



**FPT POLYTECHNIC**



---

[caodang.fpt.edu.vn](http://caodang.fpt.edu.vn)

---

## **NHẬP MÔN XỬ LÝ DỮ LIỆU**

---

### **BÀI 2: THU THẬP DỮ LIỆU**

# MỤC TIÊU

- ◎ HIỂU ĐƯỢC CÁC PHƯƠNG PHÁP THU THẬP DỮ LIỆU
- ◎ PHÂN LOẠI ĐƯỢC CÁC DẠNG DỮ LIỆU
- ◎ NẮM ĐƯỢC CÁCH THU THẬP DỮ LIỆU THỰC TẾ



- ❑ THU THẬP DỮ LIỆU
- ❑ GIỚI THIỆU CÁC PHƯƠNG PHÁP THU THẬP DỮ LIỆU
- ❑ CÁC DẠNG DỮ LIỆU CƠ BẢN
- ❑ MỘT SỐ TÌNH HUỐNG THU THẬP DỮ LIỆU THỰC TẾ

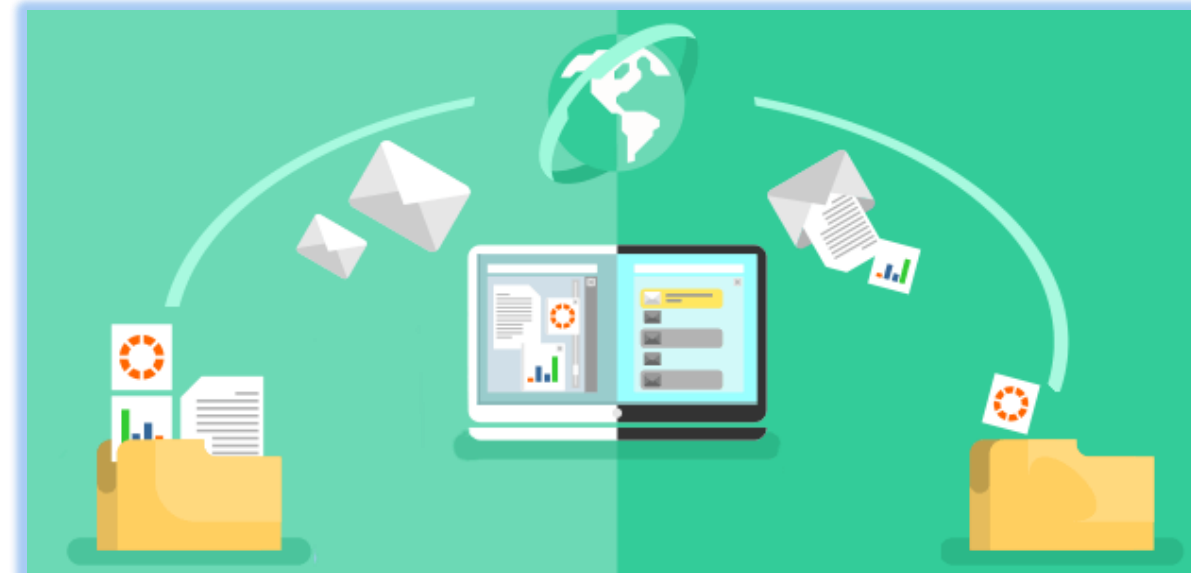


# PHẦN I: THU THẬP DỮ LIỆU

---

## ❑ THU THẬP DỮ LIỆU LÀ GÌ?

- ❖ Là một quá trình thu thập thông tin từ **tất cả các nguồn có liên quan** để tìm câu trả lời vấn đề nghiên cứu, kiểm tra giả thuyết, đánh giá kết quả.
- ❖ Từ đó, tiến hành phân tích dữ liệu để đưa ra các quyết định Marketing/ kinh doanh đúng đắn.



## ❑ THU THẬP DỮ LIỆU LÀ GÌ?

❖ Là nguồn dữ liệu đáng tin cậy cho doanh nghiệp:

- Các số liệu thống kê từ các cơ quan chính phủ.
- Báo cáo ngành từ những công ty nghiên cứu thị trường lớn.
- Các bài báo đăng trên các tạp trí uy tín.
- Kênh thông tin của chính đối thủ.



□ Có hai phương pháp thu thập dữ liệu:

- ❖ Phương pháp thu thập dữ liệu **thứ cấp**.
- ❖ Phương pháp thu thập dữ liệu **sơ cấp**.







## □ Phương pháp thu thập sơ cấp:

- ❖ Có rất nhiều những dữ liệu có sẵn trên những nguồn này về thông tin liên quan đến chủ đề hay lĩnh vực nghiên cứu đang muốn tiến hành.
- ❖ Việc áp dụng **bộ tiêu chí phù hợp để chọn dữ liệu thứ cấp<sup>1</sup>** được sử dụng trong nghiên cứu đóng vai trò quan trọng trong việc tăng mức độ tin cậy của thu thập.

## ❑ Phương pháp thu thập sơ cấp:

### ❖ Các báo cáo thị trường:

- Nhiều báo cáo về thị trường được công bố bởi các công ty nghiên cứu thị trường lớn, uy tín sẽ giúp có cái nhìn tổng quan, sâu sắc về các báo cáo và xu hướng ngành đang kinh doanh là điều vô cùng quan trọng.



## □ Phương pháp thu thập sơ cấp:

### ❖ **Dữ liệu đã được thống kê cơ quan chính phủ:**

- Là nguồn cung cấp những thông tin chính xác và những xu hướng có tầm ảnh hưởng lớn.
- Có thể lấy dữ liệu từ Data.gov hay The World Bank, website của tổng cục thống kê **<http://www.gso.gov.vn/>**



## □ Phương pháp thu thập sơ cấp:

### ❖ Lịch sử dữ liệu từ công ty:

- Có thể thu thập dữ liệu lịch sử từ công ty giúp việc phân tích xử lý dữ liệu được hiệu quả hơn.
- Tiết kiệm ngân sách đáng kể cho hoạt động marketing công ty.



## ❑ Phương pháp thu thập thứ cấp:

### ❖ Phỏng vấn cá nhân và nhóm tập trung:

- Các cuộc thảo luận trực tiếp là một nguồn dữ liệu định tính tuyệt vời.
- Thu thập dữ liệu từ các cuộc phỏng vấn qua điện thoại, hội nghị video.



## □ Phương pháp thu thập thứ cấp:

### ❖ Khảo sát thông qua bảng câu hỏi:

- Dữ liệu thu về từ cuộc khảo sát là sự lượng hóa, kiểm nghiệm lại những thông tin đã khám phá từ những cuộc phỏng vấn chuyên sâu.
- Đảm bảo xây dựng cái nhìn tổng thể, chính xác về nhóm khách hàng mục tiêu mà doanh nghiệp đang hướng đến.

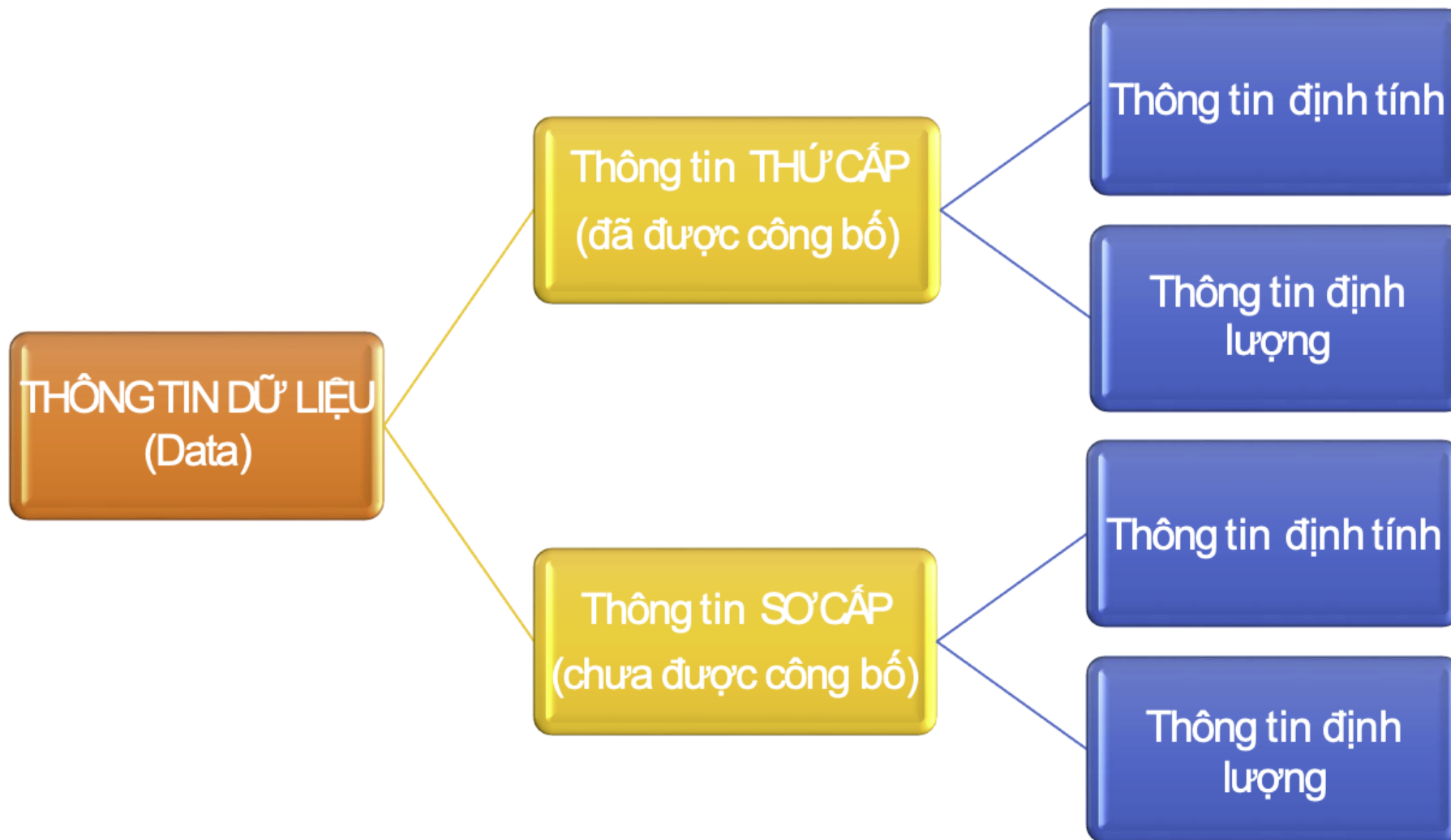






## PHẦN II: THU THẬP DỮ LIỆU (TT)

---



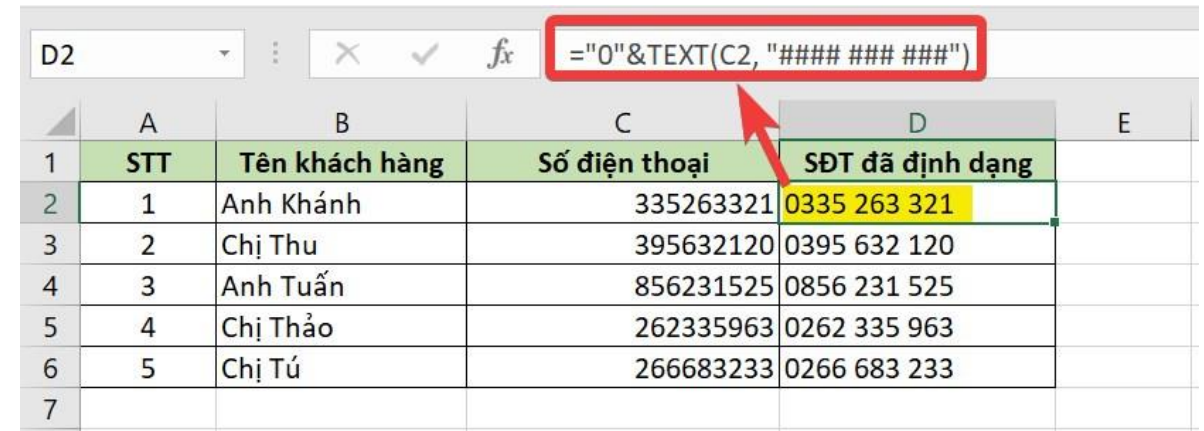


## ❑ Nguồn thông tin dữ liệu:

❖ Số liệu (Data) là những con số.

❖ Dữ liệu (Data) bao gồm:

- ✓ Số liệu.
- ✓ Những kí tự chữ (a, b, c...).
- ✓ Hình tượng (hình ảnh, sơ đồ, đồ thị, figures).
- ✓ Âm thanh.
- ✓ Video..



	A	B	C	D	E
1	STT	Tên khách hàng	Số điện thoại	SĐT đã định dạng	
2	1	Anh Khánh	335263321	0335 263 321	
3	2	Chị Thu	395632120	0395 632 120	
4	3	Anh Tuấn	856231525	0856 231 525	
5	4	Chị Thảo	262335963	0262 335 963	
6	5	Chị Tú	266683233	0266 683 233	
7					

## ❑ Nguồn thông tin dữ liệu:

- ❖ Số dạng “THÔ” (Raw data).
- ❖ Số liệu/Dữ liệu chỉ là những **giá trị thô** ban đầu, và tự nó có thể chưa có ý nghĩa và cần xử lý giá trị thô để mang lại thông tin hữu ích.
- ❖ VD:

	Organization Name URL	Last Funding Amount - ORIG	Last Funding Amount	Last Funding Amount Currenc	Last Funding Type
1	organization/app-annie	\$63,000,000	\$63,000,000		Series E
2	organization/sisense	\$50,000,000	\$50,000,000		Series D
3	organization/insightec	\$22,000,000	\$22,000,000		Series D
4	organization/cumulus-networks	\$35,000,000	\$35,000,000		Series C
5	organization/skyscanner	\$192,000,000	\$192,000,000		Venture - Series Unk
6	organization/flatiron-health	\$175,000,000	\$175,000,000		Series C
7	organization/jawbone	\$165,000,000	\$165,000,000		Private Equity
8	organization/plains-all-american-p	\$1,500,000,000	\$1,500,000,000		Post-IPO Equity
9	organization/untappd	\$3,750,000	\$3,750,000		Angel
10	organization/bitvault	\$7,600,000	\$7,600,000		Series A
11	organization/grindr	\$93,000,000	\$93,000,000		Private Equity
12	organization/nanthealth	\$52,500,000	\$52,500,000		Venture - Series Unk
13	organization/starling-3	\$70,000,000	\$70,000,000		Venture - Series Unk
14	organization/euclid	\$20,000,000	\$20,000,000		Series C
15	organization/greenwave-systems	\$45,000,000	\$45,000,000		Series C
16	organization/jwplayer	\$20,000,000	\$20,000,000		Series D

## □ Nguồn thông tin dữ liệu:

- ❖ Số liệu/dữ liệu có thể chuyển sang thông tin.
- ❖ Số liệu/dữ liệu không phải hoàn toàn là thông tin
- ❖ Cùng cơ sở dữ liệu nhưng sử dụng các phương pháp khác nhau sẽ cung cấp thông tin khác nhau.

## ❑ Thu thập dữ liệu từ trường Cao Đẳng FPT Polytechnic HCM

0.FPOLY-COM107\_TheoDoiTienDo-080122\_HCM

Home Insert Draw Page Layout Formulas Data Review View

Calibri (Body) 11 A<sup>+</sup> A<sup>-</sup> B I U Wrap Text Merge & Center General Conditional Formatting Format as Table Cell Styles Insert Delete For

Security Warning External Data Connections have been disabled

A14 fx =RIGHT(C14,7)

	A	B	C	D	E	F	G	H	I	J	K	L	U	V	W	X	Y
1	Student ID	Email	Username	Grade	Quiz 1: Qui	Quiz 2: Qui	Quiz 3: Qui	Quiz 4: Qui	Quiz 5: Qui	Quiz 6: Qui	Quiz 7: Qui	Quiz 8: Qui	Σ Quiz	Σ Quiz (Chưa học)	Σ Quiz (Đang học)	Σ Quiz (Đã đạt)	Tổng số Sv chưa tham gia học
2	2942	hanth17@fpt.edu.vn	hanth17	0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	8	0	0	Chưa tham gia học lần nào
3	21221	cuongnp@fpt.edu.vn	cuongnp	0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	8	0	0	Chưa tham gia học lần nào
4	3065	nhulh@fpt.edu.vn	nhulh	0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	8	0	0	Chưa tham gia học lần nào
5	743	hoaitt12@fpt.edu.vn	hoaitt12	0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	8	0	0	Chưa tham gia học lần nào
6	779	cuhv2@fpt.edu.vn	cuhv2	0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	8	0	0	Chưa tham gia học lần nào
7	31193	tricm4@fpt.edu.vn	tricm4	0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	8	0	0	Chưa tham gia học lần nào
8	619	tula@fpt.edu.vn	tula	0.13	1.0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	7	1	0	
9	50991	congdc5@fpt.edu.vn	congdc5	0.13	1.0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	7	1	0	
10	569	diennt@fpt.edu.vn	diennt	0.04	0.33	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	7	1	0	
11	51000	lynt51@fpt.edu.vn	lynt51	0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	8	0	0	Chưa tham gia học lần nào
12	50990	duykh6@fpt.edu.vn	duykh6	0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	8	0	0	Chưa tham gia học lần nào
13	654	danglm@fpt.edu.vn	danglm	0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	8	0	0	Chưa tham gia học lần nào
14	kietlpt	kietlpt@fpt.edu.vn	kietlpt	0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	8	0	0	Chưa tham gia học lần nào
15	ngttt14	giangttt14@fpt.edu.vn	giangttt14	0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	8	0	0	Chưa tham gia học lần nào
16	ps25414	andqps25414@fpt.edu.vn	andqps25414	0.25	1.0	1.0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	6	2	0	
17	ps24891	anhpbps24891@fpt.edu.vn	anhpbps24891	0	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	Not Attempt	8	8	0	0	Chưa tham gia học lần nào

❑ Thu thập dữ liệu từ website: <http://www.medium.com>

• Total costs

• Net Margin and Net Income

• Total revenue cumulative and Total Net income cumulative

H47  $\times$   $\checkmark$   $fx$

	A	B	C	D	E	F	G
<b>Your E-commerce business revenue</b>							
1	Price	% out of total orders	Cost to produce, 60%	Margin, 40%			
2	Product 1	\$ 25.00	20%	\$ 15.00	\$ 10.00		
3	Product 2	\$ 50.00	20%	\$ 30.00	\$ 20.00		
4	Product 3	\$ 100.00	20%	\$ 60.00	\$ 40.00		
5	Product 4	\$ 150.00	20%	\$ 90.00	\$ 60.00		
6	Product 5	\$ 200.00	20%	\$ 120.00	\$ 80.00		
7	Revenue per sale, avg	\$ 105.00					
8	Cost to produce, avg	\$ 63.00					
9							
10	<b>Metric</b>	<b>Jan-18</b>	<b>Feb-18</b>	<b>Mar-18</b>	<b>Apr-18</b>	<b>May-18</b>	<b>Jun-18</b>
11	Number of Sales	30	36	45	59	79	111
12	Sales growth MoM, %	0%	20%	25.00%	30.00%	35.00%	40.00%
13	Revenue	\$ 3,150.00	\$ 3,780.00	\$ 4,725.00	\$ 6,142.50	\$ 8,292.38	\$ 11,609.33
14	Total Cost to produce	\$ 1,890.00	\$ 2,268.00	\$ 2,835.00	\$ 3,685.50	\$ 4,975.43	\$ 6,965.60
15	Shipping costs	\$ 300.00	\$ 360.00	\$ 450.00	\$ 585.00	\$ 789.75	\$ 1,105.65
16	Transaction fee, 3%	\$ 94.50	\$ 113.40	\$ 141.75	\$ 184.28	\$ 248.77	\$ 348.28
17	Total costs	\$ 2,284.50	\$ 2,741.40	\$ 3,426.75	\$ 4,454.78	\$ 6,013.95	\$ 8,419.52
18	Add more	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
19	Add more	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
20	Net Margin	27%	27%	27%	27%	27%	27%
21	Net Income	\$ 865.50	\$ 1,038.60	\$ 1,298.25	\$ 1,687.73	\$ 2,278.43	\$ 3,189.80
22							
23	Total revenue cumulative	\$ 3,150.00	\$ 6,930.00	\$ 11,655.00	\$ 17,797.50	\$ 26,089.88	\$ 37,699.20
24	Total Net income cumulative	\$ 865.50	\$ 1,904.10	\$ 3,202.35	\$ 4,890.08	\$ 7,168.50	\$ 10,358.30

[E-commerce revenue model](#)

**Recommendations**

• Same as before, quickly adj...s, and Excel will take care of the rest.

1.1K | 14

**Nurzhan Ospanov**  
422 Followers  
Customer Success at Statsbot. I write in English and Russian. <http://nurzhan.me>

[Follow](#)

**More from Medium**

Carlos Carrero in Snowflake  
**Distributed ML with Snowpark Python UDFs**

Piyush Jain  
**Python Explained Like you are 5!- Focussed on Data**

Amit Kumar  
**LOD Expression in Tableau**

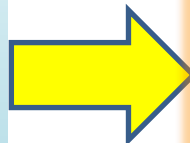
Shannon in SIOT, GovTech  
**Internship Experience—Don't DASH through data analysis**

Help Status Writers Blog Careers Privacy Terms About



## □ Thu thập dữ liệu thông qua đọc nội dung văn bản:

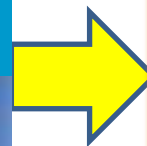
“Tính đến ngày 05/12/2021, trên thế giới, số ca nhiễm lên đến **265.801.429** người, trong đó có **5.266.133** người tử vong và **239.516.661** người khỏi bệnh. Tại Việt Nam, số ca nhiễm **1.309.092** người, số người tử vong **26.260** người, số người được điều trị khỏi bệnh **1.009.277** người”.  
( Theo nguồn từ Bộ Y tế Việt Nam)



Câu hỏi	Nội dung câu hỏi	Số ca (người)
1	<b>Trên thế giới số ca nhiễm</b> tính đến ngày 05/12/2021 là bao nhiêu?	<b>265.801.429</b>
2	<b>Tại Việt Nam số ca nhiễm</b> tính đến ngày 05/12/2021 là bao nhiêu?	<b>1.309.092</b>
3	<b>Trên thế giới số ca tử vong</b> tính đến ngày 05/12/2021 là bao nhiêu?	<b>5.266.133</b>
4	<b>Tại Việt Nam số ca khỏi bệnh</b> tính đến ngày 05/12/2021 là bao nhiêu?	<b>1.009.277</b>
5	.....tính đến ngày 05/12/2021 là bao nhiêu?	

## □ Thu thập dữ liệu thông qua đọc nội dung văn bản:











*“Theo tổng cục môi trường, Việt Nam có tổng số loài chim ghi nhận là 888 loài, trong đó có 72 loài chim hiện đang bị đe dọa tuyệt chủng ở mức độ toàn cầu, 51 loài ít xuất hiện và hiếm gặp” – Theo tạp chí môi trường Việt Nam 4/2017*

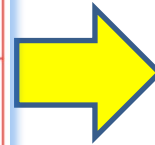


Câu hỏi	Nội dung câu hỏi	Số loài
1	<b>Việt Nam có tổng số loài chim là bao nhiêu?</b>	888
2	<b>Việt Nam có tổng số loài chim bị đe dọa tuyệt chủng là bao nhiêu?</b>	72
3	<b>Việt Nam có tổng số loài chim ít xuất hiện và hiếm gặp là bao nhiêu?</b>	51
4	<b>Việt Nam có tổng số loài thực vật là bao nhiêu?</b>	Không có thông tin

□ Thu thập dữ liệu bằng quan sát bảng:

## CÁC MÔN THỂ THAO ĐƯỢC YÊU THÍCH

Môn thể thao	Kiểm đếm	Số bạn ưa thích
Bóng đá 		18
Cầu lông 		8
Bóng bàn 		2
Đá cầu 		4
Bóng rổ 		5



a. Tựa đề của bảng bên là gì?

b. Máy cột ?

c. Máy dòng ?

d. Máy ô ?

**Các môn thể thao được yêu thích**

**3 cột**

**6 dòng**

**18 ô**



□ Thu thập dữ liệu từ thông tin của hình ảnh:

**CÁC BÌNH GA CỦA MỘT CỬA HÀNG ĐANG BÁN**



Câu	Câu hỏi	Trả lời
a.	Cửa hàng đang bán bao nhiêu bình ga?	
b.	Cửa hàng bán mấy loại bình ga?	

Câu	Màu của bình ga	Số lượng
a.		
b.		
c.		



- ☑ Phương pháp thu thập sơ cấp và thứ cấp
- ☑ Các dạng thu thập dữ liệu, chữ, số, hình ảnh, âm thanh..
- ☑ Tình huống thu thập thông tin từ website, công ty, thông tin nội dung...



FPT POLYTECHNIC

**Thank you**