

Tài liệu đọc

Phân Tích Dữ Liệu

Khóa 5: Phân tích dữ liệu để trả lời câu hỏi

Phần 1: Tài liệu đọc bổ trợ

<u>Bài đọc 1</u>	Tổ chức dữ liệu để bắt đầu phân tích
	<div><div>1.1</div><div>Giữ cho dữ liệu có tổ chức bằng sắp xếp và bộ lọc<ul style="list-style-type: none">- Nhu cầu sắp xếp và lọc- Sắp xếp- Lọc- Sắp xếp trong bảng tóm tắt</div></div> <div><div>1.2</div><div>Tải tập dữ liệu Movie lên BigQuery</div></div> <div><div>1.3</div><div>Tài liệu tham khảo</div></div>
<u>Bài đọc 2</u>	Định dạng và điều chỉnh dữ liệu
	<div><div>2.1</div><div>Chuyển đổi dữ liệu trong bảng tính<ul style="list-style-type: none">- Định dạng kiểu ngày/tháng- Chuỗi thành số- Số thành phần trăm</div></div> <div><div>2.2</div><div>Chuyển đổi dữ liệu trong SQL<ul style="list-style-type: none">- Các chuyển đổi thông dụng- Các hàm biến đổi dữ liệu</div></div> <div><div>2.3</div><div>Chuẩn bị sử dụng tập dữ liệu bike sharing trong Bigquery</div></div> <div><div>2.4</div><div>Thao tác với chuỗi trong SQL<ul style="list-style-type: none">- Các hàm CONCAT</div></div>

	2.5 Tài liệu tham khảo
Bài đọc 3	Tổng hợp dữ liệu cho phân tích
	3.1 Các khái niệm chính về VLOOKUP
	3.2 Tải tập dữ liệu Movie lên Bigquery
	3.3 Tầm quan trọng của bí danh trong SQL
	- Cú pháp cơ bản cho bí danh
	- Sử dụng bí danh trong thực tế
	3.4 Sử dụng JOIN một cách hiệu quả
	- Cú pháp JOIN tổng quát
	- Các loại JOIN
	3.5 Tải tập dữ liệu Warehouse lên Bigquery
	3.6 Các hàm và câu truy vấn con trong SQL
	- Truy vấn con-quả anh đào ở trên cùng
	- Một số quy tắc mà truy vấn con phải tuân theo
	3.7 Tài liệu tham khảo
Bài đọc 4	Thực hiện tính toán dữ liệu
	4.1 Các hàm với nhiều điều kiện
	- Từ SUMIF đến SUMIFS
	- COUNTIF đến COUNTIFS
	4.2 Các yếu tố của bảng tóm tắt
	- Sử dụng bảng tóm tắt để phân tích
	4.3 Sử dụng bảng tóm tắt trong phân tích
	- Tạo bảng tóm tắt
	4.4 Tải tập dữ liệu Avocado lên Bigquery
	4.5 Các loại xác thực dữ liệu
	4.6 Thao tác với bảng tạm
	4.7 Tài liệu tham khảo

Phần 2: Hướng dẫn trả lời câu hỏi - Quiz

Phần 1

TÀI LIỆU ĐỌC BỔ TRỢ

Bài đọc 1: Tổ chức dữ liệu để bắt đầu phân tích

1. Giữ cho dữ liệu có tổ chức bằng sắp xếp và bộ lọc

Nhu cầu sắp xếp và lọc

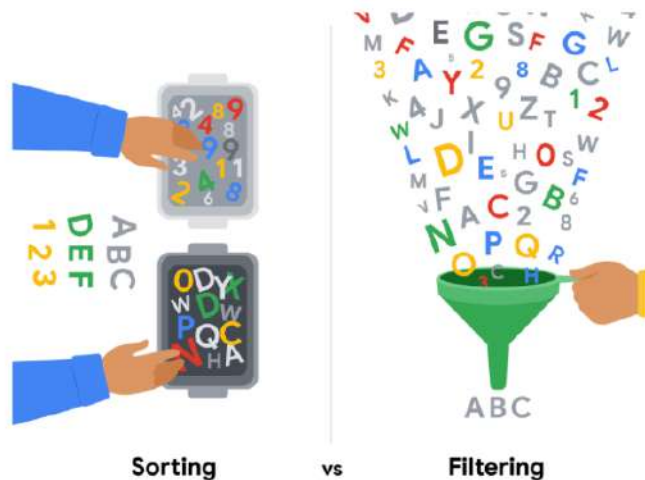
Bốn giai đoạn phân tích:

- Sắp xếp dữ liệu
- Định dạng và điều chỉnh dữ liệu
- Nhận thông tin đầu vào từ những người khác
- Chuyển đổi dữ liệu

Việc tổ chức các bộ dữ liệu thực sự quan trọng đối với các nhà phân tích dữ liệu. Hầu hết các tập dữ liệu bạn sử dụng được tổ chức dưới dạng bảng. Các bảng hữu ích vì chúng cho phép bạn thao tác và phân loại dữ liệu của mình. Bảng có các danh mục và phân loại riêng biệt, điều này cho phép bạn tập trung vào phần dữ liệu quan tâm và phân biệt giữa các dữ liệu của mình một cách nhanh chóng và dễ dàng.

Các nhà phân tích dữ liệu cũng cần định dạng và điều chỉnh dữ liệu khi thực hiện phân tích. **Sắp xếp (sorting)** và **lọc (filtering)** là hai cách bạn có thể tổ chức dữ liệu khi bạn thực hiện định dạng và điều chỉnh dữ liệu. Ví dụ: bộ lọc có thể giúp tìm ra lỗi hoặc ngoại lệ để bạn có thể sửa hoặc gỡ bỏ chúng trước khi phân tích. Các **điểm ngoại lệ (outlier)** là các điểm dữ liệu rất khác với dữ liệu được thu thập tương tự và có thể không phải là các giá trị đáng tin cậy. Lợi ích của việc lọc dữ liệu là sau khi bạn sửa lỗi hoặc xác định các ngoại lệ, bạn có thể xóa bộ lọc và trả dữ liệu về tổ chức ban đầu của nó.

Trong bài đọc này, bạn sẽ tìm hiểu sự khác biệt giữa sắp xếp và lọc. Bạn cũng sẽ được giới thiệu về cách thực hiện sắp xếp cụ thể trong bảng tổng hợp.



Sắp xếp

Sắp xếp là khi bạn tổ chức dữ liệu thành một thứ tự có ý nghĩa để dễ hiểu, dễ phân tích và dễ hình dung hơn. Dữ liệu của bạn sẽ được xếp hạng dựa trên một độ đo cụ thể mà bạn chọn. Bạn có thể sắp xếp dữ liệu trong bảng tính, cơ sở dữ liệu SQL (khi tập dữ liệu của bạn quá lớn đối với bảng tính) và các bảng trong tài liệu văn bản.

Ví dụ: nếu bạn cần xếp hạng hoặc tạo danh sách theo thứ tự thời gian, bạn có thể sắp xếp theo thứ tự tăng dần hoặc giảm dần. Nếu bạn muốn tìm ra những bộ phim yêu thích của một nhóm, bạn có thể sắp xếp theo tên phim. Sắp xếp sẽ tổ chức dữ liệu theo cách có ý nghĩa và cung cấp cho bạn thông tin chi tiết ngay lập tức. Sắp xếp cũng giúp bạn nhóm các dữ liệu tương tự lại với nhau. Đối với phim, bạn có thể sắp xếp theo thể loại - như hành động, chính kịch, khoa học viễn tưởng hoặc lãng mạn.

Lọc

Lọc được sử dụng khi bạn chỉ quan tâm đến dữ liệu đáp ứng một tiêu chí cụ thể và ẩn phần còn lại. Lọc thực sự hữu ích khi bạn có nhiều dữ liệu. Bạn có thể tiết kiệm thời gian bằng cách khai thác dữ liệu thực sự quan trọng hoặc dữ liệu có lỗi. Hầu hết các bảng tính và cơ sở dữ liệu SQL cho phép bạn lọc dữ liệu của mình theo nhiều cách khác nhau. Tính năng lọc cho phép bạn tìm thấy những gì bạn đang tìm kiếm mà không cần quá nhiều nỗ lực.

Ví dụ: nếu bạn chỉ quan tâm đến việc tìm ra những ai đã xem phim trong tháng 10, bạn có thể sử dụng bộ lọc theo ngày để chỉ hiển thị các bản ghi cho

các phim đã xem trong tháng 10. Sau đó, bạn có thể kiểm tra tên để biết ai đã xem phim trong tháng 10.

Tóm lại, cách dễ nhất để nhớ sự khác biệt giữa sắp xếp và lọc là: bạn có thể sử dụng sắp xếp để tổ chức nhanh dữ liệu và lọc để chỉ hiển thị dữ liệu đáp ứng tiêu chí bạn đã chọn. Sử dụng tính năng lọc khi bạn cần giảm lượng dữ liệu được hiển thị.

Điều quan trọng cần nhớ là, sau khi lọc dữ liệu, bạn cũng có thể **sắp xếp dữ liệu đã lọc**. Nếu bạn xem lại ví dụ về việc tìm ra những người đã xem phim vào tháng 10, sau khi bạn đã lọc các phim đã xem trong tháng 10, thì bạn có thể sắp xếp tên của những người đã xem những phim đó theo thứ tự bảng chữ cái.

Sắp xếp trong bảng tóm tắt

Các mục trong khu vực hàng và cột của **bảng tóm tắt (pivot table)** được sắp xếp theo thứ tự tăng dần bởi bất kỳ danh sách tùy chỉnh nào. Ví dụ: nếu danh sách của bạn chứa các ngày trong tuần, bảng tổng tóm tắt cho phép các tên ngày trong tuần và tháng sắp xếp như: Thứ Hai, Thứ Ba, Thứ Tư, v.v. thay vì theo thứ tự bảng chữ cái như sau: Thứ Bảy, Thứ Hai, Thứ Sáu, v.v.

Nếu các mục không có trong danh sách tùy chỉnh, chúng sẽ được sắp xếp theo thứ tự tăng dần mặc định. Tuy nhiên, nếu bạn sắp xếp theo thứ tự giảm dần, bạn đang thiết lập một quy tắc kiểm soát việc sắp xếp ngay cả sau khi các trường dữ liệu mới được thêm vào.

2. Tải tập dữ liệu Movie lên BigQuery

Một số ví dụ trong khóa học này sử dụng SQL để lọc dữ liệu trong một tập dữ liệu lớn trên BigQuery.

Nếu bạn muốn làm theo đúng các ví dụ minh họa, bạn cần đăng nhập vào tài khoản BigQuery của mình và tải lên tập dữ liệu movie được cung cấp dưới dạng file CSV. Nếu bạn chưa biết hoặc muốn ôn lại cách sử dụng BigQuery, xem lại khóa học **Chuẩn bị dữ liệu để khám phá**, nó sẽ bao gồm cách thiết lập tài khoản BigQuery.

Cách thức thực hiện:

Trước hết tải về tập dữ liệu dưới dạng file CSV tại [đây](#).

Bước 1: Mở bảng điều khiển BigQuery và nhấp vào dự án bạn muốn tải dữ liệu lên.

Bước 2: Trong Explorer ở bên trái, nhấp vào biểu tượng Action (ba chấm dọc) kế tên dự án của bạn và chọn **Create Dataset**.



Bước 3: Trong các ví dụ, tên "movie_data" sẽ được sử dụng cho tập dữ liệu. Nếu bạn định làm theo các ví dụ này, hãy nhập movie_data cho ID tập dữ liệu.

Create dataset

Dataset ID *

movie_data

Letters, numbers, and underscores allowed

Data location

Default



Default table expiration

☐ Enable table expiration ?

Default maximum table age

Days

Encryption

☒ Google-managed encryption key

No configuration required

☐ Customer-managed encryption key (CMEK)

Manage via Google Cloud Key Management Service

CREATE DATASET

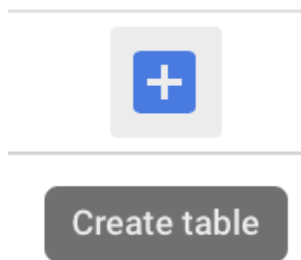
CANCEL

Bước 4: Nhấp vào **CREATE DATASET** (nút màu xanh) để thêm tập dữ liệu vào dự án của bạn.

Bước 5: Trong Explorer ở bên trái, nhấp để mở rộng dự án của bạn và sau đó nhấp vào tập dữ liệu **movie_data** bạn vừa tạo.

Bước 6: Nhấp vào biểu tượng Action (ba chấm dọc) bên cạnh **movie_data** và chọn **Open**.

Bước 7: Nhấp vào biểu tượng + màu xanh ở trên cùng bên phải để mở cửa sổ Create Table.



Bước 8: Trong Source, đối với mục Create Table, hãy chọn dữ liệu sẽ đến từ đâu.

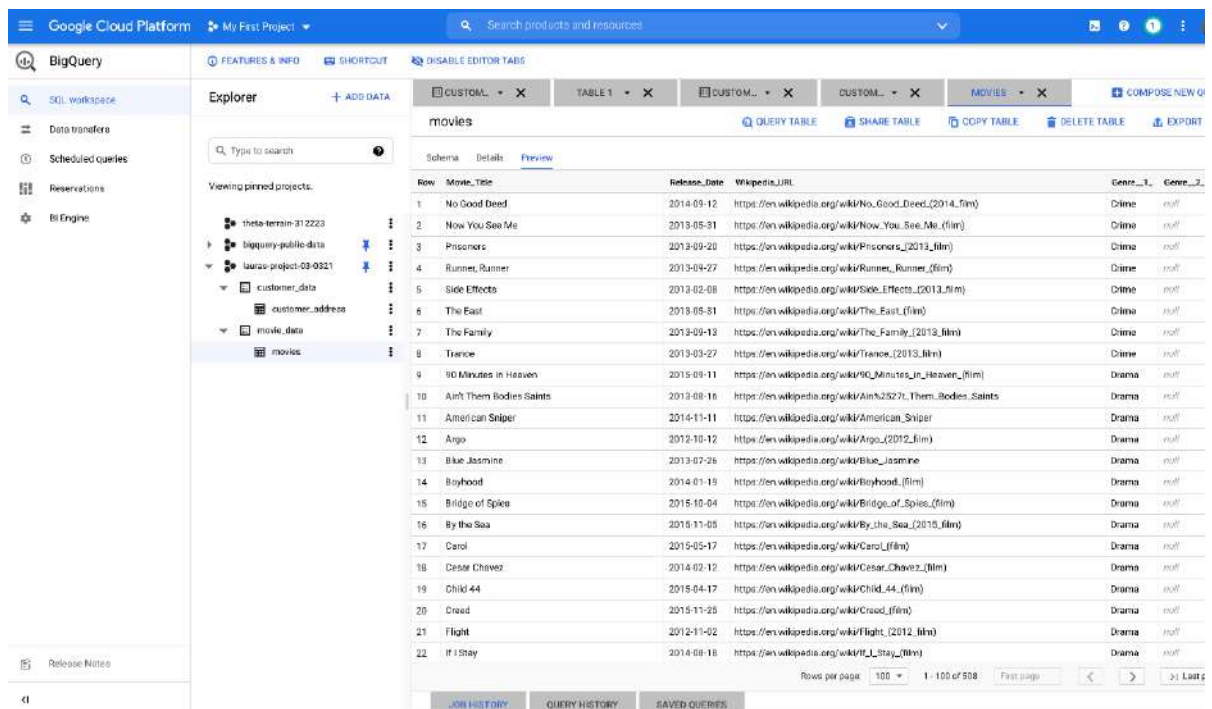
- Chọn **Upload**.
- Nhấp vào **Browse** để chọn file CSV Movie Data mà bạn đã tải xuống.
- Chọn **CSV** từ trình đơn thả xuống định dạng file.

Bước 9: Trong Destination, đối với Tên bảng, nhập **movies** để khớp với bảng trong ví dụ.

Bước 10: Đối với Schema, hãy chọn Auto Detect checkbox.

Bước 11: Nhấn **Create table** (nút màu xanh). Bây giờ bạn sẽ thấy bảng movie trong tập dữ liệu **movie_data** ở dự án của bạn.

Bước 12: Nhấp vào *movie* và sau đó chọn tab Preview. Xác nhận rằng bạn thấy dữ liệu được hiển thị bên dưới.



Row	Movie Title	Release Date	Wikipedia URL	Genre_1	Genre_2
1	No Good Deed	2014-09-12	https://en.wikipedia.org/wiki/No_Good_Deed_(2014_film)	Crime	null
2	Now You See Me	2013-05-31	https://en.wikipedia.org/wiki/Now_You_See_Me_(film)	Crime	null
3	Prisoners	2013-09-20	https://en.wikipedia.org/wiki/Prisoners_(2013_film)	Crime	null
4	Runes, Runes	2013-04-27	https://en.wikipedia.org/wiki/Runes,_Runes_(film)	Crime	null
5	Side Effects	2013-02-08	https://en.wikipedia.org/wiki/Side_Effects_(2013_film)	Crime	null
6	The East	2013-05-31	https://en.wikipedia.org/wiki/The_East_(film)	Crime	null
7	The Family	2013-09-13	https://en.wikipedia.org/wiki/The_Family_(2013_film)	Crime	null
8	Trance	2013-03-27	https://en.wikipedia.org/wiki/Trance_(2013_film)	Crime	null
9	90 Minutes in Heaven	2015-09-11	https://en.wikipedia.org/wiki/90_Minutes_in_Heaven_(film)	Drama	null
10	Ain't Them Bodies Saints	2013-08-16	https://en.wikipedia.org/wiki/Ain't_Them_Bodies_Saints	Drama	null
11	American Sniper	2014-11-11	https://en.wikipedia.org/wiki/American_Sniper	Drama	null
12	Argo	2012-10-12	https://en.wikipedia.org/wiki/Argo_(2012_film)	Drama	null
13	Blue Jasmine	2013-07-26	https://en.wikipedia.org/wiki/Blue_Jasmine	Drama	null
14	Boyhood	2014-01-19	https://en.wikipedia.org/wiki/Boyhood_(film)	Drama	null
15	Bridge of Spies	2015-10-04	https://en.wikipedia.org/wiki/Bridge_of_Spies_(film)	Drama	null
16	By the Sea	2015-11-05	https://en.wikipedia.org/wiki/By_the_Sea_(2015_film)	Drama	null
17	Carol	2015-05-17	https://en.wikipedia.org/wiki/Carol_(film)	Drama	null
18	Cesar Chavez	2014-02-12	https://en.wikipedia.org/wiki/Cesar_Chavez_(film)	Drama	null
19	Child 44	2015-04-17	https://en.wikipedia.org/wiki/Child_44_(film)	Drama	null
20	Cred	2015-11-25	https://en.wikipedia.org/wiki/Cred_(film)	Drama	null
21	Flight	2012-11-02	https://en.wikipedia.org/wiki/Flight_(2012_film)	Drama	null
22	If I Stay	2014-08-18	https://en.wikipedia.org/wiki/If_I_Stay_(film)	Drama	null

Đến đây, bạn đã có thể sẵn sàng sử dụng tập dữ liệu này.

3. Tài liệu tham khảo

[1]

<https://www.coursera.org/learn/analyze-data/supplement/RSNx9/keeping-data-organized-with-sorting-and-filters>

[2]

<https://www.coursera.org/learn/analyze-data/supplement/sBFZn/optional-upload-the-movie-dataset-to-bigquery>

Bài đọc 2: Định dạng và điều chỉnh dữ liệu

1. Chuyển đổi dữ liệu trong bảng tính

Trong bài đọc này, bạn sẽ tìm hiểu về cách **chuyển đổi dữ liệu** (converting data) từ định dạng này sang định dạng khác. Một trong những cách để đảm bảo rằng bạn thực hiện phân tích dữ liệu đúng đắn đó là đặt tất cả dữ liệu ở định dạng phù hợp. Điều này đúng ngay cả khi bạn đã làm sạch và xử lý dữ liệu của mình. Là một phần của việc chuẩn bị dữ liệu để phân tích, bạn sẽ cần phải chuyển đổi và định dạng dữ liệu của mình ngay từ đầu trong quá trình này.



Là một nhà phân tích dữ liệu, có rất nhiều tình huống khi bạn cần chuyển đổi dữ liệu trong bảng tính:

Định dạng kiểu ngày/tháng

Google Sheet lưu tất cả dữ liệu ngày tháng dưới dạng số nguyên. Không phải chuỗi các giá trị, ngày, tháng, năm, mà chỉ đơn giản là số nguyên.

- Số 1: Ngày 31/12/1899
- Số 2: Ngày 1/1/1900
- Số 102: ngày 11/4/1900 (100 ngày sau ngày 1/1/1900)
- -1: ngày 29/12/1899
- ...

Về thời gian (giờ/phút/giây) sẽ được thể hiện sau dấu thập phân.

- .00 – 12:00 AM
- .50 – 12:00 PM
- .125 – 3:00 AM
- ...

Kết hợp hai yếu tố trên:

31528.058 tương ứng với 26/4/1986, 1:23 AM

Chuyển định dạng bằng menu

Nếu các ô trong bảng tính của bạn có kiểu không phù hợp, hoặc không thống nhất, bạn có thể chuyển hóa như sau

Bước 1: Chọn các ô bạn muốn định dạng

Bước 2: Chọn Format > Number, chọn kiểu định dạng phù hợp

The screenshot shows the Google Sheets interface with the 'Format' menu open. The 'Number' option is selected, and a submenu is displayed showing various number formats. The 'Automatic' option is checked. Below it, several number formats are listed with their corresponding visual representations: Number (1,000.12), Percent (10.12%), Scientific (1.01E+03), Accounting (\$ (1,000.12)), Financial ((1,000.12)), Currency (\$1,000.12), Currency rounded (\$1,000), Date (9/26/2008), Time (3:59:00 PM), Date time (9/26/2008 15:59:00), Duration (24:01:00), 0% (123456%), Custom currency, Custom date and time, and Custom number format.

Format	Visual Representation
Automatic	✓ Automatic
Plain text	Plain text
Number	1,000.12
Percent	10.12%
Scientific	1.01E+03
Accounting	\$ (1,000.12)
Financial	((1,000.12))
Currency	\$1,000.12
Currency rounded	\$1,000
Date	9/26/2008
Time	3:59:00 PM
Date time	9/26/2008 15:59:00
Duration	24:01:00
0%	123456%
Custom currency	
Custom date and time	
Custom number format	

Ví dụ:

43783.125	11/14/2019	Date
43721.125	9/13/2019 3:00:00	DateTime
43783.125	3:00:00 AM	Time
43721.125	1049307:00:00	Duration

Các giá trị ở cột bên trái đã được chuẩn hóa thành kiểu ngày tháng tương ứng ở cột bên phải.

Chuyển chuỗi ngày tháng thành kiểu số

Với hàm DATEVALUE(data_text)



Chuỗi thành số

Dùng hàm Value:

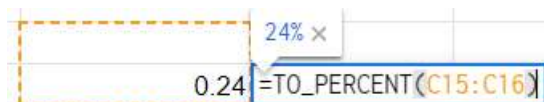
	A	B
1	Value stored as plain text	Numeric value
2	123	=VALUE(A2)
3	456	
4	789	
5	0	

Loại bỏ ký hiệu tiền tệ với hàm TO_PURE_NUMBER



Số thành phần trăm

Hàm TO_PERCENT chuyển dạng số dạng thành dạng phần trăm



Giá trị 1 sẽ tương ứng với 100%

2. Chuyển đổi dữ liệu trong SQL

Các nhà phân tích dữ liệu thường cần chuyển đổi dữ liệu từ định dạng này sang định dạng khác để hoàn thành phân tích. Nhưng điều gì sẽ xảy ra nếu bạn đang sử dụng SQL thay vì một bảng tính? Cũng giống như bảng tính, SQL sử dụng các quy tắc tiêu chuẩn để chuyển đổi loại dữ liệu này sang loại dữ liệu khác. Nếu bạn đang tự hỏi tại sao chuyển đổi dữ liệu lại là một kỹ năng quan trọng cần có với tư cách là một nhà phân tích dữ liệu, hãy nghĩ về nó giống như một người lái xe có thể thay một chiếc lốp bị xẹp. Khả năng chuyển đổi dữ liệu sang định dạng phù hợp sẽ giúp bạn tăng tốc độ trong quá trình phân tích của mình. Bạn không phải đợi người khác chuyển đổi dữ liệu cho bạn.



Trong bài đọc này, bạn sẽ học cách chuyển đổi được thực hiện bằng cách sử dụng hàm **CAST**. Ngoài ra còn có các hàm chuyên biệt hơn như **COERCION** để làm việc với các số lớn và **UNIX_DATE** để làm việc với ngày tháng. **UNIX_DATE** trả về số ngày đã trôi qua kể từ ngày 1 tháng 1 năm 1970 và được sử dụng để so sánh và làm việc với các ngày trên nhiều múi giờ. Bạn sẽ sử dụng **CAST** thường xuyên nhất.

Các chuyển đổi thông dụng

Bảng sau đây tóm tắt một số chuyển đổi thông dụng được thực hiện với hàm CAST. Tham khảo [Quy tắc chuyển đổi trong SQL chuẩn](#) để biết danh sách đầy đủ các hàm và luật liên quan.

Trong bài đọc này, chúng ta sẽ mô tả các tùy chọn để sắp xếp và lọc trong Google Sheet và Microsoft Excel. Cả hai đều cung cấp các chức năng sắp xếp và lọc cơ bản từ các menu đã cài đặt sẵn. Tuy nhiên, nếu bạn cần khả năng sắp xếp và lọc nâng cao hơn, bạn có thể sử dụng các hàm SORT và FILTER tương ứng của chúng.

Kiểu dữ liệu ban đầu	Hàm CAST có thể chuyển thành
Numeric (number)	<ul style="list-style-type: none"> - Integer - Numeric (number) - Big number - Floating integer - String
String	<ul style="list-style-type: none"> - Boolean - Integer - Numeric (number) - Big number - Floating integer - String - Bytes - Date - Date time - Time - Timestamp
Date	<ul style="list-style-type: none"> - String - Date - Date time - Timestamp

Hàm CAST (cú pháp và ví dụ)

CAST là một hàm của Viện Tiêu chuẩn Quốc gia Hoa Kỳ (ANSI) được sử dụng trong nhiều ngôn ngữ lập trình, bao gồm cả BigQuery. Phần này cung cấp cú pháp BigQuery và các ví dụ về chuyển đổi loại dữ liệu trong cột đầu tiên của bảng. Cú pháp của hàm CAST như sau:

```
CAST(expression AS typename)
```

Trong đó **expression** là dữ liệu được chuyển đổi và **typename** là kiểu dữ liệu được trả về.

Chuyển đổi một số thành một chuỗi

Câu lệnh **CAST** sau đây trả về một chuỗi từ một số được xác định bởi biến MyCount trong bảng MyTable.

```
SELECT CAST(MyCount AS STRING) FROM MyTable
```

Trong câu lệnh SQL ở trên:

- **SELECT** cho biết rằng bạn sẽ chọn dữ liệu từ một bảng
- **CAST** cho biết rằng bạn sẽ chuyển đổi dữ liệu bạn chọn sang một kiểu dữ liệu khác
- **AS** xuất hiện trước và xác định kiểu dữ liệu mà bạn đang chuyển tới
- **STRING** cho biết rằng bạn đang chuyển đổi dữ liệu thành một chuỗi
- **FROM** cho biết bạn đang chọn dữ liệu từ bảng nào

Chuyển đổi một chuỗi thành một số

Câu lệnh **CAST** sau đây trả về một số nguyên từ một chuỗi được xác định bởi biến MyVarcharCol trong bảng MyTable.

```
SELECT CAST(MyVarcharCol AS INT) FROM MyTable
```

Trong câu lệnh SQL ở trên:

- **SELECT** cho biết rằng bạn sẽ chọn dữ liệu từ một bảng
- **CAST** cho biết rằng bạn sẽ chuyển đổi dữ liệu bạn chọn sang một kiểu dữ liệu khác
- **AS** xuất hiện trước và xác định kiểu dữ liệu mà bạn đang chuyển tới
- **INT** cho biết rằng bạn đang chuyển đổi dữ liệu thành một số nguyên
- **FROM** cho biết bạn đang chọn dữ liệu từ bảng nào

Chuyển đổi ngày/tháng thành chuỗi

Câu lệnh **CAST** sau đây trả về một chuỗi từ một ngày được xác định bởi biến MyDate trong bảng được gọi là MyTable

```
SELECT CAST(MyDate AS STRING) FROM MyTable
```


Trong câu lệnh SQL ở trên:

- **SELECT** cho biết rằng bạn sẽ chọn dữ liệu từ một bảng
- **CAST** cho biết rằng bạn sẽ chuyển đổi dữ liệu bạn chọn sang một kiểu dữ liệu khác
 - **AS** xuất hiện trước và xác định kiểu dữ liệu mà bạn đang chuyển tới
 - **STRING** cho biết rằng bạn đang chuyển đổi dữ liệu thành một chuỗi
 - **FROM** cho biết bạn đang chọn dữ liệu từ bảng nào

Chuyển đổi ngày/tháng (date) thành ngày giờ (datetime)

Giá trị ngày giờ có định dạng là YYYY-MM-DD hh: mm: ss, vì vậy ngày và giờ được giữ lại cùng nhau. Câu lệnh CAST sau đây trả về một giá trị datetime từ date.

```
SELECT CAST (MyDate AS DATETIME) FROM MyTable
```

Trong câu lệnh SQL ở trên:

- **SELECT** cho biết rằng bạn sẽ chọn dữ liệu từ một bảng
- **CAST** cho biết rằng bạn sẽ chuyển đổi dữ liệu bạn chọn sang một kiểu dữ liệu khác
 - **AS** xuất hiện trước và xác định kiểu dữ liệu mà bạn đang chuyển tới
 - **DATETIME** cho biết rằng bạn đang chuyển đổi dữ liệu thành kiểu ngày giờ
- **FROM** cho biết bạn đang chọn dữ liệu từ bảng nào

Hàm SAFE_CAST

Sử dụng hàm **CAST** trong truy vấn không thành công sẽ trả về lỗi trong BigQuery. Để tránh lỗi trong trường hợp truy vấn không thành công, hãy sử dụng hàm **SAFE_CAST** để thay thế. Hàm **SAFE_CAST** trả về giá trị Null thay vì lỗi khi truy vấn không thành công.

Cú pháp cho **SAFE_CAST** giống như cho **CAST**. Chỉ cần thay thế hàm trực tiếp trong các truy vấn của bạn. Câu lệnh **SAFE_CAST** sau đây trả về một chuỗi từ một ngày.

```
SELECT SAFE_CAST(MyDate AS STRING) FROM MyTable
```

Tài liệu tham khảo khác

Các tài liệu này cung cấp thêm thông tin về chuyển đổi dữ liệu bằng các phương ngữ SQL khác (thay vì BigQuery):

- [CAST và CONVERT](#): Tài liệu tham khảo SQL Server
- [Các hàm và toán tử MySQL CAST](#): Tài liệu tham khảo MySQL
- [Cách thực hiện: Chuyển đổi kiểu SQL](#): Blog về chuyển kiểu có liên kết đến các hướng dẫn ngắn về SQL khác

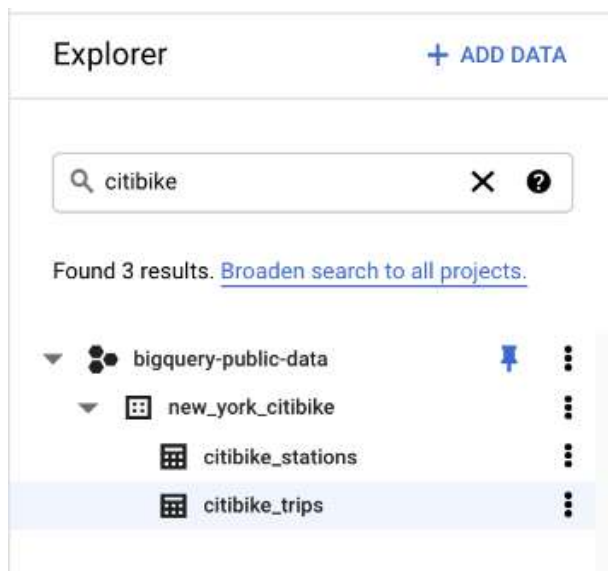
3. Chuẩn bị sử dụng tập dữ liệu bike sharing trong BigQuery

Một số ví dụ trong khóa học này sử dụng CONCAT trong truy vấn SQL để trả về dữ liệu từ hai cột thành một cột.

Nếu bạn muốn làm theo các ví dụ tương ứng, bạn sẽ cần đăng nhập vào tài khoản BigQuery của mình để sử dụng tập dữ liệu mở (công khai) có tên **new_york_citibike**. Nếu bạn cần tham khảo lại cách Sử dụng BigQuery, xem lại khóa học **Chuẩn bị dữ liệu để khám phá**, trong đó có giải thích cách thiết lập tài khoản BigQuery.

Cách sử dụng tập dữ liệu

Bước 1: Trong BigQuery Explorer, nhập **citibike** vào thanh tìm kiếm để tìm tập dữ liệu **new_york_citibike** trong **bigquery-public-data**.



Bước 2: Bấm vào bảng **citibike_trips**, sau đó bấm vào tab Preview để xem dữ liệu trong bảng.

*UNSAVE...

CITIBIKE...

CITIBIKE...

COMPOSE NEW QUERY

citibike_trips

QUERY

SHARE

COPY

DELETE

EXPORT

SCHEMA

DETAILS

PREVIEW

sn_id	start_station_name	start_station_latitude	start_station_longitude	end_station_id	end_station_name
509	9 Ave & W 22 St	40.7454973	-74.00197139	442	W 27 St & 7 Ave
280	E 10 St & 5 Ave	40.73331967	-73.99510132	254	W 11 St & 6 Ave
335	Washington Pl & Broadway	40.72903917	-73.99404649	540	Lexington Ave & E 29 St
146	Hudson St & Reade St	40.71625008	-74.0091059	387	Centre St & Chambers St
529	W 42 St & 8 Ave	40.7575699	-73.99098507	352	W 56 St & 6 Ave
470	W 20 St & 8 Ave	40.74345335	-74.00004031	252	MacDougal St & Washington Sq
3158	W 63 St & Broadway	40.77163851	-73.98261428	3167	Amsterdam Ave & W 73 St
519	Pershing Square N	40.75188406	-73.97770164	147	Greenwich St & Warren St
470	W 20 St & 8 Ave	40.74345335	-74.00004031	496	E 16 St & 5 Ave
487	E 20 St & FDR Drive	40.73314259	-73.97573881	487	E 20 St & FDR Drive
291	Madison St & Montgomery St	40.713126	-73.984844	3489	Gold St & Frankfort St

Những gì mong đợi từ truy vấn

Bạn sẽ sử dụng CONCAT để kết hợp dữ liệu trong cột **start_station_name** với dữ liệu trong cột **end_station_name** để tạo thông tin tuyến đường trong một cột khác; ví dụ: tuyến đường từ Ga 509 đến Ga 442 ở hàng đầu tiên của bảng trên sẽ là **9 Ave & W 22 St đến W 27 St & 7 Ave**, là sự kết hợp của tên ga đầu và cuối.

4. Thao tác với chuỗi trong SQL

Biết cách chuyển đổi và sử dụng dữ liệu để có một phân tích chính xác là một phần quan trọng trong công việc của nhà phân tích dữ liệu. Trong bài đọc này, bạn sẽ tìm hiểu về các hàm SQL khác nhau và cách sử dụng chúng, đặc biệt liên quan đến việc kết hợp các chuỗi.

Các hàm CONCAT

Chuỗi là một tập hợp các ký tự giúp khai báo các văn bản trong các ngôn ngữ lập trình như SQL. Các hàm chuỗi SQL được sử dụng để lấy các thông tin

khác nhau về các ký tự hoặc trong trường hợp này là thao tác chúng. Một hàm như vậy, CONCAT, thường được sử dụng. Xem lại bảng bên dưới để tìm hiểu thêm về hàm CONCAT và các biến thể của nó.

Hàm	Sử dụng	Ví dụ
CONCAT	Một hàm cộng các chuỗi với nhau để tạo chuỗi văn bản mới có thể được sử dụng làm khóa duy nhất	CONCAT ('Google', '.com');
CONCAT_WS	Hàm cộng hai hay nhiều chuỗi với nhau với một dấu phân tách	CONCAT_WS (' . ', 'www', 'google', 'com') *Dấu phân tách (dấu chấm .) sẽ được thêm vào trước và sau chuỗi google để tạo thành www.google.com
CONCAT với +	Cộng hai hay nhiều chuỗi với toán tử cộng	'Google' + '.com'

Ví dụ sử dụng CONCAT

Khi cộng hai chuỗi với nhau, chẳng hạn như "Data" và "analysis", ta nhập như sau:

- SELECT CONCAT ('Data', 'analysis');
- Kết quả sẽ là:
- Data Analysis

Tùy thuộc vào các chuỗi mình cần, bạn sẽ cần thêm một ký tự khoảng trắng, vì vậy hàm của bạn thực sự phải là:

- SELECT CONCAT ('Data', ' ', 'analysis');
- Và kết quả sẽ là:
- Data analysis

Quy tắc tương tự cũng áp dụng khi kết hợp ba chuỗi với nhau. Ví dụ,

- SELECT CONCAT ('Data', ' ', 'analysis', ' ', 'is', ' ', 'awesome!');
- Và kết quả sẽ là
- Data analysis is awesome

Tài liệu tham khảo khác

W3 School là một tài nguyên tuyệt vời để học SQL một cách tương tác và các liên kết sau sẽ hướng dẫn bạn cách chuyển đổi dữ liệu của mình bằng SQL:

- [Hàm SQL](#): Đây là danh sách đầy đủ các hàm để bạn tham khảo. Nhấp vào từng hàm, bạn sẽ tìm hiểu về định nghĩa, cách sử dụng, ví dụ và thậm chí có thể tạo và chạy truy vấn của riêng bạn để thực hành. Tự mình thử nó xem!

- [Từ khóa SQL](#): Đây là một tham chiếu từ khóa SQL hữu ích khi bạn muốn nâng cao kiến thức của mình về SQL. Danh sách từ khóa này là những từ dành riêng mà bạn sẽ sử dụng nếu bạn cần thực hiện các hành động khác nhau trong cơ sở dữ liệu.

- Mặc dù bài đọc này đã đi qua các khái niệm cơ bản của từng hàm, nhưng vẫn còn nhiều điều cần tìm hiểu và thậm chí bạn có thể kết hợp các chuỗi của riêng mình.

- ✓ Thực hành sử dụng [CONCAT](#)
- ✓ Thực hành sử dụng [CONCAT WS](#)
- ✓ Thực hành sử dụng [CONCAT với +](#)

Lưu ý: Các chức năng được trình bày trong các tài liệu ở trên có thể được áp dụng theo những cách hơi khác nhau tùy thuộc vào cơ sở dữ liệu mà bạn đang sử dụng (ví dụ: mySQL so với SQL Server). Tuy nhiên, mô tả chung được cung cấp cho mỗi hàm sẽ cho phép bạn tùy chỉnh cách bạn sử dụng các hàm này khi cần thiết.

5. Tài liệu tham khảo

[1]<https://www.coursera.org/learn/analyze-data/supplement/H7GTe/converting-data-in-spreadsheets>

[2]<https://www.coursera.org/learn/analyze-data/supplement/HqeNj/transforming-data-in-sql>

[3]

<https://www.coursera.org/learn/analyze-data/supplement/vYBY0/optional-prepare-to-use-the-bike-sharing-dataset-in-bigquery>

[4]<https://www.coursera.org/learn/analyze-data/supplement/oitMs/manipulating-strings-in-sql>

Bài đọc 3: Tổng hợp dữ liệu cho phân tích

1. Các khái niệm chính về VLOOKUP

Các hàm có thể được sử dụng để nhanh chóng tìm kiếm thông tin và thực hiện các phép tính bằng cách sử dụng các giá trị cụ thể. Trong bài đọc này, bạn sẽ tìm hiểu về tầm quan trọng của một hàm như vậy, VLOOKUP hoặc Vertical Lookup, hàm này tìm kiếm một giá trị nhất định trong cột bảng tính và trả về một phần thông tin tương ứng từ hàng mà giá trị tìm kiếm được tìm thấy.

Khi nào bạn cần sử dụng hàm VLOOKUP?

Hai lý do phổ biến để sử dụng hàm VLOOKUP là:

- Nhập dữ liệu trong bảng tính
- Hợp nhất dữ liệu từ một bảng tính với dữ liệu trong một bảng tính khác

Cú pháp VLOOKUP

Hàm Vlookup có sẵn trong cả Microsoft Excel và Google Sheet. Bạn sẽ được giới thiệu về cú pháp chung trong Google Sheet. (Bạn có thể tham khảo tài nguyên ở cuối bài đọc này để biết thêm thông tin về hàm VLOOKUP trong Microsoft Excel.)

```
VLOOKUP(10003, A2:B26, 2, FALSE)
```

Cú pháp:

```
VLOOKUP(search_key, range, index, [is_sorted])
```

Search_key

- Giá trị cần tìm kiếm.
- Ví dụ: 42, "Cats" hoặc I24.

range

- Phạm vi cần xem xét cho việc tìm kiếm.
- Cột đầu tiên trong phạm vi dùng để định vị dữ liệu khớp với giá trị được chỉ định bởi *search_key*.

index

- Chỉ mục cột của giá trị được trả về, trong đó cột đầu tiên trong phạm vi được đánh số 1.
- Nếu chỉ mục không nằm trong khoảng từ 1 đến số cột trong phạm vi, thì **#VALUE!** được trả về.

is_sorted

- Cho biết cột được tìm kiếm (cột đầu tiên của phạm vi được chỉ định) có được sắp xếp hay không. Mặc định là TRUE.
- Bạn nên đặt *is_sorted* thành FALSE. Nếu được đặt thành FALSE, một kết quả khớp chính xác sẽ được trả về. Nếu có nhiều giá trị phù hợp, nội dung của ô tương ứng với giá trị đầu tiên được tìm thấy sẽ được trả về và **#N/A** được trả về nếu không tìm thấy giá trị đó.
- Nếu *is_sorted* là TRUE hoặc bị bỏ qua, kết quả khớp gần nhất (nhỏ hơn hoặc bằng khóa tìm kiếm) sẽ được trả về. Nếu tất cả các giá trị trong cột tìm kiếm lớn hơn khóa tìm kiếm, thì **#N/A** được trả về.

Điều gì xảy ra nếu bạn nhận được giá trị #N/A

Như đã nói ở phần trên, #N/A chỉ ra rằng không thể trả về giá trị khớp cho kết quả của hàm VLOOKUP. Lỗi không có nghĩa là dữ liệu thực sự sai, nhưng mọi người có thể thắc mắc nếu họ thấy lỗi trong báo cáo. Bạn có thể sử dụng hàm IFNA để thay thế lỗi **#N/A** bằng lỗi mô tả hơn, chẳng hạn như “Không tồn tại”.

IFNA(#N/A, “Does not exist”)

Cú pháp như sau:

IFNA(value, value_if_na)

value

- Đây là một giá trị bắt buộc.
- Hàm kiểm tra xem giá trị ô có khớp với giá trị không: chẳng hạn như #N/A.

value_if_na

- Đây là một giá trị bắt buộc.

- Hàm trả về giá trị này nếu giá trị ô khớp với giá trị trong đối số đầu tiên; nó trả về giá trị này khi giá trị ô là **#N/A**.

Lưu ý liên quan đến hàm VLOOKUP

- TRUE nghĩa là đối sánh gần đúng, FALSE có nghĩa là đối sánh chính xác trên khóa tìm kiếm. Nếu dữ liệu được sử dụng cho khóa tìm kiếm được sắp xếp, TRUE có thể được sử dụng.
- Bạn muốn cột khớp với khóa tìm kiếm trong công thức VLOOKUP ở bên trái dữ liệu. VLOOKUP chỉ xem dữ liệu ở bên phải sau khi tìm thấy kết quả khớp. Nói cách khác, chỉ mục cho VLOOKUP chỉ cho biết các cột ở bên phải. Điều này có thể yêu cầu bạn di chuyển các cột trước khi bạn sử dụng hàm VLOOKUP.
- Sau khi bạn đã điền dữ liệu bằng công thức VLOOKUP, bạn chỉ có thể sao chép và dán dữ liệu dưới dạng giá trị (paste as value) để xóa công thức, cho phép bạn thao tác lại dữ liệu.

2. Tải tập dữ liệu Movie lên BigQuery

Một số ví dụ trong khóa học trình bày cách sử dụng JOINS để hợp nhất và trả về dữ liệu từ hai bảng dựa trên một thuộc tính chung được sử dụng trong cả hai bảng.

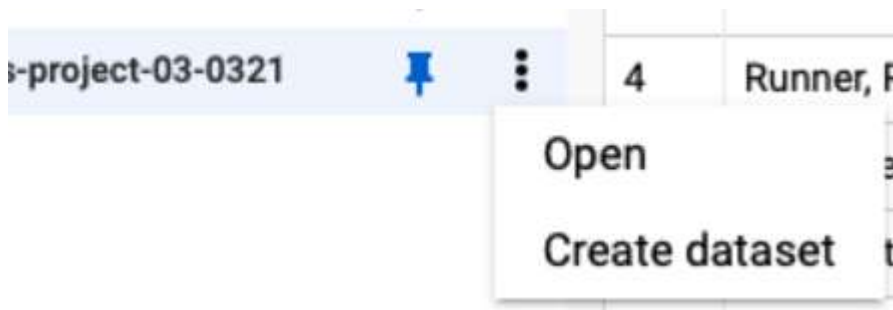
Nếu bạn muốn làm theo đúng các ví dụ minh họa, bạn cần đăng nhập vào tài khoản BigQuery của mình và tải lên tập dữ liệu movie được cung cấp dưới dạng file CSV. Nếu bạn chưa biết hoặc muốn ôn lại cách sử dụng BigQuery, xem lại khóa học Chuẩn bị dữ liệu để khám phá, nó sẽ bao gồm cách thiết lập tài khoản BigQuery.

Cách thức thực hiện:

Trước hết tải về tập dữ liệu dưới dạng file CSV: [bảng Employee](#) và [bảng Department](#)

Bước 1: Mở bảng điều khiển BigQuery và nhấp vào dự án bạn muốn tải dữ liệu lên.

Bước 2: Trong Explorer ở bên trái, nhấp vào biểu tượng Action (ba chấm dọc) kế tên dự án của bạn và chọn **Create Dataset**.



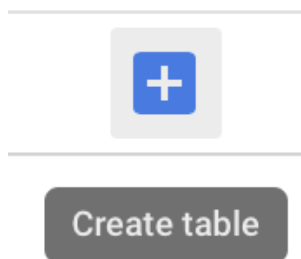
Bước 3: Nhập vào **employee_data** trong trường Dataset ID

Bước 4: Nhấp vào **CREATE DATASET** (nút màu xanh) để thêm tập dữ liệu vào dự án của bạn.

Bước 5: Trong Explore ở bên trái, hãy nhấp để mở rộng dự án của bạn, sau đó nhấp vào tập dữ liệu **employee_data** bạn vừa tạo.

Bước 6: Nhấp vào biểu tượng Action (ba chấm dọc) bên cạnh **employee_data** và chọn Open.

Bước 7: Nhấp vào biểu tượng + màu xanh ở trên cùng bên phải để mở cửa sổ Tạo bảng.



Bước 8: Trong Source, đối với mục Create Table, hãy chọn dữ liệu sẽ đến từ đâu.

- Chọn **Upload**.
- Nhấp vào **Browse** để chọn file CSV Employee Table mà bạn đã tải xuống.
- Chọn **CSV** từ trình đơn thả xuống định dạng file.

Bước 9: Trong mục tên Table, hãy nhập vào **employees** nếu bạn định làm theo như ví dụ.

Bước 10: Trong Schema, hãy chọn Auto Detect checkbox.

Bước 11: Nhấn **Create table** (nút màu xanh). Bây giờ bạn sẽ thấy bảng **employee** trong tập dữ liệu **employee_data** ở dự án của bạn.

Bước 12: Nhấp lại vào tập dữ liệu **employee_data**.

Bước 13: Nhấp vào biểu tượng để mở lại cửa sổ Create Table.

Bước 14: Trong Source, đối với mục Create Table, hãy chọn dữ liệu sẽ đến từ đâu.

- Chọn **Upload**.
- Nhấp vào **Browse** để chọn file CSV Department Table mà bạn đã tải xuống.
- Chọn **CSV** từ trình đơn thả xuống định dạng file.

Bước 15: Trong mục tên Table, hãy nhập vào **departments** nếu bạn định làm theo như ví dụ.

Bước 16: Trong Schema, hãy chọn Auto Detect checkbox.

Bước 17: Nhấn **Create table** (nút màu xanh). Bây giờ bạn sẽ thấy bảng **departments** trong tập dữ liệu **employee_data** ở dự án của bạn.

Bước 18: Nhấp vào bảng **employees** và nhấp vào tab Preview để xác nhận rằng bạn có dữ liệu được hiển thị như bên dưới.

employees
⋮
 DELETE
 EXPORT

SCHEMA
DETAILS
PREVIEW

Row	name	department_id	role
1	Dave Smith	1	Product Marketing Manager
2	Scott Tanner	1	Director of Demand Gen
3	Margaret Lane	1	VP of Marketing
4	Julie Jones	2	Software Engineer
5	Ted Connors	2	Software Engineer
6	Mary Martin	5	Receptionist

Bước 19: Nhấp vào bảng **departments** và nhấp vào tab Preview để xác minh rằng bạn có dữ liệu được hiển thị như bên dưới.

departments			
SCHEMA		DETAILS	PREVIEW
Row	name	department_id	
1	Marketing	1	
2	Engineering	2	
3	Accounting	3	
4	Sales	4	

Đến đây, bạn đã có thể sẵn sàng sử dụng tập dữ liệu này.

3. Tầm quan trọng của bí danh trong SQL

Trong bài đọc này, bạn sẽ tìm hiểu về cách sử dụng **bí danh (alias)** để đơn giản hóa các truy vấn SQL của mình. Bí danh được sử dụng trong các truy vấn SQL để tạo tên tạm cho một cột hoặc bảng. Bí danh làm cho các bảng và cột tham chiếu trong truy vấn SQL của bạn đơn giản hơn nhiều khi bạn có tên bảng hoặc cột quá dài hoặc phức tạp để sử dụng trong các truy vấn. Hãy tưởng tượng một tên bảng như sau: *special_projects_customer_negotiation_mileages*. Rất khó để gõ lại tên mỗi khi bạn sử dụng bảng này. Với một bí danh, bạn có thể tạo một biệt hiệu có ý nghĩa mà bạn có thể sử dụng để phân tích. Trong trường hợp này, “*special_projects_customer_negotiation_mileages*” có thể được đặt bí danh đơn giản là “*mileages*”. Thay vì phải viết ra bảng tên dài, bạn có thể sử dụng một biệt hiệu có ý nghĩa do bạn quyết định.

Cú pháp cơ bản cho bí danh

Trong các truy vấn SQL, bí danh được thực hiện bằng cách sử dụng lệnh **AS**. cú pháp cơ bản cho lệnh **AS** có thể được nhìn thấy trong truy vấn sau để đặt bí danh cho bảng:

```
SELECT column_name(s)
FROM table_name AS alias_name;
```

Chú ý rằng AS đứng trước tên bảng và theo sau là biệt hiệu mới. Một cách tiếp cận tương tự để tạo bí danh cho một cột:

```
SELECT column_name AS alias_name
FROM table_name;
```

Trong cả hai trường hợp, bạn có một tên mới mà bạn có thể sử dụng để tham chiếu đến cột hoặc bảng được đặt bí danh.

Cú pháp thay thế cho bí danh

Nếu việc sử dụng AS dẫn đến lỗi khi chạy truy vấn vì cơ sở dữ liệu SQL bạn đang làm việc không hỗ trợ nó, bạn có thể loại bỏ AS. Trong các ví dụ trước, cú pháp thay thế cho bí danh một bảng hoặc cột sẽ là:

- FROM table_name alias_name
- SELECT column_name alias_name

Bạn có thể dùng AS hoặc không để tạo bí danh. Tuy nhiên, sử dụng AS có lợi ích là làm cho các truy vấn dễ đọc hơn. Nó giúp làm cho bí danh nổi bật hơn, rõ ràng hơn.

Sử dụng bí danh trong thực tế

Hãy xem ví dụ về truy vấn SQL sử dụng bí danh. Giả sử bạn đang làm việc với hai bảng: một trong số đó có dữ liệu **employee** và bảng còn lại có dữ liệu **department**. Câu lệnh FROM với bí danh các bảng đó có thể là:

```
FROM work_day.employees AS employees
```

Những bí danh này vẫn cho bạn biết chính xác những gì có trong các bảng, nhưng giờ đây bạn không phải nhập các tên bảng dài đó theo cách thủ công. Bí danh thực sự hữu ích cho các truy vấn phức tạp và dài. Việc đọc và viết các truy vấn của bạn sẽ dễ dàng hơn khi bạn có các bí danh cho bạn biết những gì được bao gồm trong các bảng của bạn.

4. Sử dụng JOIN một cách hiệu quả

Trong bài đọc này, bạn sẽ xem lại cách các JOIN được sử dụng và sẽ được giới thiệu một số tài liệu mà bạn có thể sử dụng để tìm hiểu thêm về chúng. JOIN kết hợp các bảng bằng cách sử dụng khóa chính hoặc khóa ngoại để căn chỉnh thông tin đến từ cả hai bảng trong quá trình kết hợp. JOIN sử dụng các khóa này để xác định các mối quan hệ và các giá trị tương ứng giữa các bảng.

Nếu bạn cần xem lại về khóa chính và khóa ngoại, hãy tham khảo bảng thuật ngữ cho khóa học này hoặc quay lại bài **Cơ sở dữ liệu trong phân tích dữ liệu**.

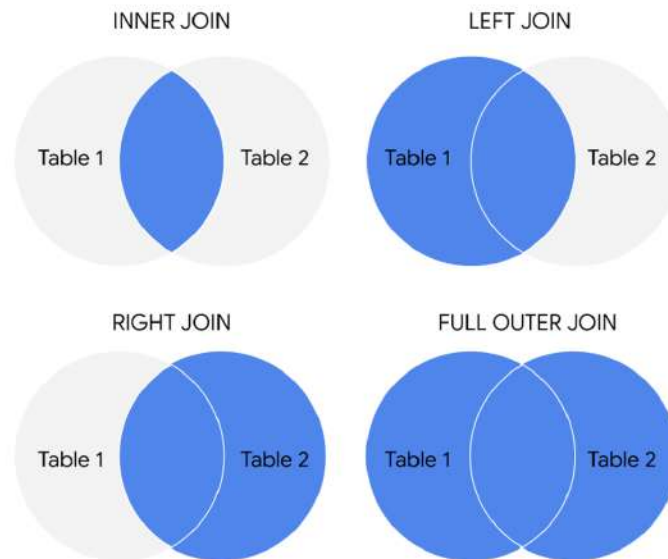
Cú pháp JOIN tổng quát

```
SELECT
    -- table columns from tables are inserted here
    table_name1.column_name
    table_name2.column_name
FROM
    table_name1
JOIN
    table_name2
ON table_name1.column_name = table_name2.column_name
```

Như bạn có thể thấy từ cú pháp, câu lệnh JOIN là một phần của mệnh đề FROM trong truy vấn. JOIN trong SQL chỉ ra rằng bạn sẽ kết hợp dữ liệu từ hai bảng. ON trong SQL xác định cách khớp các bảng để có thông tin chính xác được kết hợp từ cả hai.

Các loại JOIN

Có bốn cách chung để thực hiện các lệnh JOIN trong truy vấn SQL: INNER, LEFT, RIGHT và FULL OUTER



Sau đây là cách các câu truy vấn JOIN khác nhau thực hiện

INNER JOIN

Bạn có thể dùng **JOIN** hay **INNER JOIN** trong câu truy vấn SQL này, vì đây là loại JOIN mặc định cũng như câu lệnh JOIN thông dụng nhất. INNER JOIN trả về các bản ghi nếu dữ liệu nằm trong cả hai bảng. Ví dụ: nếu bạn sử dụng INNER JOIN cho bảng "*customers*" và "*order*" và khớp dữ liệu bằng cách sử dụng khóa *customer_id*, bạn sẽ kết hợp dữ liệu cho mỗi *customer_id* tồn tại trong cả hai bảng. Nếu *customer_id* tồn tại trong bảng khách hàng nhưng không tồn tại trong bảng đơn đặt hàng, thì dữ liệu cho *customer_id* đó sẽ không được kết hợp hoặc trả về bởi câu truy vấn.

```
SELECT
    customers.customer_name,
    orders.product_id,
    orders.ship_date
FROM
    customers
INNER JOIN
    orders
ON customers.customer_id = orders.customer_id
```

Kết quả từ truy vấn có thể trông giống như sau, trong đó `customer_name` từ bảng `customers` và `product_id` và `ship_date` từ bảng `orders`:

customer_name	product_id	ship_date
Martin's Ice Cream	043998	2021-02-23
Beachside Treats	872012	2021-02-25
Mona's Natural Flavors	724956	2021-02-28
... etc.	... etc.	... etc.

Dữ liệu từ cả hai bảng được kết hợp với nhau bằng cách so khớp `customer_id` chung cho cả hai bảng. Lưu ý rằng `customer_id` không hiển thị trong kết quả truy vấn. Nó chỉ đơn giản được sử dụng để thiết lập mối quan hệ giữa dữ liệu trong hai bảng để dữ liệu có thể được nối và trả về.

LEFT JOIN

Bạn có thể dùng **LEFT OUTER JOIN**, nhưng hầu hết người dùng thích xài **LEFT JOIN**. Cả hai đều đúng cú pháp. **LEFT JOIN** trả về tất cả các bản ghi từ bảng bên trái và chỉ các bản ghi phù hợp từ bảng bên phải. Sử dụng **LEFT JOIN** bất cứ khi nào bạn cần dữ liệu từ toàn bộ bảng đầu tiên và các giá trị từ bảng thứ hai, nếu chúng tồn tại. Ví dụ: trong truy vấn bên dưới, **LEFT JOIN** sẽ trả về `customer_name` với `sales_rep` tương ứng, nếu nó có sẵn. Nếu có một khách hàng không tương tác với đại diện bán hàng, khách hàng đó sẽ vẫn hiển thị trong kết quả truy vấn nhưng với giá trị `NULL` cho `sales_rep`.

```
SELECT
    customers.customer_name,
    sales.sales_rep
FROM
    customers
LEFT JOIN
    sales
ON customers.customer_id = sales.customer_id
```

Kết quả từ truy vấn có thể trông giống như sau trong đó `customer_name` là từ bảng `customers` và `sales_rep` là từ bảng `sales`. Một lần nữa, dữ liệu từ cả hai bảng được kết hợp với nhau bằng cách so khớp `customer_id` chung cho cả hai bảng mặc dù `customer_id` không được trả về trong kết quả truy vấn.

customer_name	sales_rep
Martin's Ice Cream	Luis Reyes
Beachside Treats	NULL
Mona's Natural Flavors	Geri Hall
...etc.	...etc.

RIGHT JOIN

Bạn có thể dùng **RIGHT OUTER JOIN** hoặc **RIGHT JOIN**. **RIGHT JOIN** trả về tất cả các bản ghi từ bảng bên phải và các bản ghi tương ứng từ bảng bên trái. Thực tế mà nói, **RIGHT JOIN** hiếm khi được sử dụng. Hầu hết mọi người chỉ cần thay đổi thứ tự bảng và dùng **LEFT JOIN**. Nhưng sử dụng cùng ví dụ cho **LEFT JOIN** ở phần trước, truy vấn sử dụng **RIGHT JOIN** sẽ trông giống như sau:


```
SELECT
    sales.sales_rep,
    customers.customer_name
FROM
    sales
RIGHT JOIN
    customers
ON sales.customer_id = customers.customer_id
```

Kết quả cuối cùng sẽ giống như LEFT JOIN

customer_name	sales_rep
Martin's Ice Cream	Luis Reyes
Beachside Treats	NULL
Mona's Natural Flavors	Geri Hall
...etc.	...etc.

FULL OUTER JOIN

Bạn cũng có thể dùng **FULL JOIN**. **FULL OUTER JOIN** trả về tất cả các bản ghi từ các bảng được chỉ định. Bạn có thể kết hợp các bảng theo cách này, nhưng hãy lưu ý rằng dữ liệu sau khi nối có thể rất lớn. **FULL OUTER JOIN** trả về tất cả các bản ghi từ cả hai bảng ngay cả khi dữ liệu không được điền vào một trong các bảng. Ví dụ: trong truy vấn bên dưới, bạn sẽ nhận được tất cả khách hàng và ngày vận chuyển sản phẩm của họ. Bởi vì bạn đang sử dụng **FULL OUTER JOIN**, bạn có thể có thông tin khách hàng mà không có ngày giao hàng tương ứng hoặc ngày giao hàng mà không có khách hàng tương ứng. Giá trị **NULL** được trả về nếu dữ liệu tương ứng không tồn tại trong một trong hai bảng.

```
SELECT
  customers.customer_name,
  orders.ship_date
FROM
  customers
FULL OUTER JOIN
  orders
ON customers.customer_id = orders.customer_id
```

Kết quả từ truy vấn có thể trông giống như sau

customer_name	ship_date
Martin's Ice Cream	2021-02-23
Beachside Treats	2021-02-25
NULL	2021-02-25
The Daily Scoop	NULL
Mountain Ice Cream	NULL
Mona's Natural Flavors	2021-02-28
...etc.	...etc.

5. Tải tập dữ liệu warehouse lên BigQuery

Một số ví dụ trong khóa học trình bày cách sử dụng COUNT và COUNT DISTINCT trong SQL để đếm và trả về số lượng giá trị nhất định trong tập dữ liệu.

Nếu bạn muốn làm theo đúng các ví dụ minh họa, bạn cần đăng nhập vào tài khoản BigQuery của mình và tải lên tập dữ liệu movie được cung cấp dưới dạng file CSV. Nếu bạn chưa biết hoặc muốn ôn lại cách sử dụng BigQuery,

xem lại khóa học **Chuẩn bị dữ liệu để khám phá**, nó sẽ bao gồm cách thiết lập tài khoản BigQuery.

Cách thức thực hiện:

Trước hết tải về tập dữ liệu warehouse dưới dạng hai file CSV: [bảng Warehouse](#) và [bảng Orders](#)

Bước 1: Mở bảng điều khiển BigQuery và nhấp vào dự án bạn muốn tải dữ liệu lên.

Bước 2: Trong Explorer ở bên trái, nhấp vào biểu tượng Action (ba chấm dọc) kế tên dự án của bạn và chọn **Create Dataset**.



Bước 3: Trong các ví dụ, tên "warehouse_orders" sẽ được sử dụng cho tập dữ liệu. Nếu bạn dự định làm theo các ví dụ, hãy nhập **warehouse_orders** cho ID tập dữ liệu.

Create dataset

Dataset ID *

warehouse_orders

Letters, numbers, and underscores allowed

Data location

Default

Default table expiration

☐ Enable table expiration ?

Default maximum table age

Days

Encryption

☒ Google-managed encryption key

No configuration required

☐ Customer-managed encryption key (CMEK)

Manage via Google Cloud Key Management Service

CREATE DATASET

CANCEL

Bước 4: Nhấp vào **CREATE DATASET** (nút màu xanh) để thêm tập dữ liệu vào dự án của bạn.

Bước 5: Trong Explore ở bên trái, hãy nhấp để mở rộng dự án của bạn, sau đó nhấp vào tập dữ liệu **warehouse_orders** bạn vừa tạo.

Bước 6: Nhấp vào biểu tượng Action (ba chấm dọc) bên cạnh **warehouse_orders** và chọn Open.

Bước 7: Nhấp vào biểu tượng + màu xanh ở trên cùng bên phải để mở cửa sổ Tạo bảng.



Create table

Bước 8: Trong Source, đối với mục Create Table, hãy chọn dữ liệu sẽ đến từ đâu.

- Chọn **Upload**.
- Nhấp vào **Browse** để chọn file CSV Warehouse mà bạn đã tải xuống.
- Chọn **CSV** từ trình đơn thả xuống định dạng file.

Bước 9: Trong mục tên Table, hãy nhập vào **Warehouse** nếu bạn định làm theo như ví dụ.

Bước 10: Trong Schema, hãy chọn Auto Detect checkbox.

Bước 11: Nhấn **Create table** (nút màu xanh). Bây giờ bạn sẽ thấy bảng **Warehouse** trong tập dữ liệu **warehouse_orders** ở dự án của bạn.

Bước 12: Nhấp lại vào tập dữ liệu **warehouse_orders**.

Bước 13: Nhấp vào biểu tượng để mở lại cửa sổ Create Table.

Bước 14: Trong Source, đối với mục Create Table, hãy chọn dữ liệu sẽ đến từ đâu.

- Chọn **Upload**.
- Nhấp vào **Browse** để chọn file CSV Orders Table mà bạn đã tải xuống.
- Chọn **CSV** từ trình đơn thả xuống định dạng file.

Bước 15: Trong mục tên Table, hãy nhập vào **Orders** nếu bạn định làm theo như ví dụ.

Bước 16: Trong Schema, hãy chọn Auto Detect checkbox.

Bước 17: Nhấn **Create table** (nút màu xanh). Bây giờ bạn sẽ thấy bảng **Orders** trong tập dữ liệu **warehouse_orders** ở dự án của bạn.

Bước 18: Nhấp vào bảng **Warehouse** và nhấp vào tab Preview để xác nhận rằng bạn có dữ liệu được hiển thị như bên dưới.

SCHEMA DETAILS <u>PREVIEW</u>					
Row	warehouse_id	warehouse_alias	maximum_capacity	employee_total	state
1	1543	Somerset Fulfillment Center	210	14	KY
2	2270	Bowling Green Warehouse	280	13	KY
3	9080	Frankfort Fulfillment Center	235	5	KY
4	2666	Lansing Fulfillment Center	290	16	MI
5	3961	Lansing Storage Warehouse	740	22	MI
6	8118	Ann Arbor Fulfillment Center	780	17	MI
7	3417	Gatlinburg Warehouse	620	6	TN
8	4338	Knoxville Fulfillment Center	215	13	TN
9	6509	Memphis Fulfillment Center	755	22	TN
10	9831	Clarsvill Warehouse	400	16	TN

Bước 19: Nhấp vào bảng **Orders** và nhấp vào tab Preview để xác minh rằng bạn có dữ liệu được hiển thị như bên dưới.

SCHEMA DETAILS <u>PREVIEW</u>					
Row	order_id	customer_id	warehouse_id	order_date	shipper_date
1	789	3731	8118	2019-01-01	2019-01-04
2	790	3486	8118	2019-01-01	2019-01-04
3	791	2623	8118	2019-01-01	2019-01-04
4	792	9869	8118	2019-01-01	2019-01-04
5	793	6866	8118	2019-01-01	2019-01-04
6	794	8055	8118	2019-01-01	2019-01-04
7	795	1152	8118	2019-01-01	2019-01-04
8	796	5765	8118	2019-01-01	2019-01-04
9	797	6709	8118	2019-01-01	2019-01-04
10	798	4866	2666	2019-01-01	2019-01-04
11	799	4515	2666	2019-01-01	2019-01-04
12	800	9618	2666	2019-01-01	2019-01-04
13	801	2337	2666	2019-01-01	2019-01-04
14	802	1166	2666	2019-01-01	2019-01-04
15	803	4376	2666	2019-01-01	2019-01-04
16	804	9832	2666	2019-01-01	2019-01-04
17	805	6046	9080	2019-01-01	2019-01-04

Rows per page: 100 1 - 100 of 9999

Đến đây, bạn đã có thể sẵn sàng sử dụng tập dữ liệu này.

6. Các hàm và câu truy vấn con trong SQL

Trong bài đọc này, bạn sẽ tìm hiểu về các hàm SQL và cách chúng được sử dụng với các truy vấn con. **Hàm SQL** là các công cụ được tích hợp sẵn trong SQL để có thể thực hiện các phép tính. **Truy vấn con** (còn được gọi là truy vấn bên trong hoặc truy vấn lồng nhau) là một truy vấn bên trong một truy vấn khác.

Truy vấn con - quả anh đào ở trên cùng

Hãy coi một câu truy vấn như một chiếc bánh. Một chiếc bánh có thể có nhiều lớp bên trong nó và thậm chí nhiều lớp bên trong những lớp đó. Mỗi lớp này là các truy vấn con và khi bạn đặt tất cả các lớp lại với nhau, bạn sẽ nhận được một chiếc bánh (câu truy vấn). Thông thường, bạn sẽ tìm thấy các truy vấn con được lồng trong các mệnh đề SELECT, FROM và WHERE. Không có cú pháp chung cho các truy vấn con, nhưng cú pháp cho một truy vấn con cơ bản như sau:

```
SELECT account_table.*  
FROM (  
    SELECT *  
    FROM transaction.sf_model_feature_2014_01  
    WHERE day_of_week = 'Friday'  
    ) account_table  
WHERE account_table.availability = 'YES'
```

Bạn thấy rằng, trong mệnh đề SELECT đầu tiên là một mệnh đề SELECT khác. Mệnh đề SELECT thứ hai đánh dấu sự bắt đầu của truy vấn con trong câu lệnh này. Có nhiều cách khác nhau mà bạn có thể sử dụng các truy vấn con và các tài nguyên được trích dẫn sẽ cung cấp thêm hướng dẫn khi bạn tìm hiểu. Nhưng trước hết, hãy tóm tắt lại các quy tắc truy vấn con.

Một số quy tắc mà truy vấn con phải tuân theo:

- Truy vấn con phải được đặt trong dấu ngoặc đơn
- Một truy vấn con chỉ có thể có một cột được chỉ định trong mệnh đề SELECT. Nhưng nếu bạn muốn một truy vấn con để so sánh nhiều cột, các cột đó phải được chọn trong truy vấn chính.
 - Truy vấn con trả về nhiều hơn một hàng chỉ có thể được sử dụng với các toán tử đa giá trị, chẳng hạn như toán tử IN cho phép bạn chỉ định đa giá trị trong mệnh đề WHERE.
 - Không thể lồng một truy vấn con trong lệnh SET. Lệnh SET được sử dụng với UPDATE để chỉ định các cột (và giá trị) nào sẽ được cập nhật trong bảng.

7. Tài liệu tham khảo

- [1]<https://www.coursera.org/learn/analyze-data/supplement/SNmqP/vlookup-core-concepts>
- [2]<https://www.coursera.org/learn/analyze-data/supplement/13KQO/optional-upload-the-employee-dataset-to-bigquery>
- [3]<https://www.coursera.org/learn/analyze-data/supplement/qURXP/secret-identities-the-importance-of-aliases>
- [4]<https://www.coursera.org/learn/analyze-data/supplement/DBOi7/using-joins-effectively>
- [5]<https://www.coursera.org/learn/analyze-data/supplement/HuXCc/optional-upload-the-warehouse-dataset-to-bigquery>
- [6]<https://www.coursera.org/learn/analyze-data/supplement/SthVU/sql-functions-and-subqueries-a-functional-friendship>

Bài đọc 4: Thực hiện tính toán dữ liệu

1. Các hàm với nhiều điều kiện

Trong bài đọc này, bạn sẽ tìm hiểu thêm về các hàm điều kiện và cách xây dựng các hàm có nhiều điều kiện. Nhắc lại rằng các hàm và công thức có điều kiện thực hiện các phép tính theo các điều kiện cụ thể. Trước đây, bạn đã học cách sử dụng các hàm như **SUMIF** và **COUNTIF** có một điều kiện. Bạn có thể sử dụng các hàm **SUMIFS** và **COUNTIFS** nếu bạn có hai điều kiện trở lên. Bạn sẽ tìm hiểu cú pháp cơ bản của chúng trong Google Sheet và xem ví dụ.

Tham khảo tài liệu ở cuối bài đọc này để biết thông tin về các hàm tương tự trong Microsoft Excel.

Từ SUMIF đến SUMIFS

Cú pháp cơ bản của hàm SUMIF là: **=SUMIF(range, criterion, sum_range)**

- **range**: nơi hàm sẽ tìm kiếm điều kiện mà bạn đã đặt.
- **criterion**: điều kiện bạn đang áp dụng
- **sum_range**: phạm vi ô sẽ được bao gồm trong phép tính.

Ví dụ, bạn có một bảng với danh sách các khoản chi phí (expense), giá trị của chúng (Price) và ngày chúng xảy ra (Date).

	A	B	C
1	Expense	Price	Date
2	Fuel	\$48.00	12/14/2020
3	Food	\$12.34	12/14/2020
4	Taxi	\$21.57	12/14/2020
5	Coffee	\$2.50	12/15/2020
6	Fuel	\$36.00	12/15/2020
7	Taxi	\$15.88	12/15/2020
8	Coffee	\$4.15	12/15/2020
9	Food	\$6.75	12/15/2020

Bạn có thể sử dụng SUMIF để tính tổng giá nhiên liệu trong bảng này, như sau:

A11			=SUMIF(A1:A9, "Fuel", B1:B9)
-----	---	---	------------------------------

Tuy nhiên, bạn cũng có thể xây dựng trong nhiều điều kiện bằng cách sử dụng hàm SUMIFS. SUMIF và SUMIFS rất giống nhau, nhưng SUMIFS có thể bao gồm nhiều điều kiện.

Cú pháp cơ bản như sau:

=SUMIFS(sum_range, criteria_range1, criterion1, [criteria_range2, criterion2, ...])

Dấu ngoặc vuông cho bạn biết rằng đây là tùy chọn. Dấu chấm lửng ở cuối câu lệnh cho bạn biết rằng bạn có thể lặp lại nhiều lần các tham số này nếu cần. Ví dụ: nếu bạn muốn tính tổng chi phí nhiên liệu cho một ngày trong bảng này, bạn có thể tạo câu lệnh SUMIFS với nhiều điều kiện, như sau:

A12			=SUMIFS(B1:B9, A1:A9, "Fuel", C1:C9, "12/15/2020")
-----	---	---	--

Công thức này cung cấp cho bạn tổng chi phí của mọi chi phí nhiên liệu kể từ ngày được liệt kê trong điều kiện. Trong ví dụ này, C1:C9 là dãy tiêu chí thứ hai và ngày 15/12/2020 là điều kiện thứ hai. Miễn là bạn tuân theo cú pháp cơ bản, bạn có thể thêm tới 127 điều kiện vào một câu lệnh SUMIFS!

COUNTIF đến COUNTIFS

Cũng giống như hàm SUMIFS, COUNTIFS cho phép bạn tạo một hàm COUNTIF với nhiều điều kiện.

Cú pháp:

=COUNTIF(range, criterion)

Cũng giống như SUMIF, bạn đặt phạm vi (**range**) và sau đó đặt điều kiện (**criterion**) cần được đáp ứng. Ví dụ: nếu bạn muốn đếm số lần Food xuất hiện trong cột Expense, bạn có thể sử dụng hàm COUNTIF như sau:

A13 ▾ | fx | =COUNTIF(A1:A9, "Food")

COUNTIFS có cùng cú pháp cơ bản như SUMIFS:

=COUNTIFS(criteria_range1, criterion1, [criteria_range2, criterion2, ...])

Phạm vi tiêu chí (criteria_range) và tiêu chí (criterion) theo cùng một thứ tự và bạn có thể thêm nhiều điều kiện hơn vào cuối hàm. Vì vậy, nếu bạn muốn tìm số lần Coffee xuất hiện trong cột Expense vào ngày 15/12/2020, bạn có thể sử dụng COUNTIFS để áp dụng các điều kiện đó, như sau:

A14 ▾ | fx | =COUNTIFS(A1:A9, "Coffee", C1:C9, "12/15/2020")

Công thức này tuân theo cú pháp cơ bản để tạo điều kiện cho "Coffee" và ngày cụ thể. Bây giờ chúng ta có thể tìm thấy mọi trường hợp mà cả hai điều kiện này đều thỏa.

2. Các yếu tố của bảng tóm tắt (pivot table)

Bạn đã biết rằng bảng tóm tắt là một công cụ được sử dụng để sắp xếp, tổ chức lại, nhóm, đếm, tính tổng hoặc trung bình dữ liệu trong bảng tính. Trong bài đọc này, bạn sẽ tìm hiểu thêm về các phần của bảng tóm tắt và cách các nhà phân tích dữ liệu sử dụng chúng để tóm tắt dữ liệu và trả lời các câu hỏi về dữ liệu của họ.

Bảng tóm tắt (pivot table) giúp bạn có thể xem dữ liệu theo nhiều cách để xác định thông tin chi tiết và xu hướng. Chúng có thể giúp bạn nhanh chóng hiểu được các tập dữ liệu lớn hơn bằng cách so sánh các chỉ số, thực hiện tính toán và tạo báo cáo. Chúng cũng hữu ích để trả lời các câu hỏi cụ thể về dữ liệu của bạn.

Bảng tóm tắt có bốn phần cơ bản: hàng (rows), cột (columns), giá trị (values) và bộ lọc (filters).

Rows

Add

Columns

Add

Values

Add

Filters

Add

Tùy chọn **hàng** của bảng tóm tắt sắp xếp và nhóm dữ liệu bạn chọn theo chiều ngang. Ví dụ: trong slide nói về bảng tóm tắt, các giá trị Release Date được sử dụng để tạo các hàng, nhóm dữ liệu theo năm.

Release Date - Year

2012

2013

2014

2015

2016

Các **cột** sắp xếp và hiển thị các giá trị từ dữ liệu của bạn theo chiều dọc. Tương tự như hàng, cột có thể được kéo trực tiếp từ tập dữ liệu hoặc được tạo bằng cách sử dụng các **giá trị**.

Giá trị được sử dụng để tính toán và đếm dữ liệu. Đây là nơi bạn nhập các biến mà bạn muốn đo lường. Đây cũng là cách bạn tạo các trường được tính

toán (**calculated field**) trong bảng tóm tắt của mình. Nhắc lại, trường được tính toán là trường mới trong bảng tóm tắt thực hiện các phép tính nhất định dựa trên giá trị của các trường khác

Trong ví dụ về *movie data* trước đó, phần Values đã tạo các cột cho bảng tóm tắt, bao gồm SUM của Box Office Revenue, AVERAGE của Box Office Revenue và COUNT của các cột Box Office Revenue.

SUM of Box Office Revenue (\$)	AVERAGE of Box Office Revenue (\$)	COUNT of Box Office Revenue (\$)
\$18,078,040,000.00	\$170,547,547.17	106
\$13,672,800,000.00	\$160,856,470.59	85
\$20,013,420,000.00	\$168,180,000.00	119
\$13,521,310,000.00	\$109,042,822.58	124
\$11,921,900,000.00	\$161,106,756.76	74
\$77,207,470,000.00	\$151,983,208.66	508

Cuối cùng, phần **bộ lọc (filter)** của bảng tóm tắt cho phép bạn áp dụng bộ lọc dựa trên các tiêu chí cụ thể - giống như bộ lọc trong bảng tính thông thường! Ví dụ: một bộ lọc đã được thêm vào bảng tóm tắt dữ liệu phim để nó chỉ bao gồm các phim tạo ra doanh thu dưới 10 triệu đô la.

Release Date - Year	SUM < \$10M	AVERAGE < \$10 M	COUNT < \$10 M
2012	\$18,078,040,000.00	\$170,547,547.17	106
2013	\$13,672,800,000.00	\$160,856,470.59	85
2014	\$20,013,420,000.00	\$168,180,000.00	119
2015	\$13,521,310,000.00	\$109,042,822.58	124
2016	\$11,921,900,000.00	\$161,106,756.76	74
Grand Total	\$77,207,470,000.00	\$151,983,208.66	508

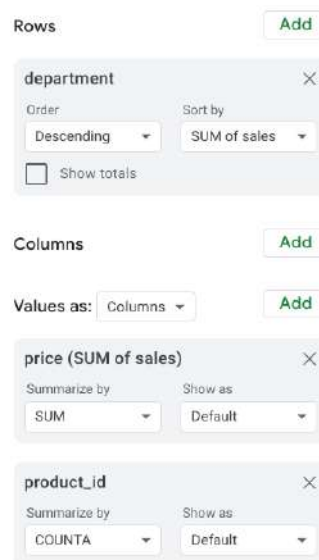
Có thể sử dụng tất cả bốn phần bảng tóm tắt sẽ cho phép bạn so sánh các chỉ số khác nhau từ dữ liệu của mình và thực hiện các phép tính, điều này sẽ giúp bạn có được những thông tin chi tiết có giá trị.

Sử dụng bảng tóm tắt để phân tích

Bảng tóm tắt là một công cụ hữu ích để trả lời các câu hỏi cụ thể về tập dữ liệu để bạn có thể nhanh chóng chia sẻ câu trả lời với các bên liên quan. Ví dụ,

một nhà phân tích dữ liệu làm việc tại một cửa hàng bách hóa được yêu cầu xác định tổng doanh số cho từng bộ phận (department) và số lượng sản phẩm mà mỗi bộ phận đã bán. Họ cũng quan tâm đến việc biết chính xác bộ phận nào tạo ra nhiều doanh thu nhất.

Thay vì thực hiện các thay đổi đối với dữ liệu bảng tính ban đầu, họ đã sử dụng bảng tóm tắt để trả lời những câu hỏi này và dễ dàng so sánh doanh thu bán hàng và số lượng sản phẩm bán ra của từng bộ phận.



The screenshot shows the configuration interface for a Google Data Studio report. It includes three main sections: Rows, Columns, and Values. The Rows section has a card for 'department' with 'Order' set to 'Descending' and 'Sort by' set to 'SUM of sales'. The Columns section has a card for 'price (SUM of sales)' with 'Summarize by' set to 'SUM' and 'Show as' set to 'Default'. The Values section has a card for 'product_id' with 'Summarize by' set to 'COUNTA' and 'Show as' set to 'Default'. Each section has an 'Add' button.

Rows Add

department ×

Order: Descending ▼ Sort by: SUM of sales ▼

☐ Show totals

Columns Add

Values as: Columns ▼ Add

price (SUM of sales) ×

Summarize by: SUM ▼ Show as: Default ▼

product_id ×

Summarize by: COUNTA ▼ Show as: Default ▼

Họ đã sử dụng bộ phận (department) làm các hàng cho bảng tóm tắt để nhóm và sắp xếp phần còn lại của dữ liệu bán hàng. Sau đó, họ nhập hai Values dưới dạng cột: tổng doanh số và tổng số sản phẩm đã bán. Họ cũng sắp xếp dữ liệu theo cột SUM của doanh thu để xác định bộ phận nào tạo ra nhiều doanh thu nhất.

department	SUM of sales	COUNTA of product_id
Toys	\$3,045.95	49
Beauty	\$2,958.37	55
Movies	\$2,880.55	57
Tools	\$2,869.96	48
Games	\$2,785.80	49
Industrial	\$2,728.90	51
Jewelry	\$2,669.25	52
Health	\$2,613.56	48
Automotive	\$2,589.56	47
Garden	\$2,586.92	45
Sports	\$2,460.12	46
Grocery	\$2,459.22	44
Outdoors	\$2,399.48	47
Electronics	\$2,353.65	41
Books	\$2,313.01	42
Baby	\$2,272.77	46
Home	\$2,222.99	38
Computers	\$2,206.20	44
Kids	\$2,117.98	41
Shoes	\$2,108.33	39
Clothing	\$1,858.47	38
Music	\$1,809.36	33

Bây giờ họ biết rằng bộ phận Toys tạo ra nhiều doanh thu nhất!

Bảng tóm tắt là một công cụ hiệu quả cho các nhà phân tích dữ liệu làm việc với bảng tính vì chúng làm nổi bật những thông tin chi tiết chính từ dữ liệu bảng tính mà không cần phải thực hiện các thay đổi đối với bảng tính. Trong tương lai, bạn sẽ tạo bảng tóm tắt của riêng mình để phân tích dữ liệu và xác định các xu hướng sẽ có giá trị cao đối với các bên liên quan

3. Sử dụng bảng tóm tắt trong phân tích

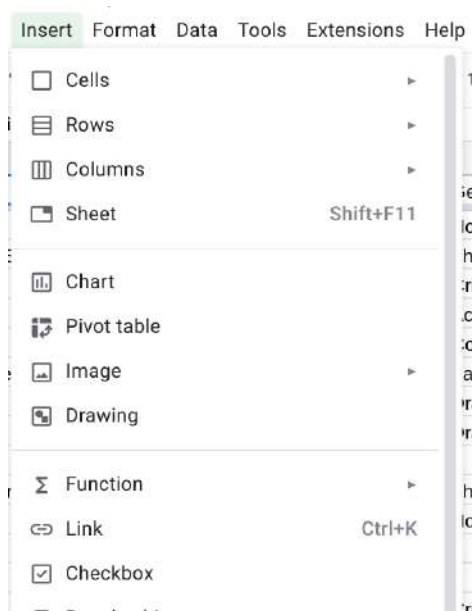
Trong bài đọc này, bạn sẽ học cách tạo và sử dụng bảng tóm tắt để phân tích dữ liệu. **Bảng tóm tắt (pivot table)** là một công cụ bảng tính cho phép bạn xem dữ liệu theo nhiều cách để tìm thông tin chi tiết và xu hướng.

Bảng tóm tắt cho phép bạn hiểu các tập dữ liệu lớn bằng cách cung cấp cho bạn các công cụ để dễ dàng so sánh các chỉ số, nhanh chóng thực hiện các phép tính và tạo các báo cáo có thể đọc được. Bạn có thể tạo bảng tóm tắt để giúp bạn trả lời các câu hỏi cụ thể về dữ liệu của mình. Ví dụ: nếu bạn đang phân tích dữ liệu bán hàng, bạn có thể sử dụng bảng tóm tắt để trả lời các câu hỏi như "Tháng nào có doanh số bán hàng nhiều nhất?" và "Sản phẩm nào tạo ra doanh thu cao nhất trong năm nay?" Khi bạn cần câu trả lời cho các câu hỏi về dữ liệu của mình, bảng tóm tắt có thể giúp bạn loại bỏ sự lộn xộn và chỉ tập trung vào dữ liệu bạn cần.

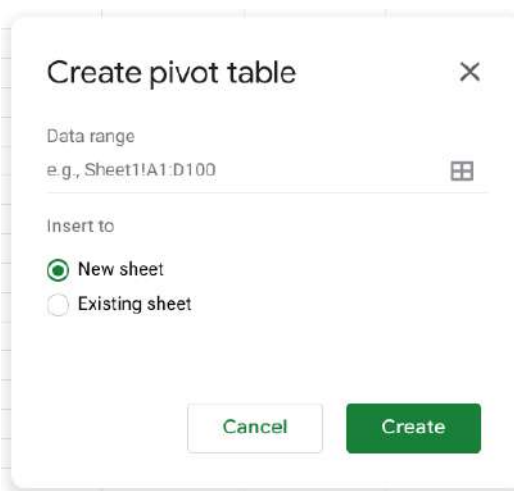
Tạo bảng tóm tắt

Trước khi có thể phân tích dữ liệu bằng bảng tóm tắt, bạn sẽ cần tạo một bảng tóm tắt với dữ liệu của mình. Phần sau bao gồm các bước để tạo bảng tóm tắt trong Google Sheet, nhưng hầu hết các chương trình bảng tính sẽ có các công cụ tương tự.

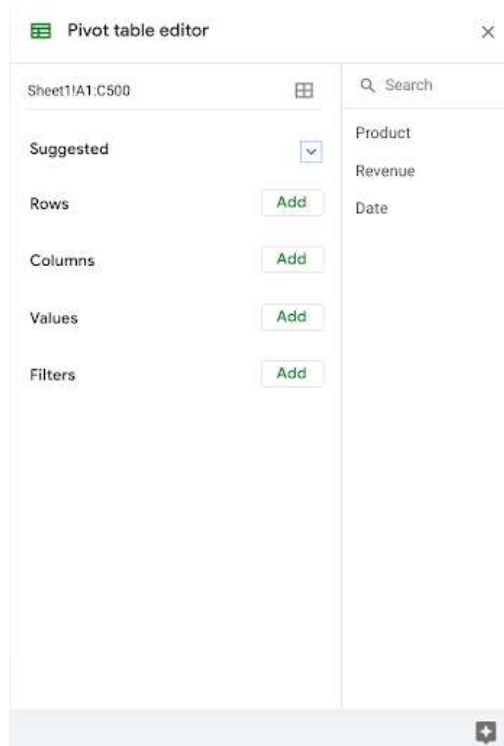
Đầu tiên, bạn sẽ mở menu Insert từ thanh công cụ, sẽ có một tùy chọn cho Pivot table.



Menu bật lên này sẽ xuất hiện:



Nói chung, bạn nên tạo một trang tính mới cho bảng tóm tắt của mình để giữ dữ liệu thô và phân tích của bạn tách biệt. Bạn cũng có thể lưu trữ tất cả các phép tính của mình ở một nơi để dễ dàng tham khảo. Khi bạn đã tạo bảng tóm tắt, sẽ có một trình chỉnh sửa bảng tóm tắt mà bạn có thể truy cập ở bên phải dữ liệu của mình.



Đây là nơi bạn sẽ có thể tùy chỉnh bảng tóm tắt của mình, bao gồm những biến bạn muốn đưa vào phân tích.

4. Tải tập dữ liệu avocado lên BigQuery

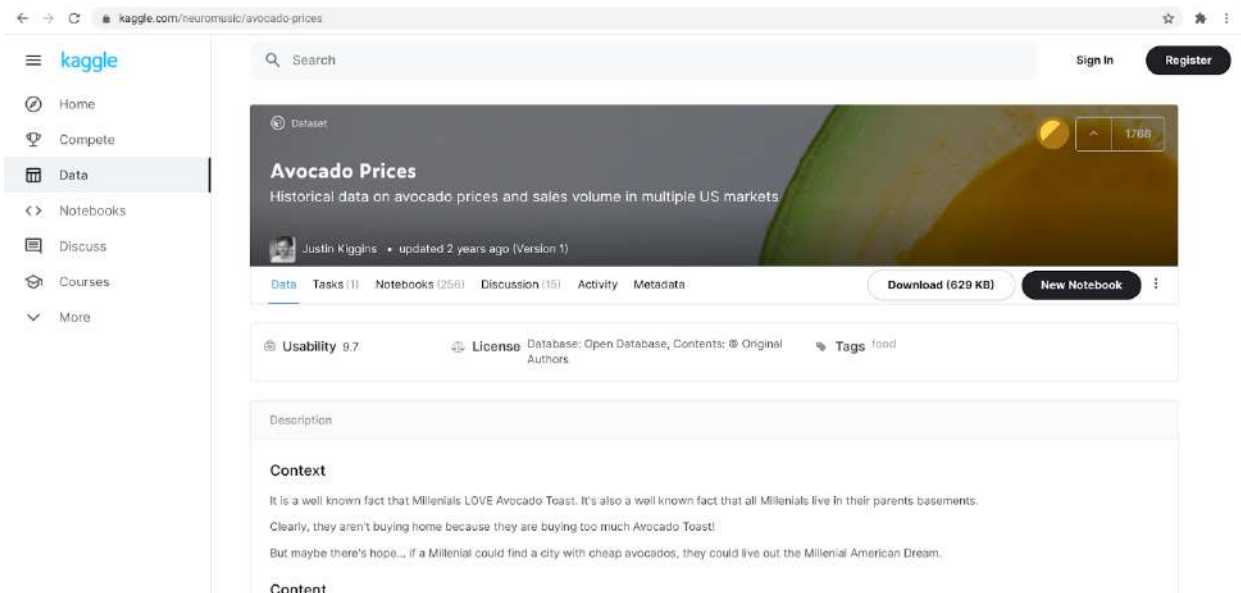
Sử dụng tập dữ liệu công khai là một cách tuyệt vời để thực hành với SQL. Một số ví dụ trong khóa học này sử dụng dữ liệu về giá bơ (avocado) để thực hiện tính toán trên BigQuery. Phần này sẽ hướng dẫn từng bước các bạn tải dữ liệu này vào bảng điều khiển BigQuery của riêng mình để bạn có thể làm theo các ví dụ minh họa

Nếu bạn chưa biết hoặc muốn ôn lại cách sử dụng BigQuery, xem lại khóa học **Chuẩn bị dữ liệu để khám phá**, nó sẽ bao gồm cách thiết lập tài khoản BigQuery.

Bước 1: Tải file CSV từ Kaggle

[Avocado dataset](#): Bộ dữ liệu giá bơ có sẵn công khai từ Kaggle mà bạn sẽ sử dụng (do Justin Kiggins cung cấp theo giấy phép Open Data Commons).

Bạn có thể tải dữ liệu này xuống thiết bị của riêng mình rồi tải lên BigQuery. Ngoài ra còn có các bộ dữ liệu công khai khác trên Kaggle mà bạn có thể tải xuống và sử dụng. Bạn có thể làm theo các bước sau để tải chúng vào bảng điều khiển của mình và tự thực hành!

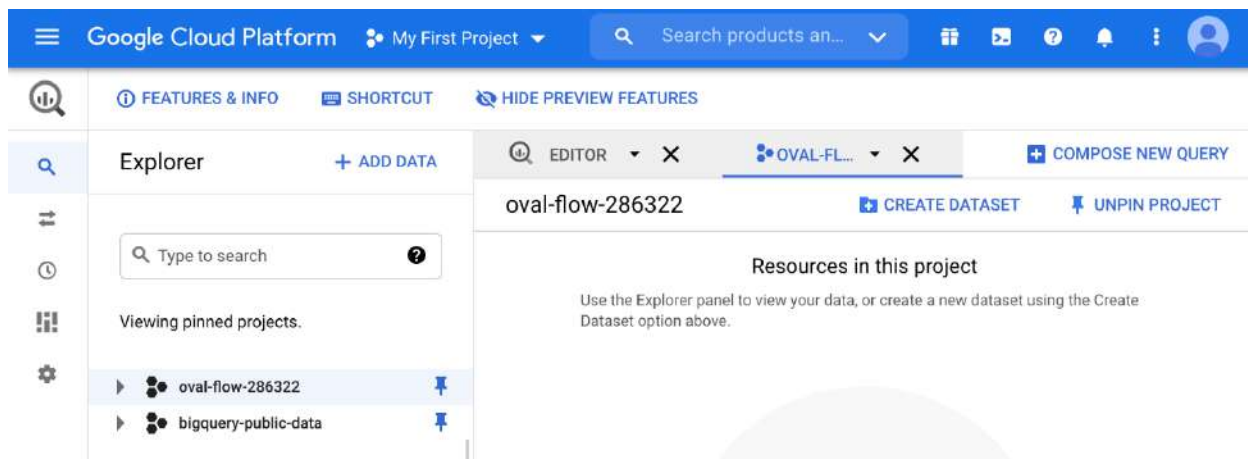


Bạn sẽ tìm thấy thêm một số thông tin về tập dữ liệu bờ, bao gồm bối cảnh, nội dung và nguồn gốc trên trang này. Hiện tại, bạn chỉ cần tải xuống tệp.

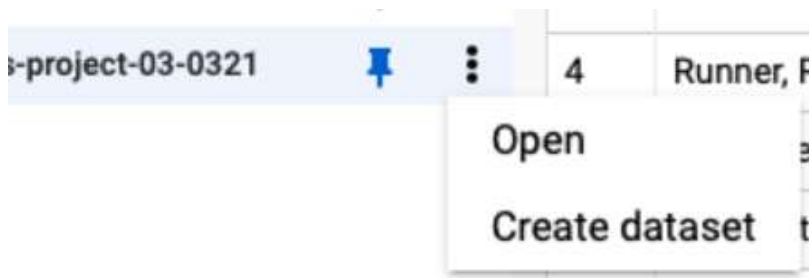
Bước 2: Mở bảng điều khiển BigQuery và tạo tập dữ liệu mới

Mở BigQuery. Sau khi đã tải xuống tập dữ liệu từ Kaggle, bạn có thể tải tập dữ liệu đó lên bảng điều khiển BigQuery của mình.

Trong Explorer ở bên trái bảng điều khiển của bạn, hãy nhấp vào dự án mà bạn muốn thêm tập dữ liệu - lưu ý rằng dự án của bạn sẽ không được đặt tên giống như dự án trong ví dụ ("oval-flow-286322"). Đừng chọn "bigquery-public-data" làm dự án của bạn vì đó là dự án công cộng mà bạn không thể thay đổi.



Nhấp vào biểu tượng Action (ba chấm dọc) bên cạnh dự án của bạn và chọn **Create Dataset**.



Tại đây, bạn sẽ đặt tên cho tập dữ liệu; trong trường hợp này, hãy nhập **avocado_data**. Sau đó, nhấp vào Create Dataset (nút màu xanh lam) ở dưới cùng để tạo tập dữ liệu mới. Thao tác này sẽ thêm dữ liệu trong Explore ở bên trái bảng điều khiển.

Create dataset

Dataset ID

Data location (Optional) ?

United States (US) ▼

Default table expiration ?

☒ Never

☐ Number of days after table creation:

Encryption

Data is encrypted automatically. Select an encryption key management solution.

☒ Google-managed key
No configuration required

☐ Customer-managed key
Manage via Google Cloud Key Management Service

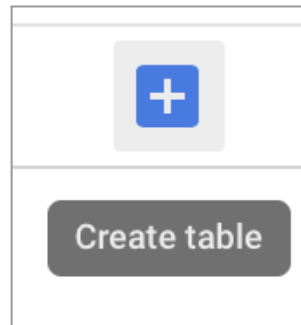
Create dataset Cancel

Bước 3: Mở tập dữ liệu mới và tạo một bảng mới

Điều hướng đến tập dữ liệu trong bảng điều khiển của bạn bằng cách nhấp để mở rộng dự án của bạn và chọn tập dữ liệu phù hợp được liệt kê. Trong trường hợp này, nó sẽ là **avocado_data**.



Nhấp vào biểu tượng Action (ba chấm dọc) bên cạnh tập dữ liệu của bạn và chọn Open. Sau đó nhấp vào biểu tượng + để tạo bảng.



Sau đó chọn:

- Trong Source, với tùy chọn Create Table, hãy chọn Upload.
- Nhấp vào Browser để chọn file CSV bạn vừa tải xuống máy tính của mình từ Kaggle. Định dạng tệp sẽ tự động thay đổi từ Avro sang CSV khi bạn chọn file.
- Đối với Table name, hãy nhập **avocado_prices**.
- Đối với Schema, chọn Auto detect check box. Sau đó, nhấp vào Create Table (nút màu xanh lam).

Create table

Source

Create table from

Upload

Empty table

Google Cloud Storage

Upload

Drive

Google Cloud Bigtable

File format

Browse

Avro

Destination

☒ Search for a project ☐ Enter a project name

Project name

My First Project

Dataset name

avocado_data

Table type

Native table

Table name

Letters, numbers, and underscores allowed

Schema

Source file defines the schema.

Partition and cluster settings

Partitioning

No partitioning

Clustering order (optional)

Clustering order determines the sort order of the data. Clustering can be used on both partitioned and non-partitioned tables.

Comma-separated list of fields to define clustering order (up to 4)

Advanced options

Create table Cancel

Trong Explorer, dữ liệu **avocado** sẽ xuất hiện trong bảng bên dưới tập dữ liệu đã tạo.

Đến đây, bạn đã có thể sẵn sàng sử dụng tập dữ liệu này.

5. Các loại xác thực dữ liệu

Bài đọc này mô tả mục đích, ví dụ và giới hạn của sáu loại xác thực dữ liệu. Năm loại đầu là các kiểu xác thực được liên kết với dữ liệu (kiểu, phạm vi, ràng buộc, nhất quán và cấu trúc) và kiểu thứ sáu tập trung vào xác thực mã ứng dụng được sử dụng để chấp nhận dữ liệu từ đầu vào của người dùng.

Là một nhà phân tích dữ liệu tập sự, bạn có thể không cần thực hiện tất cả các xác thực này. Nhưng bạn có thể hỏi liệu dữ liệu có được xác thực hay

không trước khi bạn bắt đầu làm việc với tập dữ liệu. Xác thực dữ liệu giúp đảm bảo tính toàn vẹn của dữ liệu. Nó cũng giúp bạn tự tin rằng dữ liệu bạn đang sử dụng là sạch. Danh sách sau đây phác thảo sáu loại xác thực dữ liệu và mục đích của mỗi loại, đồng thời bao gồm các ví dụ và giới hạn.

Kiểu dữ liệu

- **Mục đích:** Kiểm tra xem dữ liệu có khớp với kiểu dữ liệu được xác định cho một trường hay không.
- **Ví dụ:** Giá trị dữ liệu cho lớp 1-12 của trường phải là kiểu dữ liệu số.
- **Hạn chế:** Giá trị dữ liệu 13 sẽ vượt qua xác thực kiểu dữ liệu này nhưng đây là một giá trị không nên chấp nhận. Đối với trường hợp này, nên thực hiện xác thực phạm vi dữ liệu.

Phạm vi dữ liệu

- **Mục đích:** Kiểm tra xem dữ liệu có nằm trong phạm vi giá trị có thể chấp nhận được xác định cho trường hay không.
- **Ví dụ:** Giá trị dữ liệu cho lớp 1-12 của trường phải là giá trị từ 1 đến 12.
- **Hạn chế:** Giá trị dữ liệu 11.5 sẽ nằm trong phạm vi dữ liệu và cũng là dạng dữ liệu số. Nhưng không nên chấp nhận vì không lớp 0.5. Đối với trường hợp này, nên xác nhận rằng buộc dữ liệu

Ràng buộc dữ liệu

- **Mục đích:** Kiểm tra xem dữ liệu có đáp ứng các điều kiện hoặc tiêu chí nhất định cho một trường hay không. Điều này bao gồm loại dữ liệu được nhập cũng như các thuộc tính khác của trường, chẳng hạn như số ký tự.
- **Ví dụ:** Ràng buộc nội dung: Giá trị dữ liệu cho các lớp 1-12 của trường phải là số nguyên.
- **Hạn chế:** Giá trị dữ liệu 13 là một số nguyên và sẽ vượt qua xác thực ràng buộc nội dung. Tuy nhiên, điều đó sẽ không thể chấp nhận được vì 13 không phải là lớp học được công nhận. Đối với trường hợp này, xác thực phạm vi dữ liệu cũng cần thiết.

Nhất quán dữ liệu

- **Mục đích:** Kiểm tra xem dữ liệu có hợp lý trong ngữ cảnh của các dữ liệu liên quan khác hay không.
- **Ví dụ:** Giá trị dữ liệu cho ngày vận chuyển sản phẩm không được sớm hơn ngày sản xuất sản phẩm.

- **Hạn chế:** Dữ liệu có thể nhất quán nhưng vẫn không chính xác. Ngày vận chuyển có thể muộn hơn ngày sản xuất và vẫn bị sai.

Cấu trúc dữ liệu

- **Mục đích:** Kiểm tra xem dữ liệu có tuân theo cấu trúc đã đặt ra hay không.
- **Ví dụ:** Các trang web phải tuân theo một cấu trúc quy định để được hiển thị đúng cách.
- **Hạn chế:** Cấu trúc dữ liệu có thể đúng nhưng dữ liệu vẫn không chính xác hoặc không chính xác. Nội dung trên một trang web có thể được hiển thị đúng cách nhưng vẫn chứa thông tin sai.

Xác thực mã

- **Mục đích:** Kiểm tra xem mã ứng dụng có thực hiện một cách có hệ thống bất kỳ xác nhận nào đã đề cập trước đó trong quá trình nhập dữ liệu của người dùng hay không.
- **Ví dụ:** Các vấn đề phổ biến được phát hiện trong quá trình xác thực mã bao gồm: cho phép nhiều hơn một kiểu dữ liệu, kiểm tra phạm vi dữ liệu không được thực hiện hoặc kết thúc của chuỗi văn bản không được xác định rõ.
- **Hạn chế:** Xác thực mã có thể không xác thực tất cả các biến thể có thể có với đầu vào dữ liệu.

6. Thao tác với bảng tạm

Các **bảng tạm** giống như tên gọi của chúng — các bảng tạm thời trong cơ sở dữ liệu SQL không được lưu trữ vĩnh viễn. Trong bài đọc này, bạn sẽ tìm hiểu các phương pháp tạo bảng tạm thời bằng cách sử dụng các lệnh SQL. Bạn cũng sẽ học được một số phương pháp hay nhất để làm theo khi làm việc với các bảng tạm.

Nhắc lại về bảng tạm

- Chúng tự động bị xóa khỏi cơ sở dữ liệu khi bạn kết thúc phiên SQL của mình.
- Chúng có thể được sử dụng như một vùng lưu trữ để lưu trữ các giá trị nếu bạn đang thực hiện một loạt các phép tính. Điều này đôi khi được gọi là tiền xử lý dữ liệu.

- Chúng có thể thu thập kết quả của nhiều truy vấn riêng biệt. Điều này đôi khi được gọi là dàn dữ liệu (data staging). Dàn dữ liệu rất hữu ích nếu bạn cần thực hiện truy vấn trên dữ liệu đã thu thập hoặc hợp nhất dữ liệu đã thu thập.
- Chúng có thể lưu trữ một tập hợp con đã lọc của cơ sở dữ liệu. Bạn không cần phải chọn và lọc dữ liệu mỗi khi làm việc với nó. Ngoài ra, việc sử dụng ít lệnh SQL hơn sẽ giúp giữ dữ liệu của bạn sạch hơn.

Lưu ý rằng, mỗi cơ sở dữ liệu có một bộ lệnh riêng duy nhất để tạo và quản lý các bảng tạm. Chúng ta làm việc với BigQuery, vì vậy chúng ta sẽ tập trung vào các lệnh hoạt động tốt trong môi trường đó. Phần còn lại của bài đọc này sẽ đề cập đến các cách tạo bảng tạm, chủ yếu trong BigQuery.

Tạo bảng tạm trong Big Query

Bảng tạm có thể được tạo bằng các mệnh đề khác nhau. Trong BigQuery, mệnh đề **WITH** có thể được sử dụng để tạo bảng tạm. Cú pháp chung cho phương thức này như sau:

```
WITH
new_table_data AS (

SELECT *

FROM
Existing_table

WHERE
Tripduration >=60

)
```

Trong câu truy vấn này ta thấy:

- Câu lệnh bắt đầu bằng mệnh đề **WITH** theo sau là tên của bảng tạm mới mà bạn muốn tạo
- Mệnh đề **AS** xuất hiện sau tên của bảng mới. Mệnh đề này hướng dẫn cơ sở dữ liệu đặt tất cả dữ liệu được xác định trong phần tiếp theo của câu lệnh vào bảng mới.
- Dấu ngoặc đơn mở sau mệnh đề **AS** tạo truy vấn con lọc dữ liệu từ một bảng hiện có. Truy vấn con là một câu lệnh **SELECT** thông thường cùng với mệnh đề **WHERE** để chỉ định dữ liệu được lọc.
- Dấu ngoặc đóng kết thúc truy vấn con được tạo bởi mệnh đề **AS**.

Khi cơ sở dữ liệu thực hiện truy vấn này, trước tiên nó sẽ hoàn thành truy vấn con và gán các giá trị là kết quả từ truy vấn con đó thành “new_table_data”, đây là bảng tạm. Sau đó, bạn có thể chạy nhiều truy vấn trên dữ liệu đã lọc này mà không cần phải lọc dữ liệu ở mỗi lần.

Tạo bảng tạm trong các cơ sở dữ liệu khác (không được hỗ trợ trong BigQuery)

Phương pháp sau không được hỗ trợ trong BigQuery, nhưng hầu hết các phiên bản khác của cơ sở dữ liệu SQL đều hỗ trợ phương pháp này, bao gồm SQL Server và MySQL. Sử dụng **SELECT** và **INTO**, bạn có thể tạo một bảng tạm dựa trên các điều kiện được xác định bởi mệnh đề **WHERE** để định vị thông tin bạn cần cho bảng tạm. Cú pháp chung cho phương thức này như sau:

```
SELECT
*
INTO
AfricaSales
FROM
GlobalSales
WHERE
Region = "Africa"
```

Câu lệnh **SELECT** này sử dụng các mệnh đề tiêu chuẩn như **FROM** và **WHERE**, nhưng mệnh đề **INTO** yêu cầu cơ sở dữ liệu lưu trữ dữ liệu tạo một bảng tạm mới có tên, trong trường hợp này là “AfricaSales”.

Tạo bảng tạm thời do người dùng quản lý

Cho đến nay, chúng ta đã khám phá các cách tạo bảng tạm mà cơ sở dữ liệu chịu trách nhiệm quản lý. Tuy nhiên, bạn cũng có thể tạo các bảng tạm thời mà bạn có thể quản lý với tư cách là người dùng. Là một nhà phân tích, bạn có thể quyết định tạo một bảng tạm để phân tích mà bạn có thể tự quản lý. Bạn sẽ sử dụng câu lệnh **CREATE TABLE** để tạo loại bảng tạm thời này. Sau khi làm việc

xong với bảng, bạn sẽ xóa nó (delete) hay bỏ (drop) khỏi cơ sở dữ liệu vào cuối phiên của mình.

Lưu ý: BigQuery sử dụng CREATE TEMP TABLE thay vì CREATE TABLE, nhưng cú pháp chung là giống nhau.

```
CREATE TABLE table_name (  
    column1 datatype,  
    column2 datatype,  
    column3 datatype,  
    ....  
)
```

Sau khi hoàn tất thao tác với bảng tạm, bạn có thể bỏ bảng khỏi cơ sở dữ liệu bằng mệnh đề DROP TABLE. Cú pháp chung như sau:

```
DROP TABLE table_name
```

Các phương pháp gợi ý khi làm việc với bảng tạm:

- **Bảng tạm toàn cục và cục bộ:** Các bảng tạm toàn cục được cung cấp cho tất cả người dùng cơ sở dữ liệu và bị xóa khi tất cả các kết nối sử dụng chúng đã đóng. Bảng tạm cục bộ chỉ được cung cấp cho người dùng có truy vấn hoặc kết nối đã thiết lập bảng tạm. Rất có thể bạn sẽ làm việc với các bảng tạm cục bộ. Nếu bạn đã tạo bảng tạm cục bộ và là người duy nhất sử dụng nó, bạn có thể bỏ bảng tạm sau khi sử dụng xong.

- **Bỏ bảng tạm sau khi sử dụng:** Bỏ (drop) bảng tạm hơi khác với việc xóa (delete) bảng tạm. Bỏ bảng tạm không chỉ loại bỏ thông tin có trong các hàng của bảng mà còn loại bỏ chính các định nghĩa bảng (cột). Xóa bảng tạm thời sẽ xóa các hàng của bảng nhưng để lại định nghĩa bảng và các cột sẵn sàng được

sử dụng lại. Mặc dù các bảng tạm cục bộ bị bỏ sau khi bạn kết thúc phiên SQL của mình, nhưng nó có thể không xảy ra ngay lập tức. Nếu nhiều quá trình xử lý đang xảy ra trong cơ sở dữ liệu, thì việc loại bỏ các bảng tạm thời của bạn sau khi sử dụng chúng là một phương pháp hay để giữ cho cơ sở dữ liệu hoạt động trơn tru.

7. Tài liệu tham khảo

[1]<https://www.coursera.org/learn/analyze-data/supplement/s9khi/functions-with-multiple-conditions>

[2]<https://www.coursera.org/learn/analyze-data/supplement/j6w9Z/elements-of-a-pivot-table>

[3]<https://www.coursera.org/learn/analyze-data/supplement/qRo2l/using-pivot-tables-in-analysis>

[4]<https://www.coursera.org/learn/analyze-data/supplement/7KhvI/optional-upload-the-avocado-dataset-to-bigquery>

[5]<https://www.coursera.org/learn/analyze-data/supplement/tQAED/types-of-data-validation>

[6]<https://www.coursera.org/learn/analyze-data/supplement/oGADZ/working-with-temporary-tables>

Phần 2

HƯỚNG DẪN

TRẢ LỜI CÂU HỎI

Câu Hỏi về hiểu phân tích dữ liệu

1. Bạn hỏi các tình nguyện viên tham gia dự án về nhiệm vụ nào họ đã hoàn thành và thêm dữ liệu đó vào bảng tính chứa tất cả các nhiệm vụ cần thiết. Bạn sẽ sử dụng thông tin do các tình nguyện viên cung cấp để tìm ra những nhiệm vụ nào cần phải tiếp tục hoàn thành. Đây là một ví dụ về giai đoạn phân tích nào?

- A. Tổ chức dữ liệu (thành tập dữ liệu)
- B. Định dạng và điều chỉnh dữ liệu
- C. Nhận thông tin đầu vào từ những người khác
- D. Chuyển đổi dữ liệu

Đáp án: C

2. Bạn đang làm việc với ba tập dữ liệu về tỷ lệ cử tri đi bỏ phiếu trong quận của bạn. Đầu tiên, bạn xác định các mối quan hệ và các mẫu giữa các tập dữ liệu. Sau đó, bạn sử dụng các công thức và hàm để tính toán dựa trên dữ liệu của mình. Đây là một ví dụ về giai đoạn phân tích nào?

- A. Tổ chức dữ liệu (thành tập dữ liệu)
- B. Định dạng và điều chỉnh dữ liệu
- C. Nhận thông tin đầu vào từ những người khác
- D. Chuyển đổi dữ liệu

Đáp án: D

3. Bạn đang làm việc với tập dữ liệu từ một trường cao đẳng cộng đồng địa phương. Bạn sắp xếp các học sinh theo thứ tự bảng chữ cái của họ. Đây là một ví dụ về giai đoạn phân tích nào?

- A. Tổ chức dữ liệu (thành tập dữ liệu)
- B. Định dạng và điều chỉnh dữ liệu
- C. Nhận thông tin đầu vào từ những người khác
- D. Chuyển đổi dữ liệu

Đáp án: B

Câu hỏi về tổ chức dữ liệu

1. Điền vào chỗ trống: Một nhà phân tích dữ liệu sử dụng _____ để quyết định dữ liệu nào có liên quan đến phân tích của họ và các loại dữ liệu và biến nào là phù hợp.

- A. Tổ chức cơ sở dữ liệu
- B. Tham chiếu cơ sở dữ liệu
- C. Các mối quan hệ cơ sở dữ liệu
- D. Chuẩn hóa cơ sở dữ liệu

Đáp án: A

2. Một nhà phân tích dữ liệu muốn tổ chức cơ sở dữ liệu để chỉ hiển thị 100 giao dịch bất động sản gần đây nhất ở Stamford, Connecticut. Làm thế nào họ có thể làm điều đó?

- A. Thêm một bộ lọc để chỉ trả về các giao dịch ở Stamford, Connecticut, sau đó sắp xếp các giao dịch bán hàng gần nhất ở đầu danh sách.
- B. Lọc để loại bỏ các giao dịch ở Stamford, Connecticut, sau đó sắp xếp các giao dịch bán hàng gần nhất ở đầu danh sách.
- C. Thêm một bộ lọc để chỉ trả về các giao dịch ở Stamford, Connecticut, sau đó sắp xếp các giao dịch bán hàng xa nhất ở đầu danh sách.
- D. Lọc để loại bỏ các giao dịch ở Stamford, Connecticut, sau đó sắp xếp các giao dịch bán hàng xa nhất ở đầu danh sách.

Đáp án: A

3. Bạn đang làm việc với một bảng cơ sở dữ liệu có chứa dữ liệu khách hàng (*customer*). Cột *country* cho biết quốc gia mà mỗi khách hàng đang ở. Bạn muốn tìm những khách hàng ở Brazil.

Hãy hoàn tất câu truy vấn SQL bên dưới, thêm mệnh đề WHERE tương ứng để chỉ trả về những khách hàng ở Braxin.

```
SELECT * FROM customer
```

Đáp án: *WHERE country = "Brazil"*

Câu Hỏi về sắp xếp trong bảng tính

1. Chức năng sắp xếp nào trên thanh menu được sử dụng để sắp xếp tất cả dữ liệu trong bảng tính theo thứ tự của một cột cụ thể?

- A. Sort Sheet
- B. Sort Range
- C. Sort Data
- D. Sort By Rank

Đáp án: A

2. Trong bảng tính, nhà phân tích dữ liệu có thể sắp xếp một phạm vi từ tab Dữ liệu trong menu hoặc bằng cách gõ trực tiếp một hàm vào một ô trống.

- A. ĐÚNG
- B. SAI

Đáp án: A

3. Nhà phân tích sử dụng =SORT để sắp xếp dữ liệu bảng tính theo thứ tự giảm dần. Họ gõ gì ở cuối hàm sắp xếp này?

- A. FALSE
- B. TRUE
- C. DESCEND
- D. Z-A

Đáp án: A

Câu Hỏi về sắp xếp trong SQL

1. Một nhà phân tích dữ liệu muốn sắp xếp danh sách các loại cây bụi(shrubs) theo mức giá từ rẻ nhất đến đắt nhất. Họ nên sử dụng câu lệnh nào?Sort Sheet

- A. ORDER BY shrub_price
- B. ORDER BY shrub_price DESC
- C. WHERE shrub_price
- D. WHERE shrub_price ASC

Đáp án: A

2. Bạn đang làm việc với một bảng cơ sở dữ liệu chứa dữ liệu về các thể loại nhạc (genre). Bạn muốn sắp xếp theo tên các thể loại với thứ tự tăng dần. Các thể loại được liệt kê trong cột genre_name.

Hoàn tất câu truy vấn SQL bên dưới. Thêm mệnh đề ORDER BY để sắp xếp các thể loại theo tên với thứ tự tăng dần.

```
SELECT * FROM genre
```

Đáp án: *ORDER BY genre_name*

3. Bạn đang làm việc với một bảng cơ sở dữ liệu có chứa dữ liệu nhân viên (employee). Bạn muốn sắp xếp nhân viên với ngày thuê có thứ tự giảm dần. Ngày thuê được liệt kê trong cột hire_date.

Hoàn tất câu truy vấn SQL bên dưới. Thêm mệnh đề ORDER BY để sắp xếp nhân viên theo ngày tuyển dụng với thứ tự giảm dần.

```
SELECT * FROM employee
```

Đáp án: *ORDER BY hire_date*

Câu Hỏi tổng hợp

1. Trong quá trình phân tích dữ liệu, điều nào sau đây đề cập đến một giai đoạn phân tích? Chọn tất cả những câu phù hợp.

- A. Sắp xếp dữ liệu thành các phần dễ hiểu
- B. Nhận thông tin đầu vào từ những người khác
- C. Trực quan hóa dữ liệu
- D. Định dạng dữ liệu bằng cách sử dụng sắp xếp và bộ lọc

Đáp án: A,B,D

2. Giai đoạn nào của quá trình phân tích dữ liệu có mục tiêu xác định các xu hướng và mối quan hệ?

- A. Chuẩn bị (Prepare)
- B. Quá trình (Process)
- C. Phân tích (Analyze)
- D. Hành động (Act)

Đáp án: C

3. Mục tiêu của giai đoạn phân tích trong quá trình phân tích dữ liệu là gì?

- A. Để xác định các xu hướng và mối quan hệ trong dữ liệu
- B. Để tạo dữ liệu mới
- C. Để khái quát hóa về dữ liệu
- D. Để mô tả cấu trúc dữ liệu

Đáp án: A

4. Giai đoạn nào trong bốn giai đoạn phân tích, bạn thu thập các bộ dữ liệu liên quan cho một dự án?

- A. Tổ chức dữ liệu
- B. Định dạng và điều chỉnh dữ liệu
- C. Nhận thông tin đầu vào từ những người khác

D. Chuyển đổi dữ liệu

Đáp án: A

5. Giai đoạn nào trong bốn giai đoạn phân tích, bạn so sánh dữ liệu của mình với các nguồn bên ngoài?

A. Tổ chức dữ liệu

B. Định dạng và điều chỉnh dữ liệu

C. Nhận thông tin đầu vào từ những người khác

D. Chuyển đổi dữ liệu

Đáp án: C

6. Giai đoạn nào trong bốn giai đoạn phân tích, bạn tìm thấy mối tương quan giữa hai biến?

A. Tổ chức dữ liệu

B. Định dạng và điều chỉnh dữ liệu

C. Nhận thông tin đầu vào từ những người khác

D. Chuyển đổi dữ liệu

Đáp án: D

7. Bạn đang thực hiện một phép tính trong quá trình phân tích tập dữ liệu. Bạn đang ở giai đoạn phân tích nào?

A. Tổ chức dữ liệu

B. Định dạng và điều chỉnh dữ liệu

C. Nhận thông tin đầu vào từ những người khác

D. Chuyển đổi dữ liệu

Đáp án: D

8. Một nhà phân tích dữ liệu đang làm việc trên một tập dữ liệu và bắt đầu giai đoạn biến đổi dữ liệu trong quá trình phân tích. Họ sẽ thực hiện các hành động nào? Chọn tất cả những câu phù hợp.

A. Sắp xếp dữ liệu

- B. Lọc dữ liệu
- C. Tìm mối tương quan trong dữ liệu
- D. Thực hiện một phép tính với dữ liệu

Đáp án: C,D

9. Hành động nào sau đây có thể xảy ra khi chuyển đổi dữ liệu? Chọn tất cả những câu phù hợp.

- A. Nhận ra các mối quan hệ trong dữ liệu của bạn
- B. Xác định một mẫu trong dữ liệu của bạn
- C. Tính toán dựa trên dữ liệu của bạn
- D. Loại bỏ các dữ liệu không liên quan ra khỏi dữ liệu của bạn

Đáp án: A,B,C

10. Điền vào chỗ trống: Sắp xếp dữ liệu dựa trên _____ cụ thể mà bạn chọn.

- A. Độ đo
- B. Phép tính
- C. Mô hình
- D. Quan sát

Đáp án: A

11. Điền vào chỗ trống: Lọc chỉ bao gồm việc hiển thị dữ liệu đáp ứng _____ cụ thể, trong khi ẩn phần còn lại.

- A. Điều kiện
- B. Độ đo
- C. Mô hình
- D. Quan sát

Đáp án: A

12. Thông thường, một nhà phân tích dữ liệu sử dụng bộ lọc khi họ muốn mở rộng lượng dữ liệu mà họ đang làm việc.

A. Đúng

B. Sai

Đáp án: B

13. Một nhà phân tích dữ liệu đang sắp xếp dữ liệu trong một bảng tính. Họ chọn một tập hợp các ô cụ thể để giới hạn việc sắp xếp chỉ ở các ô được chỉ định. Họ đang sử dụng công cụ bảng tính nào?

A. Sort Range

B. Sort Sheet

C. Limit Sort

D. Limit Range

Đáp án: A

14. Một nhà phân tích dữ liệu đang sắp xếp dữ liệu bảng tính. Họ muốn đảm bảo rằng, khi họ sắp xếp lại dữ liệu, dữ liệu trên các hàng được giữ cùng nhau. Họ nên sử dụng kỹ thuật nào để sắp xếp dữ liệu?

A. Sort Sheet

B. Sort Column

C. Sort Together

D. Sort Rows

Đáp án: A

15. Một nhà phân tích dữ liệu đang sắp xếp dữ liệu trong một bảng tính. Họ đang sử dụng công cụ nào nếu tất cả dữ liệu được sắp xếp theo thứ hạng của một cột cụ thể và dữ liệu trên các hàng được lưu cùng nhau?

A. Sort Sheet

B. Sort Document

C. Sort Rank

D. Sort Together

Đáp án: A

16. Một nhà phân tích dữ liệu sắp xếp một phạm vi bảng tính giữa các ô F19 và G82. Họ sắp xếp theo thứ tự tăng dần theo cột thứ hai, Cột G. Cú pháp họ đang sử dụng là gì?

- A. =SORT(F19:G82, 2, TRUE)
- B. =SORT(F19:G82, B, TRUE)
- C. =SORT(F19:G82, 2, FALSE)
- D. =SORT(F19:G82, B, FALSE)

Đáp án: A

17. Một nhà phân tích dữ liệu sắp xếp một phạm vi bảng tính giữa các ô D5 và M5. Họ sắp xếp theo thứ tự giảm dần bởi cột thứ ba, Cột F. Cú pháp họ đang sử dụng là gì?

- A. =SORT(D5:M5, 3, FALSE)
- B. =SORT(D5:M5, C, FALSE)
- C. =SORT(D5:M5, 3, TRUE)
- D. =SORT(D5:M5, C, TRUE)

Đáp án: A

18. Một nhà phân tích dữ liệu sử dụng một chức năng để sắp xếp phạm vi bảng tính giữa các ô H1 và K65. Họ sắp xếp theo thứ tự tăng dần theo cột đầu tiên, Cột H. Cú pháp họ đang sử dụng là gì?

- A. =SORT(H1:K65, 1, TRUE)
- B. =SORT(H1:K65, A, TRUE)
- C. =SORT(H1:K65, 1, FALSE)
- D. =SORT(H1:K65, A, FALSE)

Đáp án: A

19. Bạn đang truy vấn một cơ sở dữ liệu có chứa dữ liệu về âm nhạc. Mỗi thể loại âm nhạc được cấp một số ID. Bạn chỉ quan tâm đến dữ liệu liên quan đến thể loại có ID số 7. ID thể loại được liệt kê trong cột *genre_id*.

Hoàn thành câu truy vấn SQL bên dưới bằng cách thêm mệnh đề WHERE để

chỉ trả về dữ liệu có *genre_id* bằng 7.

```
SELECT * FROM track
```

Đáp án: WHERE *genre_id* = 7

20. Bạn đang truy vấn một cơ sở dữ liệu có chứa dữ liệu về âm nhạc. Bạn chỉ quan tâm đến dữ liệu liên quan đến nhạc sĩ nhạc jazz Miles Davis. Tên của các nhạc sĩ được liệt kê trong cột *composer*

Hoàn thành câu truy vấn SQL bên dưới bằng cách thêm mệnh đề WHERE để chỉ trả về các bài nhạc sáng tác bởi Miles Davis.

```
SELECT * FROM track
```

Đáp án: WHERE *composer* = "Miles Davis"

21. Bạn đang truy vấn một cơ sở dữ liệu có chứa dữ liệu về âm nhạc. Bạn chỉ quan tâm đến album có ID là 6. Album ID được liệt kê trong cột *album_id*.

Hoàn thành câu truy vấn SQL bên dưới bằng cách thêm mệnh đề WHERE để chỉ trả về dữ liệu có *album_id* bằng 6.

```
SELECT * FROM track
```

Đáp án: WHERE *album_id* = 6

22. Bạn đang làm việc với cơ sở dữ liệu chứa dữ liệu hóa đơn về các giao dịch mua nhạc trực tuyến. Bạn chỉ quan tâm đến các hóa đơn được gửi cho khách hàng ở thành phố Paris. Bạn muốn sắp xếp các hóa đơn theo tổng đơn hàng với thứ tự tăng dần. Tổng số đơn đặt hàng được liệt kê trong cột *total*.

Hoàn thành câu truy vấn SQL bên dưới bằng cách thêm mệnh đề ORDER BY để sắp xếp các hóa đơn theo thứ tự tăng dần của *total*

```
SELECT * FROM invoice WHERE billing_city = "Paris"
```

Đáp án: ORDER BY *total*

23. Bạn đang làm việc với cơ sở dữ liệu chứa dữ liệu hóa đơn về các giao dịch mua nhạc trực tuyến. Bạn chỉ quan tâm đến các hóa đơn được gửi cho khách hàng ở thành phố Delhi. Bạn muốn sắp xếp các hóa đơn theo tổng đơn hàng với thứ tự tăng dần. Tổng số đơn đặt hàng được liệt kê trong cột *total*.

Hoàn thành câu truy vấn SQL bên dưới bằng cách thêm mệnh đề ORDER BY

để sắp xếp các hóa đơn theo thứ tự tăng dần của *total*

```
SELECT * FROM invoice WHERE billing_city = "Delhi"
```

Đáp án: ORDER BY total

24. Bạn đang làm việc với cơ sở dữ liệu chứa dữ liệu hóa đơn về các giao dịch mua nhạc trực tuyến. Bạn chỉ quan tâm đến các hóa đơn được gửi cho khách hàng ở thành phố Chicago. Bạn muốn sắp xếp các hóa đơn theo tổng đơn hàng với thứ tự tăng dần. Tổng số đơn đặt hàng được liệt kê trong cột *total*.

Hoàn thành câu truy vấn SQL bên dưới bằng cách thêm mệnh đề ORDER BY để sắp xếp các hóa đơn theo thứ tự tăng dần của *total*

```
SELECT * FROM invoice WHERE billing_city = "Chicago"
```

Đáp án: ORDER BY total

Câu Hỏi về Định dạng và chuyển đổi dữ liệu

1. Một ô bảng tính chứa nhiệt độ lạnh nhất từng được ghi nhận ở New Zealand: -22°C. Hàm nào sẽ hiển thị nhiệt độ đó bằng Fahrenheit?

- A. =CONVERT(-22, "C", "F")
- B. =CONVERT(-22, "F", "C")
- C. =CONVERT(-22, C, F)
- D. =CONVERT(-22, F, C)

Đáp án: A

2. Một nhà phân tích dữ liệu muốn đảm bảo các công thức bảng tính tiếp tục chạy chính xác, ngay cả khi ai đó nhập sai dữ liệu do nhầm lẫn. Họ nên chọn tùy chọn menu xác thực dữ liệu nào để ngăn cản các lỗi nhập dữ liệu?

- A. Từ chối đầu vào không hợp lệ (Reject Invalid Inputs)
- B. Mục nhập bị cấm (Forbid Entry)
- C. Từ chối văn bản trợ giúp (Deny Help Text)
- D. Xóa xác thực (Remove Validation)

Đáp án: A

3. Một nhà phân tích dữ liệu nhấp vào Format Cells trong trình đơn thả xuống và chọn tùy chọn Text Is Exactly November. Điều này làm thay đổi màu của tất cả các ô có chứa từ November. Người phân tích đang sử dụng công cụ bảng tính nào?

- A. Định dạng có điều kiện (Conditional formatting)
- B. Xác nhận dữ liệu (Data validation)
- C. Biến đổi (CONVERT)
- D. Lọc (Filtering)

Đáp án: A

Câu Hỏi về Kết hợp nhiều tập dữ liệu

1. Điền vào chỗ trống: Trong SQL, _____ có thể được sử dụng để kết hợp các chuỗi từ nhiều bảng nhằm tạo một chuỗi mới.

- A. CONCAT
- B. COMBINE
- C. CONCATENATE
- D. CONNECT

Đáp án: A

2. Bạn đang làm việc với một bảng cơ sở dữ liệu chứa dữ liệu về danh sách phát (playlist) cho các loại phương tiện kỹ thuật số khác nhau. Bạn chỉ quan tâm đến 4 danh sách phát đầu tiên.

Hoàn chỉnh câu truy vấn SQL bên dưới. Thêm mệnh đề LIMIT sẽ chỉ trả về 4 danh sách phát đầu tiên.

```
SELECT * FROM playlist
```

Đáp án:

```
SELECT *  
FROM playlist  
LIMIT 4
```

3. Hàm nào có thể được sử dụng để trả về số lượng ký tự trong ô B8 để bạn có

thể xác nhận rằng nó chứa chính xác 20 ký tự?

- A. =LEN(B8)
- B. =LEN(20)
- C. =LEN(20, B8)
- D. =LEN(B8, 20)

Đáp án: A

Câu Hỏi Tổng Hợp

1. Một nhà phân tích lưu ý rằng “160” trong ô A9 được định dạng là văn bản, nhưng nó phải là đô la Úc. Công cụ bảng tính nào có thể giúp họ chọn định dạng phù hợp?

- A. Format dưới dạng Currency
- B. CURRENCY
- C. EXCHANGE
- D. Format dưới dạng Dollar

Đáp án: A

2. Một nhà phân tích lưu ý rằng dữ liệu của họ được định dạng là Euro, nhưng nó phải được định dạng là peso. Công cụ bảng tính nào có thể giúp họ chọn định dạng phù hợp?

- A. Format dưới dạng Currency
- B. CURRENCY
- C. EXCHANGE
- D. Format dưới dạng peso

Đáp án: A

3. Một nhà phân tích làm việc cho hệ thống trường học của Anh vừa tải xuống một tập dữ liệu được tạo ở Hoa Kỳ. Dữ liệu số là chính xác nhưng nó được định dạng là đô la Mỹ và nhà phân tích cần nó dưới dạng bảng Anh. Công cụ bảng tính nào có thể giúp họ chọn định dạng phù hợp?

- A. Format dưới dạng Currency
- B. CURRENCY
- C. EXCHANGE
- D. Format dưới dạng bảng Anh

Đáp án: A

4. Bạn sử dụng bảng tính để sắp xếp các công việc sửa chữa nhà sắp tới. Cột A chứa danh sách cần sửa chữa và cột B ghi chú mức độ ưu tiên của từng hạng mục trong danh sách: Mức độ ưu tiên cao hoặc mức độ ưu tiên thấp. Bạn có thể sử dụng công cụ bảng tính nào để tạo danh sách thả xuống các ưu tiên cho mỗi ô trong cột B?

- A. Xác thực dữ liệu (Data validation)
- B. Menu hiện lên (Pop-up menus)
- C. Tìm (Find)
- D. Định dạng có điều kiện (Conditional formatting)

Đáp án: A

5. Bạn đang tạo bảng tính để giúp bạn tìm kiếm việc làm. Mỗi khi bạn tìm thấy một công việc thú vị, bạn thêm nó vào bảng tính. Sau đó, bạn muốn chỉ ra hai tùy chọn khả thi: “Cần gửi hồ sơ” hoặc “Đã gửi hồ sơ”. Công cụ bảng tính nào sẽ giúp bạn tiết kiệm thời gian bằng cách cho phép tạo danh sách thả xuống với các tùy chọn “Cần gửi hồ sơ” hoặc “Đã gửi hồ sơ”?

- A. Xác thực dữ liệu (Data validation)
- B. Menu hiện lên (Pop-up menus)
- C. Tìm (Find)
- D. Định dạng có điều kiện (Conditional formatting)

Đáp án: A

6. Bạn đang chuẩn bị một bảng tính theo dõi dự án. Bên cạnh mỗi nhiệm vụ dự án, bạn cần thêm tên của thành viên trong nhóm chịu trách nhiệm. Công cụ bảng tính nào sẽ giúp bạn tiết kiệm thời gian bằng cách cho phép bạn tạo danh

sách thả xuống với tên của các thành viên trong nhóm là tùy chọn?

- A. Xác thực dữ liệu (Data validation)
- B. Menu hiện lên (Pop-up menus)
- C. Tìm (Find)
- D. Định dạng có điều kiện (Conditional formatting)

Đáp án: A

7. Một nhà phân tích dữ liệu tại một dàn nhạc giao hưởng sử dụng một bảng tính để theo dõi có bao nhiêu buổi hòa nhạc yêu cầu hơn 80 nhạc sĩ. Họ sử dụng công cụ bảng tính để thay đổi cách các ô hiển thị khi các giá trị bằng 80 trở lên. Họ đang sử dụng công cụ gì?

- A. Định dạng có điều kiện (Conditional formatting)
- B. Xác thực dữ liệu (Data validation)
- C. CONVERT
- D. Thêm màu

Đáp án: A

8. Một nhà phân tích dữ liệu về nguồn nhân lực sử dụng bảng tính để theo dõi các ngày kỷ niệm làm việc của nhân viên. Họ thêm màu sắc cho bất kỳ nhân viên nào đã làm việc cho công ty hơn 10 năm. Công cụ bảng tính nào thay đổi cách các ô hiển thị khi các giá trị bằng 10 trở lên?

- A. Định dạng có điều kiện (Conditional formatting)
- B. Xác thực dữ liệu (Data validation)
- C. CONVERT
- D. Thêm màu

Đáp án: A

9. Bạn đang sử dụng một bảng tính để theo dõi các đăng ký tờ báo của mình. Bạn thêm màu sắc để cho biết đăng ký vẫn còn hay đã hết hạn. Công cụ bảng tính nào thay đổi cách các ô hiển thị khi các giá trị đã hết hạn?

- A. Định dạng có điều kiện (Conditional formatting)

- B. Xác thực dữ liệu (Data validation)
- C. CONVERT
- D. Thêm màu

Đáp án: A

10. Bạn đang làm việc với cơ sở dữ liệu SQL với các bảng cho các tuyến giao hàng ở California. Bảng chứa một cột với tên của các địa điểm đón. Một cột khác trong cùng một bảng chứa tên của các vị trí trả. Để tạo một cột mới lưu trữ kết hợp tên địa điểm đón và trả khách, bạn sử dụng hàm nào?

- A. CONCAT
- B. GROUP
- C. COMBINE
- D. JOIN

Đáp án: A

11. Bạn đang phân tích dữ liệu về thủ đô của các quốc gia khác nhau. Trong cơ sở dữ liệu SQL của bạn, bạn có một cột có tên các quốc gia và một cột khác có tên các thủ đô. Bạn có thể sử dụng hàm nào trong truy vấn của mình để kết hợp các quốc gia và thủ đô vào một cột mới?

- A. CONCAT
- B. GROUP
- C. COMBINE
- D. JOIN

Đáp án: A

12. Một nhà phân tích dữ liệu muốn viết một truy vấn SQL để kết hợp dữ liệu từ hai cột và thành một cột mới. Họ có thể sử dụng chức năng gì?

- A. CONCAT
- B. GROUP
- C. COMBINE

D. JOIN

Đáp án: A

13. Bạn đang truy vấn cơ sở dữ liệu của các viện bảo tàng để xác định nơi nào sẽ có triển lãm điêu khắc trong năm nay. Đối với dự án của bạn, bạn chỉ cần 50 bản ghi đầu tiên. Bạn nên thêm mệnh đề nào vào truy vấn SQL sau?

A. LIMIT 50

B. LIMIT = 50

C. LIMIT_50

D. LIMIT,50

Đáp án: A

14. Bạn đang truy vấn cơ sở dữ liệu về các diễn giả chính để xác định ai có chuyên môn về động vật học. Đối với dự án của bạn, bạn chỉ cần 12 bản ghi đầu tiên. Bạn nên thêm mệnh đề nào vào truy vấn SQL sau?

A. LIMIT 12

B. LIMIT = 12

C. LIMIT_12

D. LIMIT,12

Đáp án: A

15. Bạn đang truy vấn cơ sở dữ liệu về hương vị kem để xác định cửa hàng nào đang bán vị *mint_chip* nhiều nhất. Đối với dự án của bạn, bạn chỉ cần 80 bản ghi đầu tiên. Bạn nên thêm mệnh đề nào vào truy vấn SQL sau?

A. LIMIT 80

B. LIMIT = 80

C. LIMIT_80

D. LIMIT,80

Đáp án: A

16. Điền vào chỗ trống: Một nhà phân tích dữ liệu đang làm việc với một bảng

tính có các chuỗi văn bản rất dài. Họ sử dụng hàm LEN để đếm số ____ trong chuỗi văn bản.

- A. Ký tự
- B. Chuỗi con
- C. Giá trị
- D. Lĩnh vực

Đáp án: A

17. Một nhà phân tích dữ liệu đang làm việc với một bảng tính có các chuỗi văn bản rất dài. Họ sử dụng một hàm để đếm số ký tự trong ô G11. Cú pháp chính xác là gì?

- A. =LEN(G11)
- B. =LEN("G11")
- C. =LEN(G:G11)
- D. =LEN(G,11)

Đáp án: A

18. Một nhà phân tích dữ liệu đang làm việc với một bảng tính có các chuỗi văn bản rất dài. Thay vì tự đếm các ký tự để xác định số ký tự mà chúng chứa, họ có thể sử dụng công cụ nào?

- A. Hàm LEN
- B. Hàm COUNT
- C. Hàm CHAR
- D. Hàm MID

Đáp án: A

19. Ô bảng tính L6 chứa chuỗi văn bản "Function". Để trả về chuỗi con "Fun", cú pháp chính xác là gì?

- A. =LEFT(L6, 3)
- B. =RIGHT(L6, 3)

C. =LEFT(3,L6)

D. =RIGHT(3,L6)

Đáp án: A

20. Ô F2 của bảng tính chứa chuỗi văn bản “Dashboard”. Để trả về chuỗi con “board”, cú pháp đúng là gì?

A. =RIGHT(F2, 5)

B. =LEFT(F2, 5)

C. =RIGHT(5,F2)

D. =LEFT(5,F2)

Đáp án: A

21. Ô bảng tính E13 chứa chuỗi văn bản “Database”. Để trả về chuỗi con “data”, cú pháp chính xác là gì?

A. =LEFT(E13, 4)

B. =RIGHT(E13, 4)

C. =LEFT(4,E13)

D. =RIGHT(4,E13)

Đáp án: A

22. Khi làm việc với bảng tính, nhà phân tích dữ liệu có thể sử dụng hàm WHERE để định vị các ký tự cụ thể trong một chuỗi.

A. Đúng

B. Sai

Đáp án: B

23. Điền vào chỗ trống: Khi làm việc với bảng tính, nhà phân tích dữ liệu có thể sử dụng hàm _____ để định vị các ký tự cụ thể trong một chuỗi.

A. Find

B. WHERE

C. IDENTIFY

D. FROM

Đáp án: A

24. Khi làm việc với bảng tính, các nhà phân tích dữ liệu sử dụng hàm *find* để định vị các ký tự cụ thể trong một chuỗi. Hàm *find* phân biệt chữ hoa chữ thường, vì vậy cần nhập chuỗi con chính xác cách nó xuất hiện.

A. Đúng

B. Sai

Đáp án: A

Câu Hỏi Về VLOOKUP

1. Trong bảng tính, để biến một chuỗi văn bản trong ô thành một giá trị số, ta sẽ dùng hàm?

A. =VALUE(F8)

B. =CONVERT(F8)

C. =NUM(F8)

D. =MATCH(F8)

Đáp án: A

2. Mục đích của tham chiếu tuyệt đối trong một hàm, ví hạn như "\$C\$3" là?

A. Để tạo công thức và hàm không điều kiện

B. Để xóa các hướng dẫn không cần thiết khỏi công thức hoặc hàm

C. Để biểu thị các giá trị bị thiếu trong một công thức hoặc hàm

D. Để khóa các hàng và cột để chúng không thay đổi khi một hàm được sao chép

Đáp án: D

3. Trong hàm VLOOKUP, giá trị TRUE để cho hàm tìm kiếm các kết quả chính xác và FALSE cho hàm tìm kiếm các kết quả gần đúng.

A. Đúng

B. Sai

Đáp án: B

4. Cho một phần của bảng tính như bên cạnh:

Để tìm kiếm dân số của Nigeria, cú pháp VLOOKUP là?

A. =VLOOKUP("Nigeria", A2:C10, 2, false)

B. =VLOOKUP(Nigeria, A2:C10, 3, false)

C. =VLOOKUP(Nigeria, A2:C10, 3, true)

D. =VLOOKUP(Nigeria, A2,C10, 2, true)

Đáp án: A

5. Cho một phần của bảng tính như bên cạnh:

Để tìm chiều cao của tòa cao ốc tại Mecca, cú pháp VLOOKUP là?

A. =VLOOKUP("Mecca", A2:D7, 3, false)

B. =VLOOKUP(Mecca, A2:D7, 2, true)

C. =VLOOKUP(Mecca, A2:D7, 2, false)

D. =VLOOKUP(Mecca, A2,D7, 3, true)

Đáp án: A

Câu Hỏi Về Sử dụng JOIN để tổng hợp dữ liệu

1. Một nhà phân tích dữ liệu muốn rút trích các bản ghi từ cơ sở dữ liệu có các giá trị khớp nhau trong hai bảng khác nhau. Họ nên sử dụng hàm JOIN nào?

A. LEFT JOIN

B. RIGHT JOIN

C. INNER JOIN

D. OUTER JOIN

Đáp án: C

2. Bạn viết một câu truy vấn SQL để hướng dẫn cơ sở dữ liệu đếm các giá trị

trong một phạm vi được chỉ định. Bạn chỉ muốn đếm mỗi giá trị một lần, ngay cả khi nó xuất hiện nhiều lần. Bạn nên dùng hàm nào trong truy vấn của mình?

- A. COUNT
- B. COUNT DISTINCT
- C. COUNT VALUES
- D. COUNT RANGE

Đáp án: B

3. Một nhà phân tích dữ liệu muốn đặt tên tạm thời cho một cột trong truy vấn của họ để giúp đọc và ghi dễ dàng hơn. Họ nên sử dụng kỹ thuật nào?

- A. Filtering
- B. Aliasing
- C. Naming
- D. Tagging

Đáp án: B

Câu Hỏi về câu truy vấn con

1. Truy vấn nào sau đây chứa truy vấn con? Chọn tất cả các câu phù hợp.

Đáp án: A,B,C

2. Điền vào chỗ trống: Người phân tích dữ liệu sử dụng bí danh để giúp đọc và viết truy vấn dễ dàng hơn. Bí danh liên quan đến việc tạm thời _____ một bảng hoặc cột trong một truy vấn.

- A. Đặt tên
- B. Sao chép
- C. Che giấu
- D. Loại bỏ

Đáp án: A

3. Khi làm việc với truy vấn con, truy vấn bên ngoài thực hiện trước.

- A. Đúng
- B. Sai

Đáp án: B

Câu Hỏi tổng hợp

1. Điền vào chỗ trống: Tổng hợp dữ liệu bao gồm việc tạo _____ tập dữ liệu ban đầu từ nhiều nguồn.

- A. tóm tắt
- B. bản địa hóa
- C. sửa đổi
- D. mở rộng

Đáp án: A

2. Trong phân tích dữ liệu, quy trình thu thập dữ liệu từ nhiều nguồn và kết hợp nó thành một tập dữ liệu duy nhất, tóm tắt được gọi là?

- A. Ánh xạ dữ liệu (Data mapping)
- B. Thành phần dữ liệu (Data composition)
- C. Nhóm dữ liệu (Data grouping)
- D. Tổng hợp dữ liệu (Data aggregation)

Đáp án: D

3. Trong phân tích dữ liệu, tổng hợp dữ liệu là gì?

- A. Quá trình sửa đổi dữ liệu để phù hợp với việc phân tích.
- B. Quá trình di chuyển các điểm dữ liệu nhất định lên thứ hạng hoặc vị trí cao hơn.
- C. Quá trình thu thập dữ liệu từ nhiều nguồn và kết hợp nó thành một bộ sưu tập tổng hợp, duy nhất.
- D. Quy trình đảm bảo dữ liệu của công ty được lưu trữ, quản lý và duy trì đúng cách.

Đáp án: C

4. Hàm VALUE chuyển đổi giá trị số thành chuỗi văn bản trong bảng tính.

- A. Đúng
- B. Sai

Đáp án: B

5. Một nhà phân tích dữ liệu sử dụng hàm SUM để cộng các số lại với nhau từ một bảng tính. Tuy nhiên, sau khi nhận được kết quả bằng không, họ nhận ra những con số thực ra là chuỗi văn bản. Họ có thể sử dụng chức năng nào để chuyển chuỗi văn bản thành giá trị số?

- A. DIGIT
- B. VALUE
- C. FIGURE
- D. CONVERT

Đáp án: B

6. Một nhà phân tích dữ liệu muốn đảm bảo tất cả các ô chứa số trong bảng tính đều là số. Họ nên sử dụng chức năng nào để chuyển đổi văn bản thành giá trị số?

- A. VALUE
- B. CONVERT
- C. PROCESS
- D. EXCHANGE

Đáp án: A

7. Khi sử dụng hàm VLOOKUP, có một số hạn chế phổ biến mà các nhà phân tích dữ liệu cần lưu ý. Xác định những hạn chế này. Chọn tất cả câu phù hợp.

- A. Hàm VLOOKUP chỉ trả về kết quả phù hợp đầu tiên mà nó tìm thấy, ngay cả khi có nhiều kết quả phù hợp có thể xảy ra.
- B. Hàm VLOOKUP chỉ có thể trả về một giá trị từ dữ liệu ở bên trái của cột

mà nó được nhập vào.

C. Hàm VLOOKUP chỉ trả về các kết quả phù hợp mà nó tìm thấy trong khi tìm kiếm qua một hàng.

D. Hàm VLOOKUP chỉ có thể trả về một giá trị từ dữ liệu ở bên phải cột của giá trị phù hợp.

Đáp án: A, D

8. Khi sử dụng hàm VLOOKUP, có một số hạn chế phổ biến mà các nhà phân tích dữ liệu cần lưu ý. Một trong những hạn chế này là hàm VLOOKUP chỉ có thể trả về một giá trị từ dữ liệu ở bên trái giá trị đã so khớp.

A. Đúng

B. Sai

Đáp án: B

9. Khi sử dụng hàm VLOOKUP, có một số hạn chế phổ biến mà các nhà phân tích dữ liệu cần lưu ý. Một trong những hạn chế này là hàm VLOOKUP chỉ trả về kết quả phù hợp đầu tiên mà nó tìm thấy, ngay cả khi có nhiều kết quả phù hợp có thể có trong cột.

A. Đúng

B. Sai

Đáp án: A

10. Điền vào chỗ trống: Khi viết một hàm, chuyên viên phân tích dữ liệu bao bọc một mảng bằng các ký hiệu đô la. Đây là ____, được sử dụng để khóa mảng để các hàng và cột không thay đổi nếu hàm được sao chép.

A. Tham chiếu tùy ý

B. Tham chiếu xác thực

C. Tham chiếu tuyệt đối

D. Tham chiếu chính xác

Đáp án: C

11. Một nhà phân tích dữ liệu tạo một tham chiếu tuyệt đối xung quanh một

mảng hàm. Mục đích của tham chiếu tuyệt đối là gì?

- A. Để khóa mảng hàm để các hàng và cột không thay đổi nếu hàm được sao chép
- B. Để tự động thay đổi giá trị số thành giá trị tiền tệ
- C. Để sao chép một hàm và áp dụng nó cho tất cả các hàng và cột
- D. Để giữ cho một mảng hàm nhất quán để các hàng và cột sẽ tự động thay đổi nếu hàm được sao chép

Đáp án: A

12. Một nhà phân tích dữ liệu sử dụng tham chiếu tuyệt đối để khóa một mảng hàm để các hàng và cột không thay đổi nếu hàm được sao chép. Ký hiệu nào được sử dụng để tạo tham chiếu tuyệt đối?

- A. Dấu hoa thị (*)
- B. Ký hiệu và (&)
- C. Dấu thăng (#)
- D. Ký hiệu đô la (\$)

Đáp án: D

13. Cho dữ liệu từ bảng tính như bên

Để tìm kiếm dân số của Pakistan, cú pháp VLOOKUP chính xác là gì?

- A. `VLOOKUP("Pakistan", A2:B10, 2, false)`
- B. `VLOOKUP(Pakistan, A2*B10, 2, false)`
- C. `VLOOKUP(Pakistan, A2:B10, 3, false)`
- D. `VLOOKUP("Pakistan", A2:B10, 3, false)`

Đáp án: A

14. Cho dữ liệu từ bảng tính như bên

Để tìm kiếm tỉ lệ gia tăng dân số của Indonesia, cú pháp VLOOKUP là gì?

- A. `=VLOOKUP("Indonesia", A2:C10, 3, false)`
- B. `=VLOOKUP(Indonesia, A2:C10, 2, false)`

- C. =VLOOKUP(Indonesia, A2:C10, 3, false)
- D. =VLOOKUP("Indonesia", A2:C10, 2, false)

Đáp án: A

15. Cho dữ liệu từ bảng tính như bên

Để tìm kiếm dân số của Brazil, cú pháp VLOOKUP là gì?

- A. =VLOOKUP("Brazil", A2:B10, 2, false)
- B. =VLOOKUP(Brazil, A2:B10, 3, false)
- C. =VLOOKUP(Brazil, A2:B10, 3, false)
- D. =VLOOKUP(Brazil, A2:B10, 2, false)

Đáp án: A

16. INNER JOIN là một hàm trả về các bản ghi có giá trị phù hợp trong hai hoặc nhiều bảng. OUTER JOIN là một hàm kết hợp RIGHT và LEFT JOIN để trả về tất cả các bản ghi phù hợp trong cả hai bảng.

- A. Đúng
- B. Sai

Đáp án: A

17. Khi tạo câu truy vấn SQL, mệnh đề JOIN nào trả về chỉ các bản ghi có giá trị phù hợp trong hai hoặc nhiều bảng cơ sở dữ liệu?

- A. LEFT
- B. RIGHT
- C. INNER
- D. OUTER

Đáp án: C

18. Khi tạo câu truy vấn SQL, mệnh đề JOIN nào trả về tất cả các bản ghi phù hợp trong hai hoặc nhiều bảng cơ sở dữ liệu?

- A. LEFT
- B. RIGHT

- C. INNER
- D. OUTER

Đáp án: C

19. Một nhà phân tích dữ liệu viết câu truy vấn yêu cầu cơ sở dữ liệu chỉ trả về các giá trị riêng biệt trong một phạm vi được chỉ định, thay vì bao gồm các giá trị lặp lại. Họ sử dụng chức năng nào?

- A. COUNT
- B. COUNT DISTINCT
- C. RETURN
- D. RETURN VALUES

Đáp án: B

20. Hàm COUNT DISTINCT bao gồm các giá trị lặp lại khi trả về các giá trị trong một phạm vi được chỉ định.

A. Đúng

B. Sai

Đáp án: B

21. Một nhà phân tích dữ liệu viết câu truy vấn yêu cầu cơ sở dữ liệu trả về số hàng trong một phạm vi được chỉ định. Họ sử dụng chức năng nào?

- A. RANGE
- B. RETURN RANGE
- C. COUNT
- D. COUNT DISTINCT

Đáp án: C

22. Khi làm việc với truy vấn con, phần nào của phân đoạn truy vấn thực hiện đầu tiên?

- A. Truy vấn nhỏ hơn
- B. Truy vấn trong

- C. Truy vấn lớn hơn
- D. Truy vấn ngoài

Đáp án: B

23. Thuật ngữ nào sau đây mô tả một truy vấn con? Chọn tất cả câu phù hợp.

- A. Truy vấn nhỏ
- B. Truy vấn trong
- C. Truy vấn lồng nhau
- D. Lựa chọn trong

Đáp án: B, C, D

24. Điền vào chỗ trống: Trong câu lệnh SQL, _____ là tên của khối lệnh thực thi đầu tiên. Chọn tất cả những gì phù hợp.

- A. Truy vấn trung tâm
- B. Truy vấn trong
- C. Lựa chọn trung tâm
- D. Lựa chọn trong

Đáp án: B,D

Câu Hỏi về tính toán dữ liệu

1. Công thức trên bảng tính để nhân 50 và 233 là?

- A. =50*233
- B. 50*233
- C. =50x233
- D. 50x233

Đáp án: A

2. Cho bảng tính như bên.

Bạn muốn tính tỷ lệ phần trăm thu nhập hàng tháng được chi cho các mặt

hàng có giá trị lớn, chẳng hạn như tiền thuê nhà và tạp hóa. Để chỉ cộng các giá trị từ Cột B có giá hơn \$ 150, cú pháp đúng là gì?

- A. =SUMIF(B2:B12,">150")
- B. =SUMIF(B2:B12,"<150")
- C. =SUMIF(B2:B12,>150)
- D. =SUMIF(B2:B12,<150)

Đáp án: A

3. Một nhà phân tích dữ liệu đang làm việc với một bảng tính từ một công ty mỹ phẩm.

Bạn có thể tải tập dữ liệu: [Cosmetics Inc](#)

Ví dụ nào sau đây là mảng trong bảng tính này?

- A. Ô D7 và D14
- B. Các giá trị trong ô B2 đến B31
- C. Tất cả các ô có giá trị lớn hơn 100
- D. Tất cả các ô có giá trị số

Đáp án: B

Câu Hỏi về Bảng tóm tắt (Pivot table)

1. Mục đích của bảng tóm tắt trong bảng tính là gì?

- A. Để sắp xếp tất cả dữ liệu thành một định dạng nhỏ hơn
- B. Để tính tổng các giá cả cho từng loại sản phẩm
- C. Để tóm tắt dữ liệu về từng sản phẩm
- D. Để tìm giá trung bình của từng sản phẩm

Đáp án: B

2. Làm cách nào để điều chỉnh bảng tổng hợp để hiển thị cùng một dữ liệu, nhưng chỉ cho các sản phẩm được phân loại là màu be?

- A. Thêm một cột mới có nhãn màu be
- B. Thêm bộ lọc để chỉ hiển thị các sản phẩm màu be
- C. Sắp xếp hàng hiện tại theo màu sản phẩm
- D. Tóm tắt các giá trị theo sản phẩm

Đáp án: B

3. Bạn nên sử dụng công cụ bảng tính nào nếu muốn tìm giá trị trung bình bằng cách sử dụng các giá trị được tạo trong bảng tổng hợp?

- A. Định dạng có điều kiện (Conditional formatting)
- B. Một trường được tính toán (Calculated field)
- C. Xác thực dữ liệu (Data validation)
- D. Một bộ lọc (Filter)

Đáp án: B

Câu Hỏi về Tính toán trên SQL

1. Cho một bảng CSDL chứa dữ liệu hóa đơn (invoice). Bảng này gồm các cột *invoice_line_id* (các mục hàng trong mỗi hoá đơn), *invoice_id*, *unit_price*, và *quantity* (số lượng hàng trong mỗi mục hàng). Mỗi hóa đơn chứa nhiều mục hàng. Bạn muốn biết tổng giá cho từng mục trong số 5 mục đầu tiên trong bảng. Bạn quyết định nhân đơn giá với số lượng để có tổng giá cho từng mục hàng và sử dụng lệnh AS để lưu tổng số trong một cột mới có tên là *line_total*.

Hoàn thành câu truy vấn SQL để tính tổng giá cho mỗi mục hàng và lưu trữ nó trong một cột mới với *line_total*.

LƯU Ý: Dấu ba chấm (...) cho biết vị trí cần thêm câu lệnh.

```
SELECT invoice_line_id, invoice_id, unit_price, quantity, ... FROM
invoice_item LIMIT 5
```

Đáp án: *unit_price * quantity AS line_total*

2. Trong câu truy vấn SQL, toán tử modulo (%) thực hiện phép tính nào?

- A. Nó chuyển đổi một số thập phân thành phần trăm

- B. Nó trả về phần dư của một phép tính chia
- C. Nó áp tính lũy thừa cho một giá trị
- D. Nó tìm căn bậc hai của một số

Đáp án: B

3. Bạn đang làm việc với tập dữ liệu có tên cột “firstquarterexpenses”. Làm thế nào bạn có thể đổi tên cột này để làm cho nó dễ đọc hơn?

- A. first_quarter_expenses
- B. first quarter expenses
- C. Firstquarterexpenses
- D. first+quarter+expenses

Đáp án: A

Câu Hỏi về xác thực dữ liệu

1. Mục tiêu của việc kiểm tra và kiểm tra lại chất lượng dữ liệu của bạn trong quá trình xác thực dữ liệu là gì? Chọn tất cả những câu phù hợp.

- A. Dữ liệu đầy đủ và chính xác
- B. Dữ liệu được bảo mật
- C. Dữ liệu nhất quán
- D. Dữ liệu được sắp xếp và lọc

Đáp án: A,B,C

2. Bạn đang phân tích dữ liệu bệnh nhân cho một công ty chăm sóc sức khỏe. Trong quá trình xác thực dữ liệu, bạn nhận thấy rằng ngày phục vụ đầu tiên của một số bệnh nhân muộn hơn ngày phục vụ gần đây nhất. Bạn đang hoàn thành loại kiểm tra xác thực dữ liệu nào?

- A. Loại dữ liệu (Data type)
- B. Tính nhất quán dữ liệu (Data consistency)
- C. Cấu trúc dữ liệu (Data structure)
- D. Dải dữ liệu (Data range)

Đáp án: B

3. Trong quá trình phân tích, bạn hoàn thành kiểm tra xác thực dữ liệu để tìm lỗi trong số nhận dạng khách hàng (ID). ID khách hàng phải có tám ký tự và chỉ được chứa số. Loại lỗi nào sau đây sẽ được kiểm tra kiểu dữ liệu (data-type check) phát hiện?

- A. ID có chứa văn bản
- B. ID có nhiều hơn tám ký tự
- C. ID nhập sai cột
- D. ID bị lặp lại

Đáp án: A

Câu Hỏi về sử dụng SQL với bảng tạm

1. Khi nào các bảng tạm thời tự động bị xóa?

- A. Sau khi chạy một báo cáo từ bảng
- B. Sau khi kết thúc phiên trong cơ sở dữ liệu SQL
- C. Sau khi hoàn thành tất cả các phép tính trong bảng
- D. Sau khi chạy một truy vấn trong cơ sở dữ liệu SQL của bạn

Đáp án: B

2. Câu truy vấn SQL sau chứa thông tin về các chuyến xe đạp: Dữ liệu nào sẽ xuất hiện trong bảng tạm được tạo thông qua truy vấn này?

- A. Tổng số chuyến xe đạp
- B. Một tập hợp con ngẫu nhiên của các chuyến xe đạp
- C. Các chuyến xe đạp kéo dài đúng 60 phút
- D. Các chuyến xe đạp bằng hoặc hơn một giờ

Đáp án: C

3. Câu lệnh CREATE TABLE thêm vào bảng tạm mang lại lợi ích gì?

- A. Quyền truy cập cho bất kỳ ai sử dụng bảng
- B. Siêu dữ liệu về dữ liệu trong bảng
- C. Tính toán tự động
- D. Quy ước đặt tên cụ thể

Đáp án: A

Câu Hỏi Tổng Hợp

1. Một nhà phân tích dữ liệu đang làm việc với một bảng tính từ một công ty nội thất. Link đến bảng tính [này](#)

Nhà phân tích nhập vào một hàm để tìm số lượng giao dịch của các sản phẩm có màu đồng. Cú pháp của công thức nào sau đây sẽ trả về kết quả đó?

- A. =COUNTIF(E2:E30, “=brass”)
- B. =COUNTIF(F2:F30, “brass”)
- C. =SUMIF(F2:F30, “=brass”)
- D. =SUMIF(F2:F30, “brass only”)

Đáp án: B

2. Một nhà phân tích dữ liệu đang làm việc với một bảng tính từ một công ty nội thất. Link đến bảng tính [này](#)

Công thức nào sau đây cho phép nhà phân tích đếm số lần mua hàng với purchase_size lớn hơn hay bằng hai?

- A. =COUNTIF(H2:H30, “>=2”)
- B. =COUNTIF(G2:G30, “>=2”)
- C. =SUMIF(H2:H30, “=4”)
- D. =SUMIF(G2:G30, “<=1”)

Đáp án: A

3. Một nhà phân tích dữ liệu đang làm việc với một bảng tính từ một công ty nội

thất. Link đến bảng tính [này](#)

Nhà phân tích nhập một hàm để tìm số lượng sản phẩm có giá nhỏ hơn \$ 150.00. Công thức nào sẽ trả về kết quả đó?

- A. =COUNTIF(G2:G30, ">=150")
- B. =COUNTIF(G2:G30, "<150")
- C. =SUMIF(G2:G30, "<150")
- D. =SUMIF(G2:G30, ">150")

Đáp án: B

4. Bạn làm việc trong bảng tính và sử dụng hàm SUMIF trong công thức bên dưới

=SUMIF(D2:D10, ">=50", E2:E10)

Phần nào của công thức này cho biết phạm vi của giá trị sẽ được cộng?

- A. =SUMIF
- B. D2:D10
- C. >=50
- D. E2:E10

Đáp án: D

5. Điền vào chỗ trống: Khi bạn viết hàm SUMIF hoặc COUNTIF, phần đầu tiên của công thức trong dấu ngoặc đơn là _____.

- A. phạm vi (range)
- B. tình trạng (condition)
- C. tiêu chuẩn (criteria)
- D. Toán tử (operator)

Đáp án: A

6. Bạn làm việc trong bảng tính và sử dụng hàm SUMIF trong công thức bên dưới

=SUMIF(A1:A25, "<10", C1:C25)

Phần nào của công thức này cho biết tiêu chuẩn hay điều kiện

- A. =SUMIF
- B. A1:A25
- C. "<10"
- D. C1:C25

Đáp án: C

7. Một nhà phân tích dữ liệu đang làm việc trong bảng tính và sử dụng hàm SUMPRODUCT như công thức bên dưới:

=SUMPRODUCT(A2:A10,B2:B10)

Nó sẽ thêm các giá trị từ phạm vi đầu tiên (A2: A10) vào các giá trị từ phạm vi thứ hai (B2: B10). Sau đó, các tổng sẽ được nhân với nhau.

- A. Đúng
- B. Sai

Đáp án: B

8. Một nhà phân tích dữ liệu đang làm việc trong bảng tính và sử dụng hàm SUMPRODUCT như công thức bên dưới:

=SUMPRODUCT(A2:A10,B2:B10)

Hàm SUMPRODUCT hoạt động như thế nào?

- A. Nó cộng các giá trị trong phạm vi đầu, sau đó cộng các giá trị trong phạm vi thứ hai.
- B. Nó nhân các giá trị trong phạm vi đầu, sau đó nhân các giá trị trong phạm vi thứ hai.
- C. Nó cộng các phạm vi, sau đó nhân chúng với giá trị cuối cùng trong mảng thứ hai.
- D. Nó nhân các phạm vi, sau đó cộng tổng các tích của hai phạm vi.

Đáp án: D

9. Một nhà phân tích dữ liệu đang làm việc với một bảng tính từ một nhà bán lẻ. Link đến bảng tính [này](#)

Nhà phân tích muốn tìm ra giá trị của tất cả các mặt hàng (item) trong bảng tính. Công thức nào sẽ tính tổng giá của tất cả các mặt hàng?

- A. =SUM(C2:C21)
- B. =SUMPRODUCT(B2:B21,C2:C21)
- C. =SUMIF(B2:B21, "=1")
- D. =SUMIFS(C2:C21,B2:B21,"1",A2:A21,"_20")

Đáp án: B

10. Một nhà phân tích dữ liệu đang làm việc với một bảng tính chứa dữ liệu các bộ phim. Link đến bảng tính [này](#)

Để tìm ra doanh thu phòng vé cho mỗi loại phim, bạn sẽ sử dụng hàm SUM trong menu Values để tóm tắt dữ liệu.

- A. Đúng
- B. Sai

Đáp án: A

11. Một nhà phân tích dữ liệu đang làm việc với một bảng tính chứa dữ liệu các bộ phim. Link đến bảng tính [này](#)

Nếu bạn muốn tóm tắt dữ liệu bằng cách sử dụng hàm AVERAGE trong menu Values, bạn có thể thêm dữ liệu từ cột nào trong bảng tính? Chọn tất cả những câu phù hợp.

- A. Movie Title
- B. Genre
- C. Budget
- D. Box Office Revenue

Đáp án: C, D

12. Một nhà phân tích dữ liệu đang làm việc với một bảng tính chứa dữ liệu các bộ phim. Link đến bảng tính [này](#)

Nếu bạn muốn tính xem mỗi thể loại kiếm được bao nhiêu doanh thu phòng vé, bạn sẽ sử dụng chức năng nào trong menu Values để tóm tắt dữ liệu?

- A. SUM
- B. COUNTA
- C. AVERAGE
- D. PRODUCT

Đáp án: A

13. Phần nào của câu truy vấn SQL sau đây cho phép nhà phân tích kiểm soát thứ tự của các phép tính?

- A. (Yes_Responses + No_Responses)
- B. Yes_responses
- C. AS Responses_Per_Survey
- D. FROM Survey_1

Đáp án: A

14. Một nhà phân tích dữ liệu sử dụng câu truy vấn SQL sau để thực hiện các phép tính cơ bản trên dữ liệu của họ. Họ đang sử dụng loại toán tử nào? Chọn tất cả những câu phù hợp.

- A. Phép cộng
- B. Phép trừ
- C. Phép nhân
- D. Phân chia

Đáp án: A, D

15. Một nhà phân tích dữ liệu sử dụng câu truy vấn sau để thực hiện các phép tính cơ bản trên dữ liệu của họ. Các biến trong truy vấn có các giá trị sau: yes_responses = 10, no_responses = 12, Total_Survey = 22. Giá trị của biến Responses_Per_Survey là bao nhiêu?

- A. 1
- B. 11
- C. 22

D. 44

Đáp án: A

16. Cho một bảng CSDL chứa dữ liệu về nhạc (music). Bảng này gồm các cột *track_id*, *track_name*, *composer*, và *milliseconds* (thời lượng của một bài nhạc). Bạn chỉ quan tâm đến dữ liệu về nhạc sĩ cổ điển Johann Sebastian Bach. Bạn muốn biết thời lượng của mỗi bản nhạc Bach tính bằng giây. Bạn quyết định chia *milliseconds* cho 1000 để lấy thời lượng tính bằng giây và sử dụng lệnh AS để lưu trữ kết quả trong một cột mới có tên là *secs*.

Hoàn thành câu truy vấn SQL để tính độ dài tính theo giây của mỗi bản nhạc và lưu trữ trong cột mới tên là *secs*

LƯU Ý: Dấu ba chấm (...) cho biết vị trí cần thêm câu lệnh.

```
SELECT track_id, track_name, composer, milliseconds, . . . FROM track
WHERE composer = "Johann Sebastian Bach"
```

Đáp án: *milliseconds* / 1000 AS *secs*

17. Cho một bảng CSDL chứa dữ liệu về nhạc (music). Bảng này gồm các cột *track_id*, *track_name*, *composer*, và *album_id*. Bạn chỉ quan tâm đến dữ liệu về nhạc sĩ cổ điển Johann Sebastian Bach. Bạn muốn tạo ID album mới. Bạn quyết định nhân các ID album hiện tại với 10 để tạo các ID album mới và sử dụng lệnh AS để lưu trữ chúng trong một cột mới có tên là *new_album_id*.

Hoàn thành câu truy vấn SQL để tính album ID mới cho từng bài nhạc và lưu trữ trong cột mới tên là *new_album_id*

LƯU Ý: Dấu ba chấm (...) cho biết vị trí cần thêm câu lệnh.

```
SELECT track_id, track_name, composer, album_id, . . . FROM track WHERE
composer = "Johann Sebastian Bach"
```

Đáp án: *album_id* * 10 AS *new_album_id*

18. Cho một bảng CSDL chứa dữ liệu về nhạc (music). Bảng này gồm các cột *track_id*, *track_name* (tên của bài nhạc), *composer*, and *bytes* (kích thước lưu trữ của bài nhạc). Bạn muốn biết kích thước của từ bài nhạc của Bach tính bằng kilobyte. Bạn quyết định chia byte cho 1000 để có kích thước tính bằng kilobyte và sử dụng lệnh AS để lưu kết quả trong một cột mới gọi là *kilobyte*.

Hoàn thành câu truy vấn SQL để tính toán kích thước theo kilobyte cho mỗi bản nhạc và lưu trữ nó trong một cột mới dưới dạng *kilobyte*.

LƯU Ý: Dấu ba chấm (...) cho biết vị trí cần thêm câu lệnh.

```
SELECT track_id, track_name, composer, bytes, . . . FROM track WHERE  
composer = "Johann Sebastian Bach"
```

Đáp án: bytes / 1000 AS kilobytes

19. Cho một bảng CSDL chứa dữ liệu về nhạc (music). Bảng này gồm các cột *album_id* và *milliseconds* (thời lượng của từng bản nhạc trong mỗi album). Bạn muốn tìm hiểu tổng thời lượng cho mỗi album tính bằng mili giây và lưu trữ kết quả trong một cột mới có tên là *total_duration*.

Hoàn thành câu truy vấn SQL bên dưới bằng cách thêm mệnh đề GROUP BY để nhóm dữ liệu theo *album_id*

```
SELECT album_id, SUM(milliseconds) AS total_duration FROM track
```

Đáp án: GROUP BY album_id

20. Cho một bảng CSDL chứa dữ liệu về hóa đơn (invoice). Bảng này gồm các cột *customer_id* và *total* (tổng số tiền được lập hóa đơn cho mỗi hóa đơn). Một số khách hàng có nhiều hóa đơn. Bạn muốn biết tổng số tiền được lập hóa đơn cho mỗi khách hàng và lưu trữ kết quả trong một cột mới dưới dạng *total_number*.

Hoàn thành câu truy vấn SQL bên dưới bằng cách thêm mệnh đề GROUP BY để nhóm dữ liệu theo số *customer_id*

```
SELECT customer_id, SUM(total) AS total_amount FROM invoice
```

Đáp án: GROUP BY customer_id

21. Cho một bảng CSDL chứa dữ liệu về hóa đơn (invoice). Bảng này gồm các cột *invoice_id* và *quantity* (số lượng mua hàng bao gồm từng mục hàng trong hóa đơn). Mỗi hóa đơn chứa nhiều mục hàng. Bạn muốn tìm tổng số lần mua cho mỗi hóa đơn và lưu trữ kết quả trong một cột mới dưới dạng *total_purchases*.

Hoàn thành câu truy vấn SQL bên dưới bằng cách thêm mệnh đề GROUP BY để nhóm dữ liệu theo *invoice_id*

```
SELECT invoice_id, SUM(quantity) AS total_purchases FROM invoice_item
```

Đáp án: GROUP BY invoice_id

22. Cho một bảng CSDL chứa dữ liệu về hóa đơn (invoice). Bảng này gồm các cột *billing_state*, *billing_country*, và *total*. Bạn muốn biết tổng giá trung bình cho các hóa đơn được thanh toán cho tiểu bang Wisconsin. Bạn quyết định sử dụng hàm AVG để tìm tổng trung bình và sử dụng lệnh AS để lưu kết quả trong một cột mới có tên là *average_total*.

Hoàn thành câu truy vấn SQL để tính tổng trung bình và lưu trữ nó trong một cột mới có tên *average_total*

LƯU Ý: Dấu ba chấm (...) cho biết vị trí cần thêm câu lệnh.

```
SELECT billing_state, billing_country, ... FROM invoice WHERE billing_state = "WI"
```

Đáp án: AVG(total) AS average_total

23. Cho một bảng CSDL chứa dữ liệu về hóa đơn (invoice). Bảng này gồm các cột *billing_city*, *billing_country*, và *total*. Bạn muốn biết tổng giá trung bình cho các hóa đơn được thanh toán cho thành phố Vancouver. Bạn quyết định sử dụng hàm AVG để tìm tổng trung bình và sử dụng lệnh AS để lưu kết quả trong một cột mới có tên là *average_total*.

Hoàn thành câu truy vấn SQL để tính tổng trung bình và lưu trữ nó trong một cột mới có tên *average_total*

LƯU Ý: Dấu ba chấm (...) cho biết vị trí cần thêm câu lệnh.

```
SELECT billing_city, billing_country, ... FROM invoice WHERE billing_city = "Vancouver"
```

Đáp án: AVG(total) AS average_total

24. Cho một bảng CSDL chứa dữ liệu về hóa đơn (invoice). Bảng này gồm các cột *billing_country*, và *total*. Bạn muốn biết tổng giá trung bình cho các hóa đơn được thanh toán cho quốc gia Ấn Độ. Bạn quyết định sử dụng hàm AVG để tìm tổng trung bình và sử dụng lệnh AS để lưu kết quả trong một cột mới có tên là *average_total*.

Hoàn thành câu truy vấn SQL để tính tổng trung bình và lưu trữ nó trong một

cột mới có tên *average_total*

LƯU Ý: Dấu ba chấm (...) cho biết vị trí cần thêm câu lệnh.

```
SELECT billing_country, ... FROM invoice WHERE billing_country = "India"
```

Đáp án: *AVG(total) AS average_total*