nhiều hàng chứa các giá trị của thuộc tính

B.Chủ thể dữ liệu rộng có thể có nhiều hàng chứa các giá trị của thuộc tính chủ thể. Chủ thể dữ liệu dài có thể có dữ liệu trong nhiều cột

C.Mỗi chủ thể dữ liệu rộng đều có một cột duy nhất chứa các giá trị của thuộc tính chủ thể. Mỗi chủ thể dữ liệu dài đều có nhiều cột

D.Chủ thể dữ liệu rộng có nhiều cột. Chủ thể dữ liệu dài có một cột

Đáp án: A

3. Việc chuyển đổi dữ liệu cho phép các nhà phân tích dữ liệu thực hiện được điều gì?

A.Kiểm tra độ chính xác của dữ liệu

B.Khôi phục dữ liệu sau khi bị mất

C. Thay đổi cấu trúc của dữ liệu

D. Truy xuất dữ liệu nhanh hơn

Đáp án: B,C,D

### Câu hỏi tổng hợp

1. Nếu bạn chỉ có thời gian ngắn để thu thập dữ liệu và cần câu trả lời ngay lập tức, bạn có thể sẽ phải sử dụng dữ liệu lịch sử đã thu thập trong quá khứ?

A. Đúng

B. Sai

Đáp án: A

2. Một nhà phân tích dữ liệu đang thực hiện một nghiên cứu khẩn cấp về lưu lượng giao thông. Do khung thời gian gấp rút, họ nên sử dụng loại dữ liệu nào nhất?

A. Không sạch (Unclean)

B. Lịch sử (Historical)

C. Lý thuyết (Theoretical)

D. Cá nhân (Personal)

Đáp án: B

3. Một nhà phân tích dữ liệu tại một nhà xuất bản sách đang thực hiện gấp một báo cáo cho các lãnh đạo. Họ chỉ sử dụng dữ liệu lịch sử. Lý do họ chọn dữ liệu lịch sử để phân tích là gì?

A. Dữ liệu liên tục thay đổi

B. Dự án có khung thời gian rất ngắn

C. Dữ liệu không xác định

D. Có nhiều thời gian để nghiên cứu dữ liệu lịch sử

Đáp án: B

4. Dữ liệu nào sau đây là dữ liệu liên tục:

- A. Kinh phí phim
- B. Thời lượng phim
- C. Doanh thu phòng vé
- D. Diễn viên chính trong phim

Đáp án: B

5. Ví dụ nào sau đây là dữ liệu rời rạc? Chọn tất cả những câu phù hợp.

- A. Kinh phí phim
- B. Thời gian chạy phim
- C. Doanh thu phòng vé
- D. Số lượng diễn viên trong phim

Đáp án: A, C, D

6. Dữ liệu liên tục được đo lường và có số lượng giới hạn giá trị có thể có.

A. Đúng

B. Sai

Đáp án: B

7. Câu hỏi nào sau đây thu thập dữ liệu định tính không thứ tự? Chọn tất cả các câu phù hợp.

- A. Khả năng bạn giới thiệu nhà hàng này cho bạn bè là bao nhiêu?
- B. Có ai giới thiệu nhà hàng của chúng tôi cho bạn hôm nay không?
- C. Bạn đã nghe nói về bữa tối hàng ngày tại nhà hàng của chúng tôi chưa?
- D. Đây có phải là lần đầu tiên bạn dùng bữa tại nhà hàng này?

Đáp án: B, C, D

8. Câu hỏi nào sau đây thu thập dữ liệu định tính không thứ tự

- A. Trên thang điểm từ 1-10, bạn đánh giá dịch vụ của mình ngày hôm nay như thế nào?



B. Bạn đã ăn tối ở nhà hàng này bao nhiêu lần?

C. Bạn thường dùng bữa với bao nhiêu người?

D. Đây có phải là lần đầu tiên bạn dùng bữa tại nhà hàng này?

Đáp án: D

9. Dữ liệu định tính không thứ tự có một thứ tự hoặc phép đo được thiết lập.

A. Đúng

B. Sai

Đáp án: B

10. Câu nào sau đây là lợi ích của dữ liệu nội bộ?

A. Dữ liệu nội bộ đáng tin cậy hơn và dễ thu thập hơn

B. Dữ liệu nội bộ là dữ liệu duy nhất có liên quan đến vấn đề

C. Dữ liệu nội bộ ít có khả năng cần làm sạch hơn

D. Dữ liệu nội bộ ít bị thiên kiến khi thu thập

Đáp án: A

11. Tại sao dữ liệu nội bộ được coi là đáng tin cậy hơn và dễ thu thập hơn dữ liệu bên ngoài?

A. Dữ liệu nội bộ vượt qua các hạn chế về quyền riêng tư

B. Dữ liệu nội bộ đến từ những người bạn biết

C. Dữ liệu nội bộ nằm trong hệ thống riêng của công ty

D. Dữ liệu nội bộ có kích thước mẫu lớn hơn nhiều

Đáp án: C

12. Dữ liệu nội bộ đáng tin cậy hơn vì nó sạch

A. Đúng

B. Sai

Đáp án: B

13. Dữ liệu có cấu trúc có thể được tìm thấy ở định dạng nào sau đây? Chọn tất cả các câu phù hợp.

- A. Bảng
- B. File âm thanh
- C. Bảng tính
- D. Ảnh kỹ thuật số

Đáp án: A,C

14. Một bài đăng trên mạng xã hội là một ví dụ về dữ liệu có cấu trúc

- A. Đúng
- B. Sai

Đáp án: B

15. Câu nào sau đây là ví dụ về dữ liệu có cấu trúc.

- A. Cơ sở dữ liệu quan hệ
- B. Ảnh kỹ thuật số
- C. File âm thanh
- D. File video

Đáp án: A

16. Điền vào chỗ trống: Kiểu dữ liệu Boolean có thể có \_\_\_\_\_ giá trị khả dĩ

- A. Hai
- B. Ba
- C. Vô hạn
- D. Mười

Đáp án: A

17. Kiểu dữ liệu Boolean phải có giá trị số

- A. Đúng

B.Sai

Đáp án: B

18. Giá trị nào sau đây là ví dụ về kiểu dữ liệu Boolean? Chọn tất cả các câu phù hợp

A.Đúng hoặc sai

B.Một, hai hoặc ba

C.Có, không, hoặc không chắc

D.Có hoặc không

Đáp án: A,D

19. Đây là lựa chọn từ một bảng tính.

Nó chứa định dạng dữ liệu nào?

A.Rộng

B.Hẹp

C.Dài

D.Ngắn

Đáp án: A

20. Đây là lựa chọn từ một bảng tính.

Nó chứa định dạng dữ liệu nào?

A.Rộng

B.Hẹp

C.Dài

D.Ngắn

Đáp án: C

21. Trong dữ liệu dài, các cột chứa các giá trị và ngữ cảnh cho các giá trị. Mỗi cột chứa gì trong dữ liệu rộng?

A.Một biến dữ liệu duy nhất

- B. Một loại dữ liệu cụ thể
- C. Một định dạng duy nhất
- D. Một ràng buộc cụ thể

Đáp án: A

22. Việc chuyển đổi dữ liệu có thể thay đổi cấu trúc của dữ liệu. Một ví dụ về điều này là lấy dữ liệu được lưu trữ ở một định dạng và chuyển đổi nó sang một định dạng khác.

A. Đúng

B. Sai

Đáp án: A

23. Điền vào chỗ trống: Việc chuyển đổi dữ liệu cho phép các nhà phân tích dữ liệu thay đổi \_\_\_\_\_ của dữ liệu.

A. Cấu trúc

B. Giá trị

C. Ý nghĩa

D. Sự chính xác

Đáp án: A

24. Một nhà phân tích dữ liệu đang làm việc trong một ứng dụng bảng tính. Họ sử dụng Save As để thay đổi loại tệp từ .XLS thành .CSV. Đây là một ví dụ về chuyển đổi dữ liệu.

A. Đúng

B. Sai

Đáp án: A

## Đánh giá chất lượng dữ liệu

### 1. Câu hỏi về dữ liệu không thiên kiến và khách quan

1. Ví dụ nào sau đây là ví dụ về thiên kiến lấy mẫu? Chọn tất cả những câu phù hợp.

- A. Một cuộc khảo sát về học sinh ở độ tuổi đi học nhưng không bao gồm học sinh học tại nhà
- B. Một cuộc thăm dò bầu cử quốc gia chỉ phỏng vấn những người có bằng đại học
- C. Một công ty phân tích tiếp thị trực tuyến lưu trữ dữ liệu trong một bảng tính
- D. Một nghiên cứu lâm sàng bao gồm nam giới nhiều hơn nữ giới gấp ba lần

Đáp án: A,B,D

2. Điền vào chỗ trống: Xu hướng tìm kiếm hoặc giải thích thông tin theo cách xác thực những niềm tin đã có từ trước là thiên kiến \_\_\_\_\_.

- A. Lấy mẫu
- B. Quan sát
- C. Xác nhận
- D. Diễn giải

Đáp án: C

3. Thuật ngữ nào sau đây cũng là cách mô tả sự thiên kiến của người quan sát? Chọn tất cả những câu phù hợp.

- A. Thiên kiến người thử nghiệm (Experimenter bias)
- B. Thiên kiến nghiên cứu (Research bias)
- C. Thiên kiến khán giả (Spectator bias)
- D. Thiên kiến nhận thức (Perception bias)

Đáp án: A,B

## 2. Câu hỏi về sự tin cậy của dữ liệu

1. Nguồn nào sau đây thường là nguồn dữ liệu tốt? Chọn tất cả những câu phù hợp.

- A. Tập dữ liệu công khai đã được kiểm duyệt
- B. Bài báo học thuật
- C. Các trang mạng xã hội
- D. Dữ liệu cơ quan chính phủ

Đáp án: A,B,D

2. Để xác định xem một nguồn dữ liệu có được trích dẫn hay không, bạn nên hỏi câu hỏi nào sau đây? Chọn tất cả những câu phù hợp.

- A. Tập dữ liệu này có phải từ một tổ chức đáng tin cậy không?
- B. Ai đã tạo ra tập dữ liệu này?
- C. Dữ liệu có liên quan đến vấn đề tôi đang cố gắng giải quyết không?
- D. Tập dữ liệu này đã được làm sạch đúng cách chưa?

Đáp án: A,B

3. Một nhà phân tích dữ liệu đang phân tích dữ liệu bán hàng cho phiên bản mới nhất của sản phẩm. Họ sử dụng dữ liệu của bên thứ ba về phiên bản cũ hơn của sản phẩm. Không nên làm như vậy cho phân tích dữ liệu. Vì sao? Chọn tất cả những câu phù hợp.

- A. Dữ liệu không phải lấy từ gốc
- B. Dữ liệu không chính xác
- C. Dữ liệu bị sai lệch
- D. Dữ liệu không mới

Đáp án: A,D

### 3. Câu hỏi về đạo đức và tính riêng tư của dữ liệu

1. Điền vào chỗ trống: \_\_\_\_\_ nói rằng tất cả các hoạt động và thuật toán xử lý dữ liệu phải hoàn toàn có thể giải thích và hiểu được bởi cá nhân cung cấp dữ liệu

- A. Giao dịch minh bạch (Transaction transparency)
- B. Tính mở (Openness)
- C. Sự riêng tư (Privacy)
- D. Tiền tệ (Currency)

Đáp án: A

2. Một nhà phân tích dữ liệu xóa thông tin nhận dạng cá nhân khỏi tập dữ liệu. Họ đang thực hiện công việc gì?

- A. Thu thập dữ liệu
- B. Sắp xếp dữ liệu
- C. Ẩn danh dữ liệu
- D. Trực quan hóa dữ liệu

Đáp án: C

3. Trước khi thực hiện một cuộc khảo sát, người được khảo sát phải đọc thông tin về cách thức và lý do sử dụng dữ liệu cá nhân của họ. Khái niệm này được gọi là gì?

- A. Sự tùy ý
- B. Sự riêng tư
- C. Tiền tệ
- D. Sự đồng thuận

Đáp án: D

#### 4. Câu hỏi về dữ liệu mở

1. Khía cạnh nào của đạo đức dữ liệu thúc đẩy quyền truy cập, sử dụng và chia sẻ dữ liệu miễn phí?

- A. Giao dịch trong suốt (Transaction transparency)
- B. Tính mở (Openness)
- C. Sự riêng tư (Privacy)
- D. Tiền tệ (Currency)

Đáp án: B

2. Những lợi ích chính của dữ liệu mở là gì? Chọn tất cả những câu phù hợp.

- A. Dữ liệu mở làm cho dữ liệu tốt được phổ biến rộng rãi hơn
- B. Dữ liệu mở làm tăng lượng dữ liệu có sẵn để mua
- C. Dữ liệu mở giúp kết hợp dữ liệu từ các nguồn khác nhau
- D. Dữ liệu mở hạn chế quyền truy cập dữ liệu đối với một số nhóm người nhất định

Đáp án: A,C

3. Cho phép mọi người tham gia là một tiêu chuẩn của dữ liệu mở. Các khía cạnh chính của cho phép mọi người tham gia là gì? Chọn tất cả những câu phù hợp.

- A. Tất cả các công ty được phép bán dữ liệu mở
- B. Mọi người phải có thể sử dụng, tái sử dụng và phân phối lại dữ liệu mở
- C. Một số nhóm người nhất định phải chia sẻ dữ liệu cá nhân của họ
- D. Không ai có thể đặt ra các hạn chế trên dữ liệu nhằm phân biệt đối xử với một người hoặc một nhóm.

Đáp án: B,D



## 5. Câu hỏi tổng hợp

1. Điền vào chỗ trống: Ưu tiên ủng hộ hoặc ghét bỏ một người, một nhóm người hoặc một sự vật được gọi là \_\_\_\_\_. Đó là một lỗi trong phân tích dữ liệu có thể làm sai lệch kết quả một cách có hệ thống theo một hướng nhất định.

- A. Ẩn danh dữ liệu
- B. Thu thập dữ liệu
- C. Thiên kiến dữ liệu
- D. Khả năng tương tác dữ liệu

Đáp án: C

2. Tình huống nào sau đây là ví dụ về sự thiên kiến? Chọn tất cả các câu phù hợp

- A. Một học giả chỉ đọc các nguồn hỗ trợ lập luận của họ
- B. Một nhà trẻ không thuê nam giới cho các vị trí trông trẻ
- C. Một nhà nghiên cứu khảo sát một nhóm mẫu đại diện cho tập hợp
- D. Giám khảo cuộc thi khiêu vũ là bạn thân của vũ công thắng cuộc thi

Đáp án: A,B,D

3. Một phòng khám khảo sát một nhóm bệnh nhân nam và nữ về trải nghiệm của họ với vật lý trị liệu. Cuộc khảo sát không bao gồm người khuyết tật. Dữ liệu khảo sát có bị thiên kiến không

- A. Có
- B. Không

Đáp án: A

4. Câu nào sau đây là loại thiên kiến dữ liệu thường gặp trong phân tích dữ liệu? Chọn tất cả các câu phù hợp

- A. Thiên kiến quan sát
- B. Thiên kiến diễn giải

C.Thiên kiến xác nhận

D.Thiên kiến giáo dục

Đáp án: A,B,C

5.Một trường đại học khảo sát sinh viên-vận động viên của mình về kinh nghiệm của họ trong các môn thể thao đại học. Cuộc khảo sát chỉ bao gồm sinh viên-vận động viên có học bổng. Đây là một ví dụ về kiểu thiên kiến nào?

A.Thiên kiến xác nhận

B.Thiên kiến diễn giải

C.Thiên kiến lấy mẫu

D.Thiên kiến quan sát

Đáp án: C

6.Loại thiên kiến nào có xu hướng luôn lý giải các tình huống mơ hồ theo hướng tích cực hay tiêu cực?

A.Thiên kiến quan sát

B.Thiên kiến xác nhận

C.Thiên kiến diễn giải

D.Thiên kiến lấy mẫu

Đáp án: C

7.Nói chung, tính hữu ích của dữ liệu giảm dần theo thời gian.

A.Đúng

B.Sai

Đáp án: A

8. Điều nào sau đây là đặc điểm của dữ liệu không đáng tin cậy? Chọn tất cả các câu phù hợp

A.Thiếu chính xác

B.Chưa toàn diện

C. Thiên kiến

D. Đã kiểm định

Đáp án: A,B,C

9. Câu nào sau đây mô tả tính chất của dữ liệu tốt? Chọn tất cả những gì phù hợp

A. Toàn diện

B. Có trích dẫn

C. Là kết quả logic

D. Tính mới

Đáp án: A,B,D

10. Nếu một công ty sử dụng dữ liệu cá nhân của bạn để kiếm tiền, bạn nên biết về điều này và quy mô của nó. Đây là khái niệm đạo đức dữ liệu nào?

A. Quyền sở hữu (Ownership)

B. Sự đồng thuận (Consent)

C. Tiền tệ (Currency)

D. Sự riêng tư (Privacy)

Đáp án: C

11. Trong đạo đức dữ liệu, sự đồng thuận cho phép một cá nhân có quyền biết câu trả lời cho câu hỏi nào sau đây? Chọn tất cả các câu phù hợp?

A. Tại sao dữ liệu của tôi được thu thập?

B. Dữ liệu của tôi sẽ được sử dụng như thế nào?

C. Dữ liệu của tôi sẽ được lưu trữ trong bao lâu?

D. Tại sao tôi buộc phải chia sẻ dữ liệu của mình?

Đáp án: A,B,C

12. Điền vào chỗ trống: \_\_\_\_\_ dữ liệu để cập đến các tiêu chuẩn có cơ sở về thế nào là đúng và sai, quy định cách dữ liệu được thu thập, chia sẻ và sử dụng

A. Đạo đức

- B.Sự tin cậy
- C.Sự riêng tư
- D.Sự ẩn danh

Đáp án: A

13. Cá nhân cung cấp dữ liệu có quyền biết và hiểu tất cả các hoạt động và thuật toán xử lý dữ liệu được sử dụng trên dữ liệu đó. Đây được gọi là quyền sở hữu

- A.Đúng
  - B.Sai
- Đáp án: B

14. Quyền sở hữu là một vấn đề then chốt trong đạo đức dữ liệu. Ai sở hữu dữ liệu?

- A.Tổ chức đầu tư thời gian và tiền bạc để thu thập, xử lý và phân tích dữ liệu
- B.Cá nhân tạo dữ liệu ban đầu
- C.Chính phủ nơi thông qua luật bảo vệ dữ liệu
- D.Các cơ quan thực thi pháp luật thực thi luật bảo vệ dữ liệu

Đáp án: B

15. Cá nhân cung cấp dữ liệu có quyền biết và hiểu tất cả các hoạt động và thuật toán xử lý dữ liệu được sử dụng trên dữ liệu đó. Khái niệm này đề cập đến khía cạnh nào của đạo đức dữ liệu?

- A.Quyền sở hữu
- B.Giao dịch minh bạch
- C.Bằng lòng
- D.Tiền tệ

Đáp án: B

16. Tính riêng tư của dữ liệu là gì?

- A.Tìm kiếm hoặc giải thích các thông tin hỗ trợ

B. Áp dụng các tiêu chuẩn có cơ sở về tính đúng và sai, quy định cách dữ liệu được thu thập, chia sẻ và sử dụng

C. Bảo vệ thông tin và hoạt động của chủ thể dữ liệu cho tất cả các giao dịch dữ liệu

D. Cung cấp quyền truy cập, sử dụng và chia sẻ dữ liệu miễn phí

Đáp án: C

17. Người sử dụng lao động truy cập báo cáo tín dụng của nhân viên mà không có sự đồng ý của họ. Điều này không vi phạm quyền riêng tư của nhân viên vì họ làm việc tại công ty

A. Đúng

B. Sai

Đáp án: B

18. Quyền kiểm tra, cập nhật hoặc sửa dữ liệu của chính bạn thuộc về khía cạnh nào của đạo đức dữ liệu?

A. Tính mở của dữ liệu

B. Quyền riêng tư dữ liệu

C. Sự đồng thuận dữ liệu

D. Quyền sở hữu dữ liệu

Đáp án: B

19. Ẩn danh dữ liệu áp dụng cho cả văn bản và hình ảnh

A. Đúng

B. Sai

Đáp án: A

20. Quy trình bảo vệ dữ liệu nhạy cảm hoặc riêng tư của người khác bằng cách loại bỏ thông tin nhận dạng là gì

A. Ẩn danh dữ liệu

B. Đạo đức dữ liệu

C. Thiết kế dữ liệu

## D.Quản trị dữ liệu

Đáp án: A

21. Phương pháp nào sau đây thường được sử dụng để ẩn danh dữ liệu? Chọn tất cả các câu phù hợp

A.Xóa trắng (blanking)

B.Băm (hashing)

C.Áp mặt nạ (masking)

D.Xóa (deleting)

Đáp án: A,B,C

22. Khả năng tương tác là chìa khóa thành công của dữ liệu mở. Điều nào sau đây là một ví dụ về khả năng tương tác?

A.Một trang web tính phí để truy cập cơ sở dữ liệu

B.Các cơ sở dữ liệu khác nhau sử dụng các định dạng và thuật ngữ chung

C.Nhà phân tích xóa tất cả thông tin nhận dạng cá nhân khỏi cơ sở dữ liệu

D.Một công ty chỉ cho chính nhân viên của mình sử dụng cơ sở dữ liệu

Đáp án: B

23. Khía cạnh quan trọng của dữ liệu mở là quyền truy cập miễn phí vào thông tin cá nhân của mọi người.

A.Đúng

B.Sai

Đáp án: B

24. Chính quyền của một thành phố lớn thu thập dữ liệu về chất lượng cơ sở hạ tầng của thành phố. Mọi doanh nghiệp, tổ chức phi lợi nhuận hoặc cá nhân đều có thể truy cập cơ sở dữ liệu của chính phủ và sử dụng lại hoặc phân phối lại dữ liệu. Đây là một ví dụ về dữ liệu mở không?

A.Đúng

B.Sai

Đáp án: A

## Làm việc với cơ sở dữ liệu

### 1. Câu hỏi về CSDL

1. Điền vào chỗ trống: \_\_\_\_\_ là chỉ số tham chiếu đến cột cơ sở dữ liệu trong đó mỗi giá trị là duy nhất?

- A. Khóa chính
- B. Trường
- C. Khóa ngoại
- D. Quan hệ

Đáp án: A

2. Điền vào chỗ trống: Cơ sở dữ liệu quan hệ chứa một tập hợp các \_\_\_\_\_ có thể được kết nối để tạo thành quan hệ.

- A. Trường
- B. Ô
- C. Bảng
- D. Bảng tính

Đáp án: C

3. Lợi ích chính khi làm việc với cơ sở dữ liệu chuẩn hóa là giảm dư thừa dữ liệu. Điều nào sau đây là một ví dụ về dư thừa?

- A. Cùng một dữ liệu nhưng được lưu trữ ở hai nơi khác nhau
- B. Cơ sở dữ liệu chứa hai khóa ngoại
- C. Cơ sở dữ liệu tạo thành hai hoặc nhiều mối quan hệ
- D. Các thành viên trong nhóm ở các vị trí văn phòng khác nhau làm việc với cùng một dữ liệu

Đáp án: A

## 2. Câu hỏi về siêu dữ liệu

1. Một công ty lớn có một nhiều bộ dữ liệu trên nhiều phòng ban của mình. Loại siêu dữ liệu nào cho biết chính xác có bao nhiêu bộ dữ liệu tồn tại?

- A. Mô tả
- B. Cấu trúc
- C. Hành chính
- D. Đại diện

Đáp án: B

2. Ngày và giờ chụp ảnh là một ví dụ về loại siêu dữ liệu nào?

- A. Mô tả
- B. Cấu trúc
- C. Quản trị
- D. Đại diện

Đáp án: C

3. Một trường trung học ở thành phố lớn cấp cho mỗi học sinh của mình một số ID để phân biệt các em trong cơ sở dữ liệu của trường. Số ID là loại siêu dữ liệu nào?

- A. Mô tả
- B. Cấu trúc
- C. Hành chính
- D. Đại diện

Đáp án: A

4. Một công ty cần hợp nhất dữ liệu của bên thứ ba với dữ liệu của chính mình. Hành động



nào sau đây sẽ giúp làm cho quá trình này thành công? Chọn tất cả các câu phù hợp

- A. Sử dụng siêu dữ liệu để chuẩn hóa dữ liệu.
- B. Thay đổi siêu dữ liệu của công ty để phù hợp hơn với siêu dữ liệu của bên thứ ba
- C. Thay thế siêu dữ liệu của dữ liệu bên thứ ba bằng siêu dữ liệu của công ty
- D. Sử dụng siêu dữ liệu để đánh giá chất lượng và độ tin cậy của dữ liệu bên thứ ba

Đáp án: A

### 3. Câu hỏi về truy xuất nguồn dữ liệu

1. File CSV lưu dữ liệu ở định dạng bảng. CSV là viết tắt của gì?

- A. Comma-separated values
- B. Cell-structured variables
- C. Compatible scientific variables
- D. Calculated spreadsheet values

Đáp án: A

2. Một nhà phân tích dữ liệu muốn đưa dữ liệu từ tệp CSV vào bảng tính. Đây là một ví dụ về quá trình nào?

- A. Lưu trữ dữ liệu
- B. Nạp dữ liệu
- C. Chỉnh sửa dữ liệu
- D. Chuẩn hóa dữ liệu

Đáp án: B

3. File CSV giúp nhà phân tích dữ liệu hoàn thành tác vụ nào dễ dàng hơn? Chọn tất cả những câu phù hợp.

- A. Kiểm tra một tập con nhỏ của tập dữ liệu lớn
- B. Phân biệt các giá trị với nhau
- C. Nhập dữ liệu vào một bảng tính mới

D. Quản lý nhiều tab trong một trang tính

Đáp án: A,B,C

#### 4. Câu hỏi về lọc và sắp xếp

1. Quá trình sắp xếp dữ liệu thành một thứ tự có ý nghĩa để dễ hiểu, phân tích và hình dung là gì?

- A. Lọc
- B. Sắp xếp
- C. Chỉnh khung
- D. Ưu tiên

Đáp án: B

2. Một nhà phân tích dữ liệu đang xem xét cơ sở dữ liệu quốc gia về mua bán bất động sản. Họ chỉ quan tâm đến việc mua bán nhà chung cư. Làm thế nào nhà phân tích có thể thu hẹp phạm vi của họ?

- A. Sắp xếp theo doanh số mua bán chung cư
- B. Sắp xếp theo doanh số mua bán không phải chung cư
- C. Lọc ra các giao dịch mua bán chung cư
- D. Lọc ra các giao dịch mua bán không phải chung cư

Đáp án: D

3. Một nhà phân tích dữ liệu làm việc cho một công ty cho thuê xe hơi. Họ có một bảng tính liệt kê biển số xe ô tô và ngày những chiếc ô tô được trả lại. Làm cách nào để họ có thể sắp xếp bảng tính để tìm những chiếc xe được trả lại gần đây nhất?

- A. Theo biển số xe, theo thứ tự tăng dần
- B. Theo biển số xe, theo thứ tự giảm dần
- C. Theo ngày trả, theo thứ tự tăng dần
- D. Theo ngày trả, theo thứ tự giảm dần

Đáp án: D

4. Điền vào chỗ trống: Để giữ hàng tiêu đề ở đầu bảng tính, hãy đánh dấu hàng và chọn \_\_\_\_\_ từ view menu.

- A. Đông cứng (Freeze)
- B. Khóa (Lock)
- C. Ghim (Pin)
- D. Thiết lập (Set)

Đáp án: B

## 5. Câu hỏi tổng hợp

1. Cơ sở dữ liệu quan hệ chứa các bảng được kết nối với nhau để tạo ra các mối quan hệ. Hai loại trường nào dùng để kết nối hai bảng?

- A. Khóa chính và khóa ngoại
- B. Dữ liệu bên trong và bên ngoài
- C. Lược đồ sao và bông tuyết
- D. Siêu dữ liệu mô tả và cấu trúc

Đáp án: A

2. Cơ sở dữ liệu quan hệ minh họa mối quan hệ giữa các bảng. Trường nào đại diện cho kết nối giữa các bảng này? Chọn tất cả những câu phù hợp.

- A. Khóa chính
- B. Khóa ngoại
- C. Khóa phụ
- D. Khóa ngoài

Đáp án: B

3. Khóa chính và khóa ngoại là hai mã định danh được kết nối trong các bảng riêng biệt. Các bảng này tồn tại trong loại cơ sở dữ liệu nào?

- A. Quan hệ
- B. Bình thường hóa
- C. Chính
- D. Siêu dữ liệu

Đáp án: A

4. Nhà phân tích dữ liệu sử dụng siêu dữ liệu cho những công việc gì? Chọn tất cả các câu phù hợp

- A. Để kết hợp dữ liệu từ nhiều nguồn
- B. Để đánh giá chất lượng dữ liệu
- C. Để thực hiện phân tích dữ liệu
- D. Để diễn giải nội dung của cơ sở dữ liệu

Đáp án: A,B,D

5. Siêu dữ liệu là dữ liệu về dữ liệu. Siêu dữ liệu có thể cung cấp những loại thông tin nào về một tập dữ liệu cụ thể? Chọn tất cả các câu phù hợp

- A. Dữ liệu sạch và đáng tin cậy hay không
- B. Cách kết hợp dữ liệu với một tập dữ liệu khác
- C. Chứa những loại dữ liệu nào
- D. Phân tích nào được thực hiện trên dữ liệu

Đáp án: A,B,C

6. Khi làm việc với dữ liệu từ một nguồn bên ngoài, siêu dữ liệu có thể giúp các nhà phân tích dữ liệu làm gì? Chọn tất cả các câu phù hợp

- A. Đảm bảo dữ liệu sạch và đáng tin cậy
- B. Kết hợp dữ liệu từ nhiều nguồn
- C. Chọn loại phân tích để chạy

D. Hiểu nội dung của cơ sở dữ liệu

Đáp án: A,B,D

7. Ví dụ dữ liệu về học sinh tại một trường trung học phổ thông. Câu nào sau đây là về siêu dữ liệu? Chọn tất cả các câu phù hợp

- A. Mã số của học sinh
- B. Điểm của học sinh
- C. Các lớp học mà học sinh đã đăng ký
- D. Ngày nhập học của học sinh

Đáp án: A,C,D

8. Hãy nghĩ về dữ liệu giống như việc lái một chiếc taxi. Trong phép ẩn dụ này, câu nào sau đây là siêu dữ liệu? Chọn tất cả những câu phù hợp.

- A. Biển số xe
- B. Công ty sở hữu taxi
- C. Hành khách mà taxi đón
- D. Chế tạo và mô hình xe taxi

Đáp án: A,B,D

9. Siêu dữ liệu cấu trúc cho biết cách dữ liệu được tổ chức và liệu nó có phải là một phần của một hay nhiều bộ dữ liệu hay không.

- A. Đúng
- B. Sai

Đáp án: A

10. Quy trình mà các nhà phân tích dữ liệu sử dụng để đảm bảo việc quản lý chính thức tài sản dữ liệu của công ty họ là

- A. Quản trị dữ liệu (Data governance)
- B. Tổng hợp dữ liệu (Data aggregation)
- C. Toàn vẹn dữ liệu (Data integrity)

D. Ánh xạ dữ liệu (Data mapping)

Đáp án: A

11. Điền vào chỗ trống: \_\_\_\_\_ dữ liệu là quy trình đảm bảo việc quản lý chính thức tài sản dữ liệu của công ty.

- A. Quản trị (governance)
- B. Tổng hợp (aggregation)
- C. Toàn vẹn (integrity)
- D. Ánh xạ (mapping)

Đáp án: A

12. Điền vào chỗ trống: Quản trị dữ liệu là quá trình đảm bảo rằng \_\_\_\_\_ của công ty được quản lý một cách chính thức.

- A. Tài sản dữ liệu
- B. Chiến lược kinh doanh
- C. Nhiệm vụ kinh doanh
- D. Kỹ sư dữ liệu

Đáp án: A

13. Một số lợi ích chính của việc sử dụng dữ liệu bên ngoài là gì? Chọn tất cả những câu phù hợp.

- A. Dữ liệu bên ngoài luôn đáng tin cậy
- B. Dữ liệu bên ngoài có phạm vi tiếp cận rộng
- C. Dữ liệu bên ngoài được sử dụng miễn phí
- D. Dữ liệu bên ngoài có thể cung cấp tầm nhìn cho toàn ngành

Đáp án: B,D

14. Một nhà phân tích dữ liệu không sử dụng dữ liệu bên ngoài vì nó đại diện cho các góc nhìn đa dạng. Ý kiến này?

- A. Đúng

B. Sai

Đáp án: B

15. Trong trường hợp nào một nhà phân tích dữ liệu có thể chọn không sử dụng dữ liệu bên ngoài trong phân tích của họ?

- A. Dữ liệu quá đầy đủ
- B. Dữ liệu không thể được xác nhận là đáng tin cậy
- C. Dữ liệu đại diện cho các góc nhìn đa dạng
- D. Dữ liệu miễn phí truy cập cho mọi người

Đáp án: B

16. Một nhà phân tích dữ liệu xem xét cơ sở dữ liệu quốc gia về các suất chiếu tại rạp chiếu phim. Họ muốn tìm những bộ phim đầu tiên được chiếu ở San Francisco vào năm 2001. Làm cách nào để họ có thể sắp xếp dữ liệu để trả về 10 bộ phim đầu tiên được chiếu ở đầu danh sách? Chọn tất cả những câu phù hợp

- A. Lọc ra các buổi chiếu bên ngoài San Francisco
- B. Sắp xếp theo ngày theo thứ tự tăng dần
- C. Sắp xếp theo ngày theo thứ tự giảm dần
- D. Lọc ra các bộ phim không chiếu trong năm 2001

Đáp án: A,B,D

17. Một tổ chức phi lợi nhuận duy trì danh sách số lượng máy tính xách tay mà họ cung cấp cho mỗi trường học trong quận. Trong bảng, có một cột được gọi là `number_of_laptops`. Một nhà phân tích dữ liệu muốn xác định trường nào được cấp ít máy tính xách tay nhất. Họ nên sắp xếp dữ liệu như thế nào để trả về các trường này trước?

- A. Sắp xếp số theo thứ tự tăng dần
- B. Sắp xếp số theo thứ tự giảm dần
- C. Sắp xếp theo thứ tự bảng chữ cái theo thứ tự tăng dần
- D. Sắp xếp theo thứ tự bảng chữ cái theo thứ tự giảm dần

Đáp án: A

18. Một nhà phân tích dữ liệu xem xét cơ sở dữ liệu về doanh số bán xe của Wisconsin để tìm ra những mẫu xe cuối cùng được bán ở Milwaukee vào năm 2019. Làm cách nào để họ có thể sắp xếp và lọc dữ liệu để trả về năm chiếc xe cuối cùng được bán ở đầu danh sách của họ? Chọn tất cả các câu phù hợp?

- A. Lọc ra doanh số bán hàng bên ngoài Milwaukee
- B. Sắp xếp theo ngày bán theo thứ tự tăng dần
- C. Sắp xếp theo ngày bán theo thứ tự giảm dần
- D. Lọc ra doanh số bán hàng không có trong năm 2019

Đáp án: A,C,D

19. Khi viết một câu truy vấn, tên của tập dữ liệu phải nằm trong hai dấu gạch ‘ ‘ để truy vấn chạy đúng

A. Đúng

B. Sai

Đáp án: B

20. Khi viết một câu truy vấn, tên của tập dữ liệu có thể nằm trong hai dấu gạch ‘ ‘ hoặc không, để truy vấn chạy đúng

A. Đúng

B. Sai

Đáp án: A

21. Khi viết một câu truy vấn, bạn phải xóa hai dấu gạch ‘ ‘ nằm xung quanh tên của tập dữ liệu, để truy vấn chạy đúng

A. Đúng

B. Sai

Đáp án: B

22. Bạn đang làm việc với một bảng cơ sở dữ liệu tên là **customer** chứa dữ liệu khách hàng. Cột `first_name` liệt kê tên của mỗi khách hàng. Bạn chỉ quan tâm đến những khách hàng có tên Mark



Hãy hoàn chỉnh câu truy vấn SQL. Thêm mệnh đề WHERE để chỉ trả về những khách hàng có tên Mark.

**SELECT \* FROM customer**

Đáp án: **SELECT \* FROM customer**  
**WHERE first\_name = 'Mark'**

23. Bạn đang làm việc với một bảng cơ sở dữ liệu tên là **customer** chứa dữ liệu khách hàng. Cột **city** liệt kê thành phố nơi mỗi khách hàng ở. Bạn muốn tìm những khách hàng đang ở Berlin.

Hãy hoàn chỉnh câu truy vấn SQL. Thêm mệnh đề WHERE để chỉ trả về những khách hàng đang sống ở Berlin.

**SELECT \* FROM customer**

Đáp án: **SELECT \* FROM customer**  
**WHERE city = 'Berlin'**

24. Bạn đang làm việc với một bảng cơ sở dữ liệu tên là **customer** chứa dữ liệu khách hàng. Cột **state** liệt kê tiểu bang nơi mỗi khách hàng ở. Tên tiểu bang được viết tắt. Bạn muốn tìm những khách hàng đang ở Florida (FL).

Hãy hoàn chỉnh câu truy vấn SQL. Thêm mệnh đề WHERE để chỉ trả về những khách hàng đang sống ở FL.

**SELECT \* FROM customer**

Đáp án: **SELECT \* FROM customer**  
**WHERE state = 'FL'**

## Quản lý và bảo mật dữ liệu

### 1. Câu hỏi về làm thế nào tổ chức dữ liệu

1. Các nhà phân tích dữ liệu sử dụng các nguyên tắc để mô tả phiên bản, nội dung và ngày được tạo của file. Những nguyên tắc này được gọi là gì?

A. Xác minh đặt tên

- B. Thuộc tính đặt tên
- C. Tham chiếu đặt tên
- D. Quy ước đặt tên

Đáp án: D

2. Các nhà phân tích dữ liệu sử dụng phương pháp sắp xếp thư mục để đạt được những mục tiêu nào? Chọn tất cả những câu phù hợp.

- A. Để chuyển file từ nơi này sang nơi khác
- B. Để tổ chức các file thành các thư mục con
- C. Để gán siêu dữ liệu về các thư mục
- D. Để giữ các file liên quan đến dự án cùng nhau

Đáp án: B,D

3. Điền vào chỗ trống: Để tách biệt công việc hiện tại và quá khứ, giảm bớt sự lộn xộn, các nhà phân tích dữ liệu tạo \_\_\_\_\_. Điều này liên quan đến việc di chuyển các file từ các dự án đã hoàn thành đến một vị trí riêng biệt.

- A. Sao lưu (backups)
- B. Cấu trúc (structures)
- C. Lưu trữ (archives)
- D. Sao chép (copies)

Đáp án: C

4. Quá trình cấu trúc các thư mục với chủ đề rộng ở trên cùng, sau đó chia nhỏ các thư mục đó thành các chủ đề cụ thể hơn là gì?

- A. Phát triển siêu dữ liệu
- B. Chỉ định quy ước đặt tên
- C. Tạo một hệ thống phân cấp
- D. Tạo một bản sao lưu

Đáp án: C

5. Quy ước đặt tên file tốt bao gồm thông tin hữu ích giúp định vị hoặc cập nhật file. Tên nào sau đây là tên file tốt?

- A. AirportCampaign\_2013\_10\_09\_V01
- B. CampaignData\_03
- C. May30-2019\_AirportAdvertisingCampaignResults\_Terminals3-5\_InclCustSurveyResponses\_PLUS\_IdeasforJune
- D. Data\_519

Đáp án: A

## 2. Câu hỏi về bảo mật dữ liệu

1. Điền vào chỗ trống: Bảo mật dữ liệu liên quan đến việc sử dụng \_\_\_\_\_ để bảo vệ dữ liệu khỏi việc truy cập trái phép hoặc phá hoại.

- A. Xác nhận dữ liệu
- B. Sắp xếp thư mục
- C. Các biện pháp an toàn
- D. Siêu dữ liệu

Đáp án: C

2. Khi sử dụng các biện pháp bảo mật dữ liệu, nhà phân tích có thể chọn giữa bảo vệ toàn bộ bảng tính hoặc bảo vệ các ô nhất định trong bảng tính.

- A. Đúng
- B. Sai

Đáp án: A

3. Các nhà phân tích dữ liệu có thể sử dụng công cụ nào để kiểm soát những ai có thể truy cập hoặc chỉnh sửa bảng tính? Chọn tất cả những câu phù hợp.

- A. Mã hóa

- B. Chia sẻ quyền truy cập
- C. Bộ lọc
- D. Các tab

Đáp án: A,B

### 3. Câu hỏi tổng hợp

1. Nhóm phân tích dữ liệu gắn nhãn các file của mình để cho biết nội dung, ngày tạo và số phiên bản của chúng. Nhóm đang sử dụng công cụ tổ chức dữ liệu nào?

- A. Xác minh đặt tên file
- B. Quy ước đặt tên file
- C. Tham chiếu đặt tên file
- D. Thuộc tính đặt tên file

Đáp án: B

2. Điền vào chỗ trống: Quy ước đặt tên file là \_\_\_\_\_ mô tả nội dung, ngày tạo hoặc phiên bản của file.

- A. Xác minh chung
- B. Hướng dẫn nhất quán
- C. Gợi ý thường xuyên
- D. Thuộc tính chung

Đáp án: B

3. Để thống nhất cách đặt tên và lưu trữ file, nên xây dựng thủ tục siêu dữ liệu với nhóm phân tích dữ liệu

- A. Đúng
- B. Sai

Đáp án: A

4. Nhóm phân tích dữ liệu sử dụng dữ liệu về dữ liệu để chỉ ra các quy ước đặt tên nhất quán cho một dự án. Đây là loại dữ liệu gì?

- A. Siêu dữ liệu
- B. Dữ liệu lớn
- C. Dữ liệu dài
- D. Dữ liệu tổng hợp

Đáp án: A

5. Điền vào chỗ trống: Nhóm phân tích dữ liệu sử dụng \_\_\_\_ để chỉ ra các quy ước đặt tên nhất quán cho một dự án. Đây là một ví dụ về việc sử dụng dữ liệu về dữ liệu.

- A. Phân cấp thư mục
- B. Kiểm soát phiên bản
- C. Phân loại
- D. Siêu dữ liệu

Đáp án: D

6. Các nhà phân tích dữ liệu sử dụng các quy ước đặt tên để giúp họ xác định hoặc định vị một file. Tên file nào sau đây là cách đặt tên tốt?

- A. Elementary\_Students\_20090221\_V03
- B. ElementarySchoolStudents\_EnrollingSeptember2021\_PlusRisingMiddleSchool\_FJP SKVND
- C. Elem\_9
- D. Sept\_ElemtaryStudents\_V1

Đáp án: A

7. Một nhà phân tích dữ liệu đang làm việc với một file từ cuộc khảo sát mức độ hài lòng của khách hàng. Bản khảo sát đã được gửi cho bất kỳ ai đã trở thành khách hàng trong

khoảng thời gian từ tháng 4 đến tháng 6 năm 2020. Tên file nào sau đây là cách đặt tên tốt?

- A. Survey\_Responses
- B. Apr-June2020\_CustSurvey\_V
- C. NewCustomerSurvey\_2020-6-20\_V03
- D. April\_May\_June\_2020\_Responses\_to\_New\_Customer\_Survey\_ANALYSISDATA\_928310

Đáp án: C

8. Một nhà phân tích dữ liệu tạo một file danh sách những người đã quyên góp vào quỹ của tổ chức của họ. Cách đặt tên file như sau có tốt hay không: FundDriveDonors\_20210216\_V01.

- A. Có
- B. Không

Đáp án: A

9. Các nhà phân tích dữ liệu sử dụng phương pháp sắp xếp thư mục để tổ chức các thư mục thành cái gì?

- A. Phiên bản
- B. Bảng
- C. Thư mục con
- D. Cơ sở dữ liệu

Đáp án: C

10. Các nhà phân tích dữ liệu sử dụng một quy trình được gọi là mã hóa để tổ chức các thư mục thành các thư mục con.

- A. Đúng
- B. Sai

Đáp án: B

11. Các nhà phân tích dữ liệu sử dụng quy trình nào để giữ các file liên quan đến dự án lại

với nhau và sắp xếp chúng thành các thư mục con?

- A. Chỉnh sửa
- B. Sắp xếp thư mục
- C. Mã hóa
- D. Đặt tên

Đáp án: B

12. Khi một nhà phân tích dữ liệu hoàn thành dự án, họ di chuyển các file dự án đến một vị trí khác để giữ chúng tách biệt với công việc hiện tại của họ. Đây là một ví dụ về quá trình?

- A. Hủy các file
- B. Nhân bản file
- C. Lưu trữ file
- D. Đổi tên file

Đáp án: C

13. Các nhà phân tích dữ liệu sử dụng tính năng lưu trữ (archive) để tách biệt công việc hiện tại với công việc trước đây. Quá trình này bao gồm những gì?

- A. Di chuyển file từ các dự án đã hoàn thành sang một vị trí khác
- B. Sử dụng phần mềm xóa dữ liệu an toàn để hủy các file cũ
- C. Sắp xếp lại và đổi tên các file hiện tại
- D. Xem xét các file dữ liệu hiện tại để xác nhận rằng chúng đã được làm sạch

Đáp án: A.

14. Các nhà phân tích dữ liệu sử dụng tính năng lưu trữ (archive) để sao chép và giữ các bản sao lưu của các file quan trọng. Các bản sao lưu này được sử dụng nếu các file gốc bị mất.

- A. Đúng
- B. Sai

Đáp án: B

15. Các nhà phân tích dữ liệu tạo ra các cấu trúc phân cấp để tổ chức các thư mục của họ. Sự phân cấp thư mục được tổ chức như thế nào?

- A. Các chủ đề cụ thể ở ngoài, các chủ đề rộng hơn ở bên trong
- B. Các chủ đề rộng ở bên phải, các chủ đề cụ thể hơn ở bên trái
- C. Các chủ đề rộng ở bên trái, các chủ đề cụ thể hơn ở bên phải
- D. Các chủ đề rộng ở ngoài, các chủ đề cụ thể hơn ở bên trong

Đáp án: D

16. Điền vào chỗ trống: Các nhà phân tích dữ liệu tạo \_\_\_\_\_ để cấu trúc các thư mục của họ.

- A. Hệ thống chia độ (scales)
- B. Thang (ladders)
- C. Hệ thống phân cấp (hierarchies)
- D. Trình tự (sequences)

Đáp án: A

17. Các nhà phân tích dữ liệu tạo ra các cấu trúc phân cấp để tổ chức các thư mục của họ. Họ thực hiện điều này bằng cách cấu trúc các thư mục theo các chủ đề cụ thể ở ngoài cùng, sau đó chủ đề rộng hơn ở bên trong.

- A. Đúng
- B. Sai

Đáp án: B

18. Một nhà phân tích dữ liệu muốn đảm bảo chỉ những người trong nhóm phân tích của họ mới có thể truy cập, chỉnh sửa và tải xuống bảng tính. Họ có thể sử dụng công cụ nào sau đây? Chọn tất cả các câu đúng

- A. Mẫu (Templates)
- B. Quyền chia sẻ (Sharing permissions)



- C. Lọc (Filtering)
- D. Mã hóa (Encryption)

Đáp án: B,D

19. Nhà phân tích dữ liệu thêm quyền chia sẻ để giới hạn người có thể chỉnh sửa dữ liệu có trong file. Đây là một ví dụ về?

- A. Bảo mật dữ liệu (Data security)
- B. Đạo đức dữ liệu (Data ethics)
- C. Toàn vẹn dữ liệu (Data integrity)
- D. Xác nhận dữ liệu (Data validation)

Đáp án: A.

20. Sử dụng mã hóa để bảo vệ dữ liệu là một ví dụ về?

- A. Bảo mật dữ liệu (Data security)
- B. Đạo đức dữ liệu (Data ethics)
- C. Toàn vẹn dữ liệu (Data integrity)
- D. Xác nhận dữ liệu (Data validation)

Đáp án: A

21. Một nhà phân tích dữ liệu tạo một bảng tính có năm tab. Họ muốn chia sẻ dữ liệu trong tab 1-4 với khách hàng. Tab 5 chứa thông tin cá nhân về các máy khách khác. Chiến thuật nào sau đây sẽ giúp họ giữ tab 5 ở chế độ riêng tư? Chọn tất cả những câu phù hợp.

- A. Ẩn tab 5, sau đó chia sẻ bảng tính với khách hàng.
- B. Đổi tên tab 5 để thêm từ “riêng tư” vào tên file, sau đó chia sẻ bảng tính với khách hàng.
- C. Sao chép các tab 1-4 vào một bảng tính riêng biệt, sau đó chia sẻ file mới với khách hàng.
- D. Tạo một bản sao của bảng tính, xóa tab 5, sau đó chia sẻ file mới với khách hàng.

Đáp án: C,D

22. Một nhà phân tích dữ liệu muốn chia sẻ tab A của bảng tính với nhóm của mình. Anh ta vẫn đang làm việc với các tab B và C và chưa muốn các thành viên trong nhóm mình truy cập chúng. Ẩn các tab B và C sẽ bảo vệ chúng khỏi bị truy cập.

- A. Đúng
- B. Sai

Đáp án: B

23. Để giảm bớt sự lộn xộn, một nhà phân tích dữ liệu ẩn các ô chứa các công thức phức tạp và dài. Để xem lại các công thức này, họ sẽ cần điều chỉnh cài đặt mã hóa hoặc chia sẻ bảng tính.

- A. Đúng
- B. Sai

Đáp án: B

---