



## Bài 5.1. Cơ bản về CSDL thống kê



# CHƯƠNG 5. AN TOÀN CƠ SỞ DỮ LIỆU THỐNG KÊ

TS. Trần Thị Lượng  
\* Khoa An toàn thông tin \*

# NỘI DUNG



1

Giới thiệu

2

Các khái niệm cơ bản

3

Các tấn công vào cơ sở dữ liệu thống kê



## □ Kiến thức

- Hiểu được khái niệm về CSDL thống kê
- Nắm được các dạng biểu diễn của CSDL thống kê
- Hiểu được ứng dụng của các CSDL thống kê trong thực tế
- Các tấn công vào CSDL thống kê



## □ Kỹ năng

- Thực hiện tạo CSDL thống kê, thực hiện các truy vấn thống kê trên CSDL thống kê
- Thực hành các truy vấn để thực hiện các tấn công đơn giản vào CSDL thống kê.



1. Giáo trình “*An toàn cơ sở dữ liệu*”// Chương 4 “*An toàn cơ sở dữ liệu thống kê*”
2. Nabil R.Adam, John C. Wortmann, *Security-Control Methods for Statistical Databases: A comparative Study*, ACM Computing Surveys, Vol. 21, No.4, December 1989.
3. Shiuh-Pyng Shieh, Chern-Tang Lin, *Information Protection in Dynamic Statistical*, National Chiao Tung University Databases.





- **Website:**

- **Tổng cục thống kê – Việt Nam:**

- <http://www.gso.gov.vn/default.aspx?tabid=228&ItemID=1915>

- **SDB Liên hợp quốc (UNO)**

- <http://unstats.un.org/unsd/databases.htm>

- **SDB kinh tế khối Châu Âu (UNECE)**

- <http://w3.unece.org/pxweb/Dialog/>

- **SDB WTO**

- <http://stat.wto.org/Home/WSDBHome.aspx?Language=>

- V.V

# NỘI DUNG



1

Giới thiệu

2

Các khái niệm cơ bản

3

Các tấn công vào cơ sở  
dữ liệu thống kê

# GIỚI THIỆU



- Khái niệm về CSDL thống kê
- Các ví dụ về CSDL thống kê
- Một số câu truy vấn thống kê
- Ứng dụng của CSDL thống kê
- Vấn đề an toàn trong CSDL thống kê



# CÂU HỎI



- CSDL thống kê là gì?
- CSDL thống kê khác CSDL quan hệ như thế nào?



# KHÁI NIỆM VỀ CSDL THỐNG KÊ



- CSDL thống kê (SDB- Statistical database)
  - Là một CSDL được sử dụng cho mục đích phân tích thống kê.
  - Là một CSDL chứa các bản ghi nhạỵ cảm mô tả về các cá nhân nhưng chỉ các câu truy vấn thống kê (như: **COUNT, SUM, AVERAGE, MAX, MIN...**) mới được trả lời, ngoài các câu truy vấn này thì những truy vấn vào các mục dữ liệu riêng sẽ không được đáp lại.

# KHÁI NIỆM VỀ CSDL THỐNG KÊ



SELECT COUNT...

SELECT SUM...



Success

SELECT MIN...



SELECT MAX...

(SDB)

Database CMYK icon



SELECT Eno...

Eno, Ename, Salary

SELECT Ename...



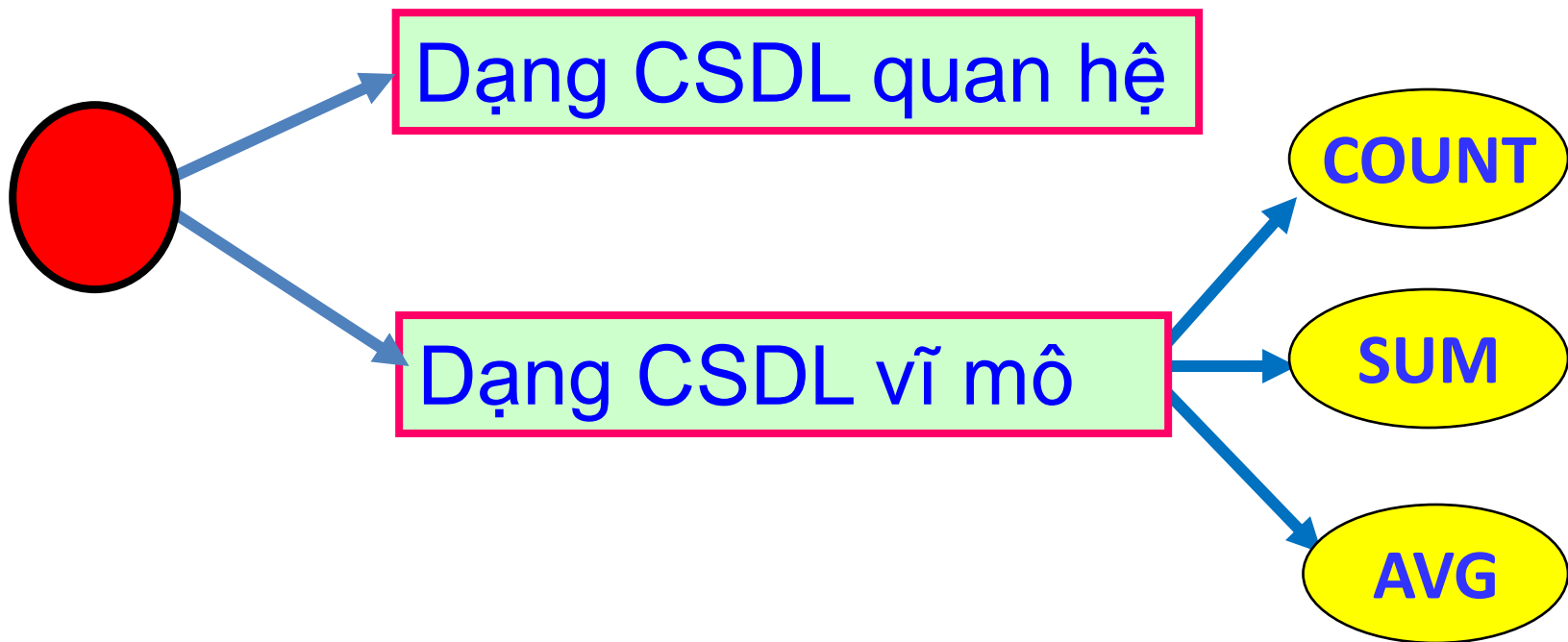
SELECT Salary...

Failed!

# GIỚI THIỆU



- Có hai dạng SDB cơ bản:



# GIỚI THIỆU



- Khái niệm về CSDL thống kê
- Các ví dụ về CSDL thống kê
- Một số câu truy vấn thống kê
- Ứng dụng của CSDL thống kê
- Vấn đề an toàn trong CSDL thống kê



# VÍ DỤ VỀ SDB



Dạng quan hệ

SDB về công nhân

ID	Tên	Chức vụ	Phòng	Tuổi	Giới tính	Lương
01	Nam	Nhân viên	Maketing	29	M	3500
02	Lan	Trưởng phòng	Kế hoạch	33	F	6200
03	Huệ	Nhân viên	Kế hoạch	27	F	4000
04	Minh	Giám sát viên	Maketing	24	M	3600
05	Quỳnh	Nhân viên	Kế hoạch	24	F	2900

# VÍ DỤ VỀ SDB ...



**Dạng quan hệ**

**SDB về công nhân**

ID	Tên	Chức vụ	Phòng	Tuổi	Giới tính	Lương
01	Nam	Nhân viên	Maketing	29	M	3500
02	Lan	Trưởng phong	Kế hoạch	33	F	6200
03	Huệ	Nhân viên	Kế hoạch	27	F	4000
04	Minh	Giám sát viên	Maketing	24	M	3600
05	Quỳnh	Nhân viên	Kế hoạch	24	F	2900



# VÍ DỤ VỀ SDB ...



**Dạng quan hệ**

*SDB về các vụ tai nạn ô tô*

HoTen	Tuoi	Đ/C	MauXe	LoaiXe	ThoiGian	CoLoi	SayRuou
Nguyễn Văn Tài	25	HN	Xanh	Honda	13.30	1	1
Lê sỹ Hoàng	37	HD	Đỏ	Toyota	6.25	1	0
Hoàng Văn Minh	42	PT	Trắng	Audi	17.45	0	0
Vũ Bình Minh	32	PT	Vàng	Volkswagen	3.30	0	1
Trần Quang Hòa	22	HN	Xanh	Honda	6.30	1	0

# VÍ DỤ VỀ SDB ...



## Dạng quan hệ

## *SDB về các Sinh viên*

Tên	Giới tính	Địa chỉ	Phụ cấp	Lớp
Minh	M	HN	500	Toán1
Hải	M	HD	0	Toán2
Tuyết	F	NĐ	300	Tin1
Nam	M	BG	100	Tin2
Phương	F	NA	200	Toán2
Hạnh	F	HT	100	Toán1

# VÍ DỤ VỀ SDB ...



**Dạng quan hệ**

*SDB về đảng viên*

MaDV	HoTen	DiaChi	ChucVu	Luong	DangVie n
MA01	Trần Văn Nguyên	Hà Nội	Trưởng phòng	3000	1
MA02	Nguyễn Thị Hoa	Hải Phòng	Nhân viên	2000	0
MA03	Vũ Văn Hiền	Hà Nội	Phó Giám đốc	4000	1
MA04	Trần Thị Mai	Nghệ An	Trưởng phòng	3000	1
MA05	Nguyễn Quang Huy	Hải Phòng	Giám đốc	5000	1
MA06	Trần Văn Hải	Hà Nam	Nhân viên	2000	1
MA07	Lê Minh Sơn	Nam Định	Nhân viên	2500	0



# VÍ DỤ VỀ SDB ...



Dạng vĩ mô

*SDB vĩ mô về các Sinh viên*

SUM

	AT4A	AT4B	AT4C	AT3
M	500	0	0	100
F	100	200	300	0
Tổng cộng	600	200	300	100

*Tổng phụ cấp theo giới tính và theo lớp*

# VÍ DỤ VỀ SDB ...



Dạng vĩ mô

*SDB vĩ mô về Công nhân*

COUNT

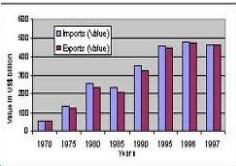
Năm sinh	Giới tính	Mã phòng		
		Phòng1	Phòng2	Phòng3
1941-1951	M	10	12	0
	F	1	0	3
1952-1962	M	12	10	5
	F	20	2	8
>1962	M	15	0	1
	F	20	10	0

# GIỚI THIỆU



- Khái niệm về CSDL thống kê
- Các ví dụ về CSDL thống kê
- Một số câu truy vấn thống kê
- Ứng dụng của CSDL thống kê
- Vấn đề an toàn trong CSDL thống kê

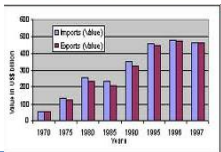
# VÍ DỤ MỘT SỐ CÂU TRUY VẤN THỐNG KÊ



## COUNT

- **Select** count(\*) **from** Nhanvien  
(Trả lại tổng số lượng các bg trong table)
- **Select** count(\*) **from** nhanvien **where**  
Luong<=1000
- **Select** count(Luong) **AS** count\_Luong **from**  
Nhanvien
- **Select** count(Distinct Luong) **from** Nhanvien  
(Trả lại số lượng các loại lương phân biệt nhau)

# VÍ DỤ MỘT SỐ CÂU TRUY VẤN THỐNG KÊ

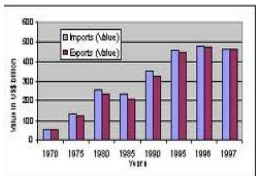


**SUM**

- **Select** SUM(Luong) **as** sum\_Luong **from** Nhanvien
- **Select** SUM(Distinct Luong) **as** sum\_Luong **from** Nhanvien
- **Select** Chucvu, Sum(Luong) **from** Nhanvien **GROUP BY** chucvu
- **Select** HoTen, chucvu, Luong **from** nhanvien **ORDER** by chucvu  
**Compute** SUM(Luong) **by** chucvu  
(Thêm cột tổng lương với từng kiểu chức vụ)



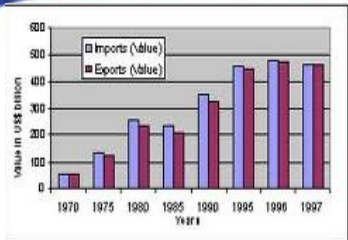
# VÍ DỤ MỘT SỐ CÂU TRUY VẤN THỐNG KÊ



**AVG**

- **Select** AVG(Luong) **AS** avg\_Luong **from** Nhanvien
  - **Select** AVG(Luong) **AS** avg\_Luong **from** Nhanvien **where** Luong>1000
  - **Select** AVG(distinct Luong) **AS** avg\_Luong **from** Nhanvien
  - **Select** chucvu, AVG(Luong) **AS** avg\_Luong, SUM(Luong) **as** sum\_luong **from** Nhanvien
- Group by** chucvu
- Order by** chucvu

# VÍ DỤ MỘT SỐ CÂU TRUY VẤN THỐNG KÊ



**MIN**

- **Select** MIN(Luong) **from** Nhanvien
- **Select** MIN(Distinct Luong) **from** Nhanvien

**MAX**

- **Select** MAX(Distinct Luong) **from** Nhanvien
- **Select** MAX(Luong) **from** Nhanvien

# GIỚI THIỆU

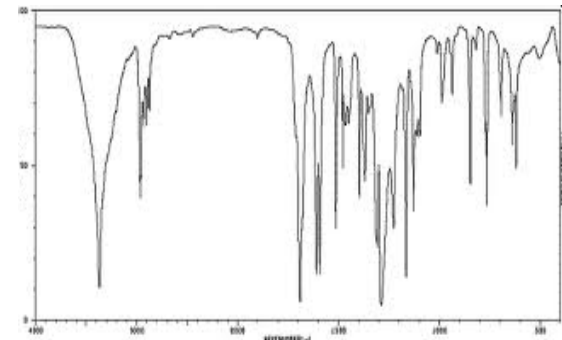


- Khái niệm về CSDL thống kê
- Các ví dụ về CSDL thống kê
- Một số câu truy vấn thống kê
- Ứng dụng của CSDL thống kê
- Vấn đề an toàn trong CSDL thống kê

# ỨNG DỤNG CỦA SDB



- Điều tra dân số
- Thống kê về số người tử vong
- Về kế hoạch kinh tế
- Thống kê về khám chữa bệnh
- Về các vụ tai nạn ô tô
- Thống kê về tội phạm
- ...



# ỨNG DỤNG CỦA SDB...



- Thống kê nông nghiệp, lâm nghiệp, thủy sản
- Thống kê ngành nghề kinh doanh
- Giáo dục và nghiên cứu
- Môi trường
- Thị trường tài chính
- Giá cả và tiêu dùng
- Tài chính công
- Thương mại hàng hoá và dịch vụ
- ...



**Phân tích và đưa ra chiến lược!**



# CÂU HỎI NGHIÊN CỨU



- Ứng dụng SDB ở Việt Nam?
- Ứng dụng SDB trên thế giới?



# GIỚI THIỆU



- Khái niệm về CSDL thống kê
- Các ví dụ về CSDL thống kê
- Một số câu truy vấn thống kê
- Ứng dụng của CSDL thống kê
- Vấn đề an toàn trong CSDL thống kê

# CÂU HỎI



- Tại sao phải bảo vệ SDB?

## Dàn xếp giữa:

- Yêu cầu bảo vệ thông tin riêng tư của các cá nhân
- Và quyền truy xuất và xử lý thông tin của các tổ chức

Vì SDB chứa dữ liệu thống kê liên quan đến **thông tin nhạy cảm** của nhiều cá nhân

# CÂU HỎI



- Tấn công vào các SDB bằng cách nào?

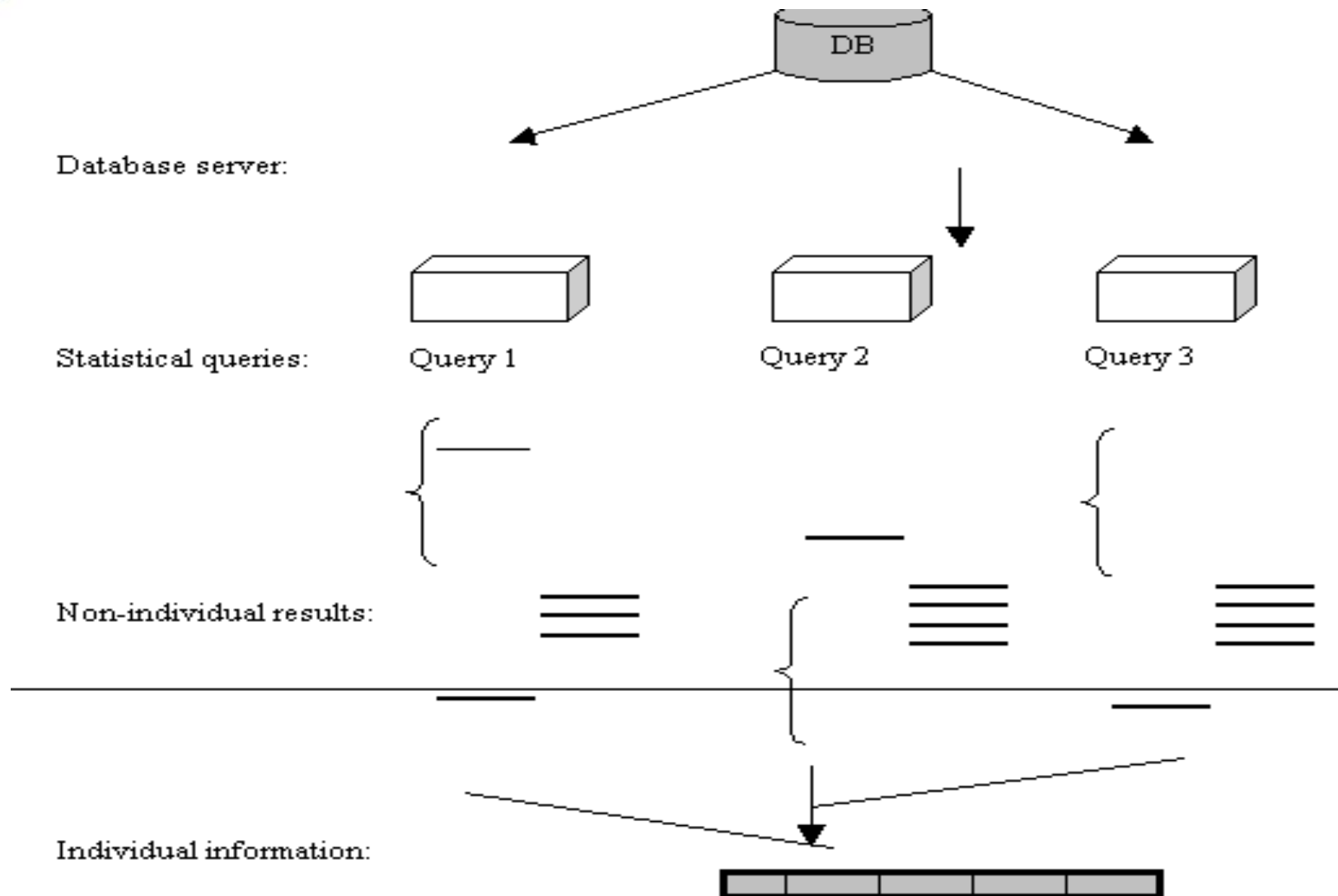
Kết hợp các  
câu truy vấn  
thống kê

Tấn công suy  
diễn  
(Interference  
attack)

Thu được  
thông tin bí  
mật về một cá  
nhân

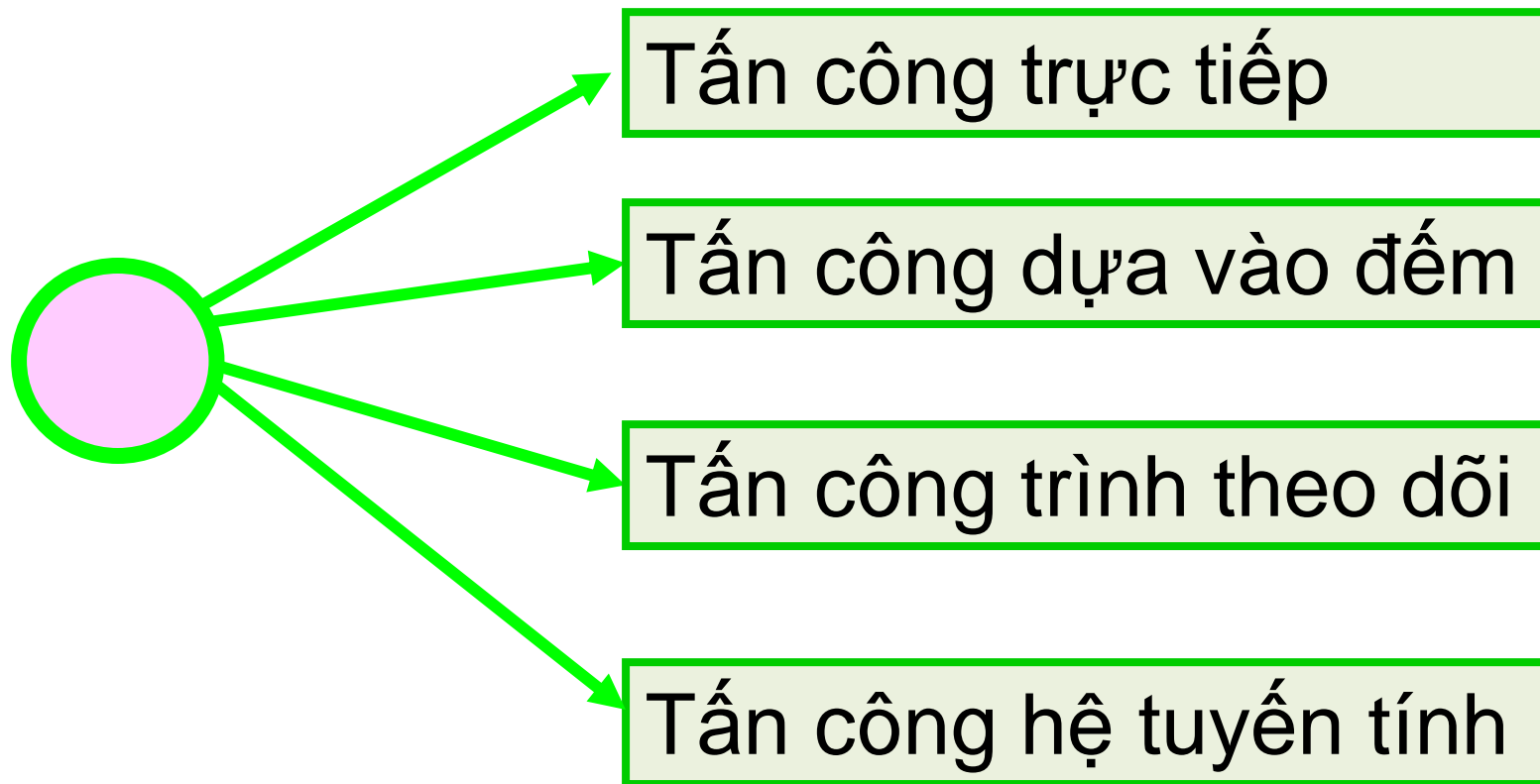
COUNT SUM MIN MAX AVG

# KIẾN TRÚC MỘT TẦN CÔNG SUY DIỄN





# MỘT SỐ KIỂU TẤN CÔNG SUY DIỄN



# NỘI DUNG



1

Giới thiệu

2

Các khái niệm cơ bản

3

Các tấn công vào cơ sở dữ liệu thống kê

# CÁC KHÁI NIỆM CƠ BẢN



- Các đặc tính của SDB cần được bảo vệ:

- *SDB tĩnh:*

- SDB không thay đổi trong suốt thời gian tồn tại của chúng.
- Ví dụ: CSDL thống kê dân số

# CÁC KHÁI NIỆM CƠ BẢN



- Các đặc tính của SDB cần được bảo vệ:
  - *SDB động*:
    - Thay đổi liên tục theo sự thay đổi của dữ liệu thực, cho phép sửa đổi để phản ánh các thay đổi động của thế giới thực
    - Ví dụ các CSDL nghiên cứu trực tuyến, lớp học trực tuyến khi bổ sung thành viên,....

# CÁC KHÁI NIỆM CƠ BẢN



- Các đặc tính của SDB cần được bảo vệ:
  - SDB trực tuyến (online):
    - Người sử dụng nhận được các phản hồi thời gian thực cho các câu truy vấn thống kê của mình.
  - SDB ngoại tuyến (offline):
    - Người sử dụng không biết khi nào các thống kê của họ được xử lý, việc SDB bị lộ sẽ khó khăn.



# CÁC KHÁI NIỆM CƠ BẢN



- Các đặc tính của SDB cần được bảo vệ:
  - Kiến thức làm việc (working knowledge)
    - Là tập các mục thông tin (field) và giá trị thuộc tính trong SDB và các kiểu thống kê có sẵn trong SDB mà người dùng có thể biết một cách hợp lệ.

# CÁC KHÁI NIỆM CƠ BẢN



- Các đặc tính của SDB cần được bảo vệ:
  - Kiến thức bổ sung của người sử dụng (supplementary knowledge):
    - Người sử dụng không biết khi nào các thống kê của họ được xử lý, việc SDB bị lộ sẽ khó khăn.

# MÔ HÌNH LÀM LỘ SDB

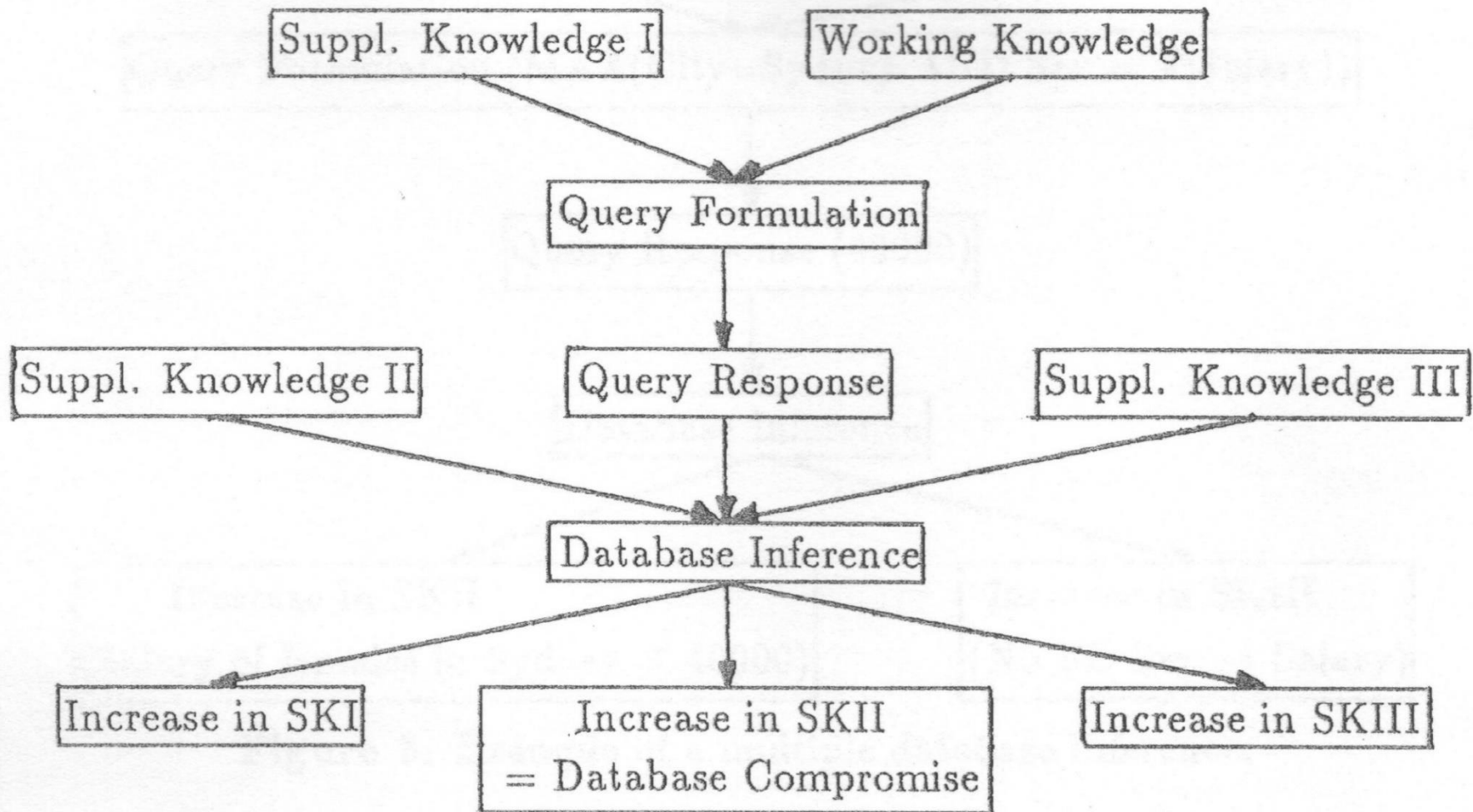


Figure 1. Database compromise and supplementary knowledge.



# VÍ DỤ



## Nhan Viên

ID	Ten	ChucVu	PhongLV	Que	GioiTinh	Lương
01	Nam	Nhân viên	Maketing	Hải Phòng	F	3500
02	Lan	Trưởng phong	Kế hoạch	Hà Nội	M	6200
03	Huệ	Nhân viên	Kế hoạch	Nam Định	M	4000
04	Minh	Giám sát viên	Maketing	Bắc Giang	F	3600
05	Quỳnh	Nhân viên	Kế hoạch	Hà Nội	F	2900



# VÍ DỤ VỀ LÀM LỘ MỘT SDB (LỘ CHÍNH XÁC)



WK = Nam Định  $\in$  Que

SK1 = Chỉ có 1 công nhân  
ở Nam Định

Query { MIN(Luong, que= "Nam Định" ) }

Result (4000)

Database  
Interference

SK2 = Lộ SDB (Lương của 1  
công nhân ở Nam Định là 4000)



# VÍ DỤ VỀ LÀM LỘ MỘT SDB (LỘ XẤP XỈ)



WK = Giới Tính gồm {M, F}

SK1 = Lương trong khoảng  
[0, 7000]

Query { MAX(Luong, Giới Tính= "F" ) }

Result (4500)

Database  
Interference

SK2 = Lộ SDB (Lương của tất cả các nữ  
công nhân đều  $\leq 4500$ )

Tiếp tục tìm  
Min(Lương,  
Giới Tính="F")

# CÁC KHÁI NIỆM CƠ BẢN



- Công thức đặc trưng:
  - Là một công thức lôgíc, được ký hiệu bởi một chữ cái viết hoa ( $A, B, C, \dots$ ),, trong đó các giá trị thuộc tính được kết hợp với nhau thông qua các toán tử Boolean như OR, AND, NOT ( $\vee, \wedge, \neg$ ).
  - Ví dụ:
$$C = (GioiTinh=F) \wedge [(MaPhong="Kế hoạch") \vee (MaPhong="Tài vụ")] \wedge (NamSinh < 1965)$$

# CÁC KHÁI NIỆM CƠ BẢN



- Tập truy vấn (query set): của một công thức đặc trưng  $C$  là tập tất cả các bản ghi thỏa mãn  $C$ .
  - Ký hiệu là  $X(C)$ .
- Thống kê trên  $C$ : là các câu truy vấn thống kê trên  $C$ 
  - Ký hiệu:  $q(C)$
  - Chẳng hạn:  $COUNT(C)$ ,  $SUM(C, A_j)$ ,  
 $MIN(C, A_j)$ ,  $MAX(C, A_j)$



# VÍ DỤ



## Nhan Viên

ID	Ten	ChucVu	PhongLV	Que	GioiTinh	Lương
01	Nam	Nhân viên	Maketing	Hải Phòng	F	3500
02	Lan	Trưởng phong	Kế hoạch	Hà Nội	M	6200
03	Huệ	Nhân viên	Kế hoạch	Nam Định	M	4000
04	Minh	Giám sát viên	Maketing	Bắc Giang	F	3600
05	Quỳnh	Nhân viên	Kế hoạch	Hà Nội	F	2900

# VÍ DỤ



- Công thức đặc trưng C

$$C = \{(GioiTinh = F) \wedge [(MaPhong = \text{"Kế hoạch"} \vee (MaPhong = \text{"Tài vụ"}))]\}$$

- Tập truy vấn X(C)

ID	Ten	ChucVu	PhongLV	Que	GioiTinh	Lương
03	Huệ	Nhân viên	Tài vụ	Nam Định	F	4000
05	Quỳnh	Nhân viên	Kế hoạch	Hà Nội	F	2900

- Thống kê trên C: Count(C), Sum(C, Lương), MAX(C, Lương), MIN(C, Lương)



# CÁC KHÁI NIỆM CƠ BẢN



- **Khái niệm bậc:** Một thống kê gồm  $m$  thuộc tính khác nhau được gọi là thống kê **bậc  $m$** .
  - Ví dụ:
    - +  $Count ((GioiTinh = F) \wedge (MaPhong = Phong1))$  là một thống kê bậc 2.
    - +  $Count(*)$  là thống kê bậc 0.

# CÁC KHÁI NIỆM CƠ BẢN



- **Khái niệm thống kê nhạy cảm:** Thống kê được tính toán trên một *thuộc tính bí mật* trong tập truy vấn có kích cỡ bằng 1 là thống kê nhạy cảm.
  - Ví dụ:  $\text{COUNT}(\text{AGE} > 50) = 1$   
 $\Rightarrow \text{SUM}(\text{Salary}, \text{age} > 50)$  là thống kê nhạy cảm



1

Giới thiệu

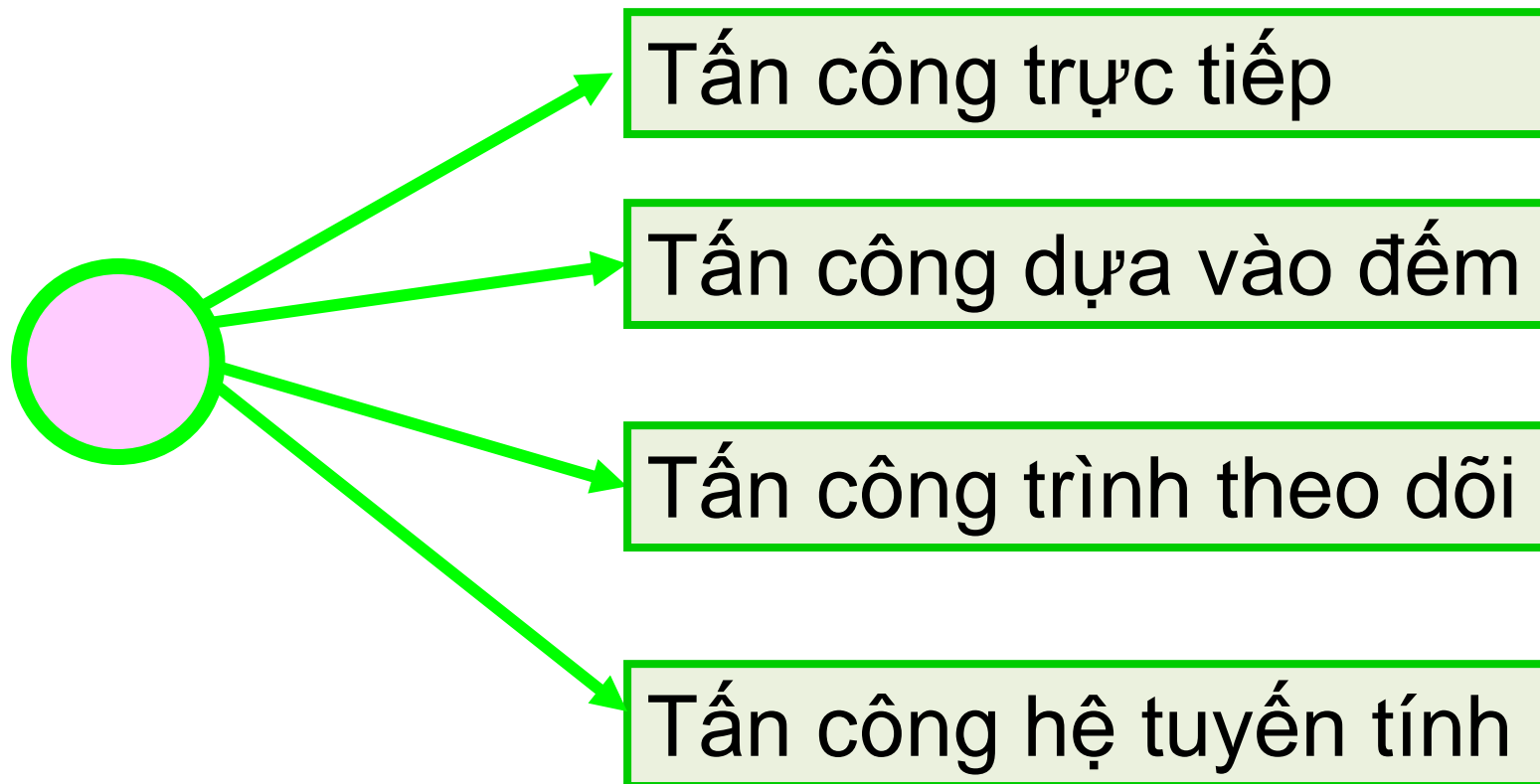
2

Các khái niệm cơ bản

3

Các tấn công vào cơ sở dữ liệu thống kê

# MỘT SỐ KIỂU TẤN CÔNG SUY DIỄN





# MỘT SỐ KIỂU TẦN CÔNG SUY DIỄN...

## Nhân Viên

ID	Ten	ChucVu	PhongLV	Que	GioiTinh	Lương
01	Nam	Nhân viên	Maketing	Hải Phòng	F	3500
02	Lan	Trưởng phong	Kế hoạch	Hà Nội	M	6200
03	Huệ	Nhân viên	Kế hoạch	Nam Định	M	4000
04	Minh	Giám sát viên	Maketing	Bắc Giang	F	3600
05	Quỳnh	Nhân viên	Kế hoạch	Hà Nội	F	2900



# MỘT SỐ KIỂU TẤN CÔNG SUY DIỄN...



## *Tấn công trực tiếp:*

- Sử dụng các câu truy vấn thông thường, không phải truy vấn thống kê.
- Ví dụ:

```
SELECT Ten FROM NhanVien WHERE Luong>4.360
```



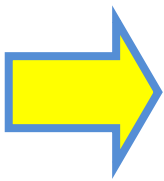
**Bộ lọc - Filter (loại các truy vấn không hợp lệ)**

# MỘT SỐ KIỂU TẤN CÔNG SUY DIỄN...

## *Tấn công dựa vào đếm*

- Đây là loại tấn công bằng cách kết hợp giá trị đếm với giá trị tổng để thu được thông tin bí mật.
- Ví dụ:

COUNT ( ChucVu = “Trưởng phòng”, Phong= “Kế hoạch” ) = 1



*SUM (Luong, (ChucVu= “Trưởng phòng”,  
Phong= “Kế hoạch”))*

# MỘT SỐ KIỂU TẤN CÔNG SUY DIỄN...



*Tấn công trình theo dõi (...)*

*Tấn công hệ tuyến tính (...)*

Sau Kiểm  
soát kích  
cỡ tập truy  
vấn

# NỘI DUNG



1

Giới thiệu

2

Các khái niệm cơ bản

3

Các tấn công vào cơ sở dữ liệu thống kê

# BÀI TẬP VỀ NHÀ



- So sánh SDB và CSDL quan hệ
- Thực hành các câu lệnh truy vấn thống kê
- Tìm hiểu ứng dụng SDB ở Việt Nam và trên thế giới?
- Tìm hiểu các tấn công suy diễn vào SDB



