FPT Education

FPT POLYTECHNIC

caodang.fpt.edu.vn

NHẬP MÔN XỬ LÝ DỮ LIỆU

BÀI 4: CHUYỂN ĐỔI DỮ LIỆU

Mục tiêu

- Hiểu tầm quan trọng chuyển đổi dữ liệu
- Thao tác các kỹ thuật chuyển đổi dữ liệu
- Các bước chuẩn hoá dữ liệu cơ bản

NỘI DUNG

- GIỚI THIỆU CHUYỂN ĐỔI DỮ LIỆU
- KỸ THUẬT CHUYỂN ĐỔI DỮ LIỆU

PHẦN I: CHUYỂN ĐỔI DỮ LIỆU

❑ Tại sao phải chuyển đổi dữ liệu thô ?

- ❖ Dữ liệu thô rất hiếm khi được **cấu trúc dữ liệu** rõ ràng hoặc được định dạng một cách hiệu quả để phục vụ cho việc **phân tích dữ liệu** trong một doanh nghiệp.

GIỚI THIỆU CHUYỂN ĐỔI DỮ LIỆU

❑ Tại sao phải chuyển đổi dữ liệu thô ?

- ❖ Vì thế, dữ liệu thô cần được áp dụng các **kỹ thuật chuyển đổi dữ liệu** để chuyển dữ liệu thô sang dạng có cấu trúc giúp cho việc **phân tích dữ liệu tiếp theo được dễ dàng và hiệu quả.**

GIỚI THIỆU CHUYỂN ĐỔI DỮ LIỆU

❑ Định nghĩa:

- ❖ Là quá trình **sửa đổi, tính toán, phân tách và kết hợp dữ liệu thô** thành dữ liệu có cấu trúc rõ ràng hơn để việc phục vụ cho việc phân tích dữ liệu được dễ dàng và hiệu quả hơn trong doanh nghiệp.

	A	B	C	D	E	F	G	H	I	J	K
1	Oder ID	SF County	Customer	Customer	Product	Product in	Product ID	Price	Quantity	Total price	Comments
2	42	Napa	marthy15(	1042	Denver sar	[object Ob	107	11	2	22	
3	43	Solano	vclemetts1	1043	Veggie bur	[object Ob	101	10	8	80	
4	43	Solano	vclemetts1	1043	Veggie bur	[object Ob	101	10	8	80	
5	44	Sonoma	amacfadin	1044	Sausage sa	[object Ob	103	9	14	126	
6	45	Contra Costa	lpaybody1	1045	Cheesebur	[object Ob	110	9	7	63	
7	46	Sonoma	jkhoter19(	1046	Veggie bur	[object Ob	101	10	7	70	
8	47	San Mateo	cquinane1	1047	Veggie bur	[object Ob	101	10	9	90	
9	48	San Mateo	pcornford1	1048	Cheesebur	[object Ob	110	9	7	63	
10	49	Santa Clara	dcupper1c	1049	Crab cake	[object Ob	109	13	11	143	
11	49	Santa Clara	dcupper1c	1049	Crab cake	[object Ob	109	13	11	143	
12	50	Solano	bdillingsto	1050	Ham and c	[object Ob	106	12	8	96	
13	51	Contra Costa	dkeasy1e(	1051	Crab cake	[object Ob	109	13	8	104	
14	52	Santa Clara	mbonnese	1052	Cheesebur	[object Ob	110	9	13	117	
15	53	Santa Clara	djore1g@g	1053	Jucy Lucy	[object Ob	104	8	9	72	
16	54	Alameda	cdunseath	1054	Jucy Lucy	[object Ob	104	8	11	88	
17	55	San Francisco	tkennerma	1055	Denver sar	[object Ob	107	11	9	99	
18	56	Alameda	apetrasek1	1056	Crab cake	[object Ob	109	13	9	117	

GIỚI THIỆU CHUYỂN ĐỔI DỮ LIỆU

❑ Định nghĩa:

- ❖ Cấu trúc dữ liệu đó là những đại diện thực tế dễ dàng **chuyển thành chỉ số**, **báo cáo** và trang tổng quan để giúp người dùng hoàn thành các mục tiêu cụ thể.

Chuẩn hóa dữ liệu:

- Đảm bảo các giá trị chính xác và được tổ chức theo cách hỗ trợ mục đích sử dụng của chúng.
- **Chuẩn hóa cơ sở dữ liệu** là một hình thức sửa đổi dữ liệu bằng cách giảm mô hình dữ liệu về dạng **không có dư thừa** hoặc một-nhiều giá trị trong một cột.

❑ Chuẩn hóa dữ liệu:

❖ **Chuẩn hóa** làm giảm nhu cầu lưu trữ và làm cho mô hình dữ liệu ngắn gọn hơn và dễ đọc hơn đối với các nhà phân tích.

isbn	title	author	nationality	format	price	subject
194503040502	Soviet Infantry Doctrine in WWII	Chet Rogers	American	Hardcover	49.99	Military Science
						History
						Firearms

isbn	title	author	nationality	format	price
194503040502	Soviet Infantry Doctrine in WWII	Chet Rogers	American	Hardcover	49.99

isbn	subject
194503040502	military science
194503040502	history
194503040502	firearms

❑ Chuyển đổi định dạng:

- ❖ Thay thế các ký tự không tương thích.
- ❖ Chuyển đổi đơn vị.
- ❖ Chuyển đổi định dạng ngày tháng.
- ❖ Chuyển đổi kiểu dữ liệu.

name	adoption_fee
Maisie	"380"

→

name	adoption_fee
Maisie	380.00

❑ Xóa các hàng trùng nhau:

❖ Sử dụng **Deduplication** có nghĩa xác định và loại bỏ các bản ghi trùng lặp.

name	breed	date_of_birth	color	weight_lbs
Maisie	pitbull	07/14/2017	brown	47
maisie	pibble	07/14/2017	brown	47

→

name	breed	date_of_birth	color	weight_lbs
Maisie	pitbull	07/14/2017	brown	47

❑ Việc loại bỏ các cột không sử dụng và lặp lại:

❖ Giúp cải thiện hiệu suất và tính dễ đọc tổng thể của cấu trúc dữ liệu.

name	breed	color	weight_lbs	weight_kilos
Maisie	pitbull	brown	47	21.4

→

name	breed	color	weight_kilos
Maisie	pitbull	brown	21.4

❑ Xác thực dữ liệu:

❖ Đánh giá tính hợp lệ của một bản ghi bằng tính **đầy đủ của dữ liệu**, thường bằng cách loại trừ các bản ghi không đầy đủ.

name	breed	date_of_birth	color	weight_lbs
Maisie	pitbull	07/14/2017	brown	47
NULL	*NULL*	*NULL*	merle	62

→

name	breed	date_of_birth	color	weight_lbs
Maisie	pitbull	07/14/2017	brown	47

Tin học:

- Tính toán các **giá trị dữ liệu mới** từ dữ liệu hiện có là tính toán tỷ lệ, thống kê tóm tắt và các số liệu quan trọng khác.
- Biến dữ liệu phi cấu trúc từ các tệp phương tiện thành **dữ liệu có cấu trúc**.

admissions	applications
345	14556

→

admissions	applications	acceptance_rate
345	14556	0.0237

PHẦN II: CHUYỂN ĐỔI DỮ LIỆU (TT)

❑ Xoay vòng:

- ❖ Chuyển các giá trị hàng thành cột và ngược lại.

time	activity
1/1/2020	purchase
1/1/2020	return
1/1/2020	purchase
1/3/2020	return
1/3/2020	purchase
1/3/2020	purchase
1/4/2020	purchase
1/4/2020	return
1/4/2020	purchase
1/4/2020	purchase

time	count_purchase	count_return
1/1/2020	2	1
1/2/2020	0	1
1/3/2020	2	0
1/4/2020	3	1
1/5/2020	4	0
...	...	...

❑ Sắp xếp và lập chỉ mục:

❖ Tổ chức các dòng theo một số thứ tự để cải thiện hiệu suất tìm kiếm.

student_id	first_name	last_name
4321	Archibald	Barry
2534	Brittany	Columbus
6633	Chad	Daniels
7787	Desmond	Ephram
1235	Eleanor	Fox
5432	Florence	Graham
5155	Grant	Hammond
3151	Helen	Ines
6675	Isabelle	Jackson
4515	Janet	King
5151	Katya	Luther
5167	Lance	Mondale
5566	Martin	Newman
1423	Nestor	Osbourne
6677	Olivia	Partridge
8897	Peyton	Quinn

student_id	first_name	last_name
1235	Eleanor	Fox
1423	Nestor	Osbourne
2534	Brittany	Columbus
3151	Helen	Ines
4321	Archibald	Barry
4515	Janet	King
5151	Katya	Luther
5155	Grant	Hammond
5167	Lance	Mondale
5432	Florence	Graham
5566	Martin	Newman
6633	Chad	Daniels
6675	Isabelle	Jackson
6677	Olivia	Partridge
7787	Desmond	Ephram
8897	Peyton	Quinn

❑ Chia tỷ lệ và chuẩn hóa:

- ❖ Thiết lập các con số trên một thang đo nhất quán.
- ❖ Như các phân số của độ lệch chuẩn trong chuẩn hóa điểm Z.
- ❖ Điều này cho phép các con số khác nhau được so sánh với nhau.

student_id	cum_sat
4321	1350
2534	1220
6633	1600
7787	1550
1235	1440
5432	1410
5155	1040
3151	800

→

student_id	cum_sat	sat_z_score	sat_min_max_scaling
4321	1350	0.926417163	0.7474747475
2534	1220	0.544086270	0.6161616162
6633	1600	1.66166888	1
7787	1550	1.514618537	0.9494949495
1235	1440	1.191107781	0.8383838384
5432	1410	1.102877575	0.8080808081
5155	1040	0.01470503434	0.4343434343
3151	800	-0.6911366138	0.1919191919

❑ Vectơ hóa dữ liệu:

❖ Chuyển đổi dữ liệu không có trúc không phải số thành dữ liệu bảng .

About the bird, the bird, bird bird bird
You heard about the bird
The bird is the word

→

phrase	about	bird	heard	is	the	word	you
"About the bird, the bird, bird bird bird"	1	5	0	0	2	0	0
"You heard about the bird"	1	1	1	0	1	0	1
"The bird is the word"	0	1	0	1	2	1	0

Tách biệt:

- Phân chia các giá trị thành các phần cấu thành của chúng.
- Các giá trị dữ liệu thường được kết hợp trong cùng một cột vì tính riêng trong thu thập dữ liệu, nhưng có thể cần được tách riêng để thực hiện phân tích chi tiết hơn.

name	breed_mix
Maisie	pitbull | australian shepherd | labrador retriever | australian cattle dog
Tacoma	husky | pitbull | australian shepherd | australian cattle dog

name	australian_cattle_dog	australian_shepherd	husky	labrador_retriever	pitbull
Maisie	1	1	0	1	1
Tacoma	1	1	1	0	1

❑ Lọc dữ liệu:

❖ Loại trừ dữ liệu trên cơ sở các giá trị hàng hoặc cột nhất định.

time	activity	location
1/1/2020	purchase	New York, NY
1/1/2020	return	Chicago, IL
1/1/2020	purchase	Atlanta, GA
1/2/2020	return	Atlanta, GA
1/3/2020	purchase	New York, NY
1/3/2020	purchase	New York, NY
1/4/2020	purchase	New York, NY
1/4/2020	return	New York, NY
1/4/2020	purchase	New York, NY
1/4/2020	purchase	Washington, DC
1/5/2020	purchase	Washington, DC
1/5/2020	purchase	Washington, DC

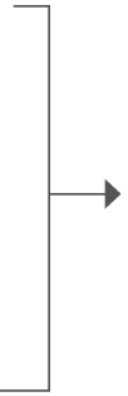
time	activity	location
1/1/2020	purchase	New York, NY
1/3/2020	purchase	New York, NY
1/3/2020	purchase	New York, NY
1/4/2020	purchase	New York, NY
1/4/2020	return	New York, NY
1/4/2020	purchase	New York, NY

❑ Kết hợp:

- ❖ kết hợp các dòng từ nhiều bảng khác nhau và dữ liệu từ nhiều nguồn để xây dựng bức tranh đầy đủ về các hoạt động của tổ chức.

id	name	city
1337	Elite Academy	New York
8455	Lakeside Academy	Chicago
4377	Armitage High	Saint Louis
8088	Mountaintop High	Denver

id	acceptance_rate
1337	0.67
8455	0.23
4377	0.45
8088	0.56



id	name	city	acceptance_rate
1337	Elite Academy	New York	0.67
8455	Lakeside Academy	Chicago	0.23
4377	Armitage High	Saint Louis	0.45
8088	Mountaintop High	Denver	0.56

❑ Các bước chuẩn hoá dữ liệu:

1. Lên ý tưởng, khảo sát nghiệp vụ
2. Xác định các yếu tố đầu vào
3. Xác định các yếu tố đầu ra (báo cáo, dashboard)
4. Xây dựng bố cục dữ liệu quan hệ trong Excel

TÓM TẮT BÀI HỌC

- ☑ Chuyển đổi dữ liệu đóng vai trò quan trọng trong xử lý dữ liệu.
- ☑ Áp dụng các kỹ thuật chuyển đổi dữ liệu để chuyển dữ liệu dạng thô không cấu trúc sang dạng dữ liệu có cấu trúc, từ đó giúp cho phân tích dữ liệu được dễ dàng và hiệu quả.
- ☑ Cần áp dụng qui trình chuẩn hoá dữ liệu để việc chuẩn hoá được hiểu quả


FPT Education
FPT POLYTECHNIC
Thank you
TIN HỌC VĂN PHÒNG QUỐC TẾ THEO CHUẨN ICDL
Syllabus 5 - Office 2007
LUẬT KINH TẾ
KỸ THUẬT SOẠN THẢO VÀ TRÌNH BÀY VĂN BẢN