# Tổng quan
# Tích hợp dữ liệu

Nguyễn Hồng Phương
Email: phuong.nguyenhong@hust.edu.vn
Site: http://is.hust.edu.vn/~phuongnh
Bộ môn Hệ thống thông tin
Viện Công nghệ thông tin và Truyền thông
Đại học Bách Khoa Hà Nội

1

## Nội dung

1. Nhu cầu tích hợp dữ liệu (THDL)
2. Mục tiêu của tích hợp dữ liệu
3. Tại sao THDL là vấn đề khó?
4. Các kỳ vọng
5. Kiến trúc THDL
6. Ví dụ về THDL
7. Các bước xử lý truy vấn

2

## 1. Nhu cầu tích hợp dữ liệu

- FullServe
- Website
- Các ngành khác

3

## FullServe

- Tập đoàn FullServe (Hoa Kỳ) cung cấp dịch vụ truy cập internet cho các gia đình và bán một số sản phẩm hạ tầng tính toán trong gia đình như modem, wireless router, voice-over-IP phone, máy pha cà phê.
- FullServe mở rộng thị trường sang châu Âu, sát nhập công ty EuroCard – công ty chuyên cung cấp thẻ tín dụng, cũng đang muốn tham gia thị trường internet

4

## FullServe

- Số lượng CSDL của FullServe là 100

**Employee Database**
FullTimeEmps(ssn, empID, firstName, middleName, lastName)
Hire(empID, hireDate, recruiter)
TempEmployees(ssn, hireStart, hireEnd, name, hourlyRate)

**Resume Database**
Interviews(interviewDate, pID, recruiter, hireDecision, hireDate)
CVs(ID, resume)

**Training Database**
Courses(courseID, name, instructor)
Enrollments(courseID, empID, date)

5

## FullServe

**Sales Database**
Products(prodName, prodID)
Sales(prodID, customerID, custName, address)

**Services Database**
Services(packName, textDescription)
Customers(name, ID, zipCode, streetAdr, phone)
Contracts(custID, packName, startDate)

**HelpLine Database**
Calls(date, agent, custID, text, action)

6

## FullServe

- CSDL của EuroCard

**Employee Database**
Emp(ID, firstNameMiddleInitial, lastName, salary)
Hire(ID, hireDate, recruiter)

**Resume Database**
Interviews(ID, date, location, recruiter)
CVs(candID, resume)

**Credit Card Database**
Cards(CustID, cardNum, expiration, currentBalance)
Customers(CustID, name, address)

**HelpLine Database**
Calls(date, agent, custID, description, followUp)

7

## Website

- http://www.monster.com/
- http://www.careerbuilder.com/
- =>vào một trang web tích hợp dữ liệu từ các trang tương ứng trên web

8

## Các ngành khác

- Sinh học, sinh thái học, quản lý nguồn nước
- Các nhà khoa học thu thập dữ liệu một cách độc lập và muốn cộng tác với nhau.

9

## 2. Mục tiêu của tích hợp dữ liệu

- Cung cấp truy cập đồng bộ tới một tập các nguồn dữ liệu tự trị và không đồng nhất
  - Truy vấn: truy vấn trên các nguồn dữ liệu riêng biệt
  - Số lượng nguồn dữ liệu: số lượng nguồn dữ liệu tăng lên? THDL Web-scale?
  - Tính không đồng nhất: các nguồn dữ liệu được phát triển độc lập, trên những hệ thống khác nhau: CSDL, hệ quản trị nội dung, file trong thư mục. Một số nguồn có cấu trúc, một số phi cấu trúc hoặc bán cấu trúc
  - Tự trị: các nguồn dữ liệu không nhất thiết thuộc về cùng một thực thể quản trị, mà có thể thuộc về các tổ chức con khác nhau.

10

## 3. Tại sao THDL là vấn đề khó?

- Lý do hệ thống
  - khác nền, khác chuẩn
  - CSDL phân tán
  - khả năng xử lý truy vấn trên các nguồn dữ liệu
- Lý do logic
  - dữ liệu được tổ chức logic trong các nguồn dữ liệu, thông qua lược đồ. Các lược đồ thường khác nhau
  - dữ liệu ở các nguồn khác nhau cũng được biểu diễn khác nhau

11

## 3. Tại sao THDL là vấn đề khó?

- Lý do xã hội và quản trị
  - dữ liệu có được lưu trữ trên thiết bị điện tử?
  - có dễ dàng tiếp cận với các nguồn dữ liệu?
  - việc cho phép hệ thống tích hợp dữ liệu truy cập và sử dụng nguồn dữ liệu của tổ chức có thể thêm tải cho hệ thống của tổ chức.
  - các vấn đề an ninh, bảo mật

12

## 4. Các kỳ vọng

- Xây dựng công cụ làm giảm công sức khi tích hợp các nguồn dữ liệu.
- Cải thiện khả năng trả lời các câu hỏi trong môi trường không chắc chắn của hệ thống.

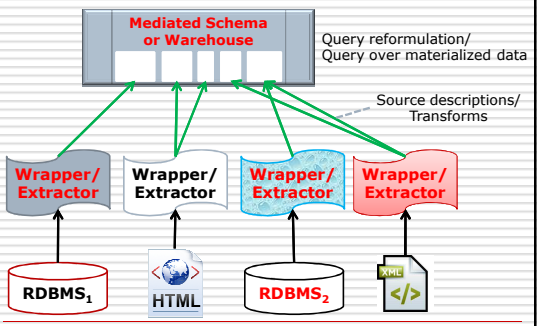
13

## 5. Kiến trúc tích hợp dữ liệu

- Có 2 kiến trúc
  - warehousing
    - dữ liệu từ các nguồn riêng biệt được nạp vào một CSDL vật lý (gọi là warehouse – kho dữ liệu), và trả lời truy vấn được thực hiện trên kho dữ liệu này.
  - virtual integration
    - dữ liệu vẫn nằm ở các nguồn, và được truy cập khi cần thiết lúc xử lý truy vấn.

14

## Kiến trúc tích hợp dữ liệu




15

## Các thành phần của hệ THDL

- Hệ tích hợp ảo
  - Nguồn dữ liệu
  - Wrapper
    - là chương trình làm nhiệm vụ: gửi các truy vấn tới nguồn dữ liệu, nhận câu trả lời và có thể áp dụng một số biến đổi cơ bản trên câu trả lời.
  - Mediated schema
    - chỉ chứa những gì liên quan đến miền ứng dụng, không nhất thiết chứa tất cả các thuộc tính của các nguồn

16

## Các thành phần của hệ THDL

- Source descriptions
  - cầu nối giữa mediated schema và lược đồ của nguồn
  - xác định các thuộc tính của nguồn mà hệ thống cần biết để dùng dữ liệu của chúng
  - thành phần chính là ánh xạ ngữ nghĩa
  - Ánh xạ ngữ nghĩa:
    - xác định cách các thuộc tính của nguồn tương ứng với các thuộc tính của mediated schema.
    - hợp giải các thuộc tính khác nhau ở các nguồn
    - xác định cách hợp giải các giá trị dữ liệu khác nhau ở các nguồn

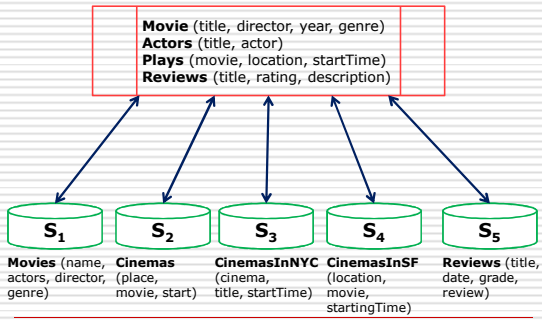
17

## Các thành phần của hệ THDL

- Hệ warehousing
  - Người dùng đưa câu truy vấn vào lược đồ warehouse
  - Đây là lược đồ vật lý, có dữ liệu thể hiện.
  - Hệ thống có tool ETL (Extract-Transform-Load) định kỳ trích rút dữ liệu từ các nguồn và nạp nó vào warehouse.
  - ETL áp dụng nhiều phép biến đổi dữ liệu phức tạp hơn Wrapper nhiều: làm sạch, tổng hợp và biến đổi giá trị.

18

## 6. Ví dụ hệ THDL




19

## Ví dụ truy vấn

- Giả sử người dùng đặt câu truy vấn tìm thời gian chiếu bộ phim ở New York được đạo diễn bởi Woody Allen

```
SELECT title, startTime
FROM Movie, Plays
WHERE Movie.title = Plays.movie AND
location = 'New York' AND director = 'Woody
Allen'
```

20

## 7. Các bước xử lý truy vấn

- Phân hệ viết lại truy vấn (Query reformulation)
  - Viết lại truy vấn này thành các truy vấn tham chiếu tới các lược đồ của các nguồn dữ liệu. Kết hợp các truy vấn này sẽ cho câu trả lời cho truy vấn ban đầu.
  - Cần sử dụng các mô tả nguồn (source descriptions)
  - Kết quả của phân hệ viết lại truy vấn là một kế hoạch truy vấn logic (logical query plan)

21

- Trong ví dụ này:
  - Các bộ Movie có thể thu được từ nguồn S1, nhưng thuộc tính 'title' cần viết lại thành 'name'
  - Các bộ Plays có thể thu được từ S2 và S3. S3 chứa đầy đủ dữ liệu về các show ở New York nên ta chọn S3.
  - Nguồn S3 cần title làm tham số đầu vào, nhưng chưa có title tường minh trong câu hỏi ban đầu, do đó, trước tiên, query plan phải truy cập nguồn S1 trước, sau đó trích rút ra thông tin title làm đầu vào cho S3.

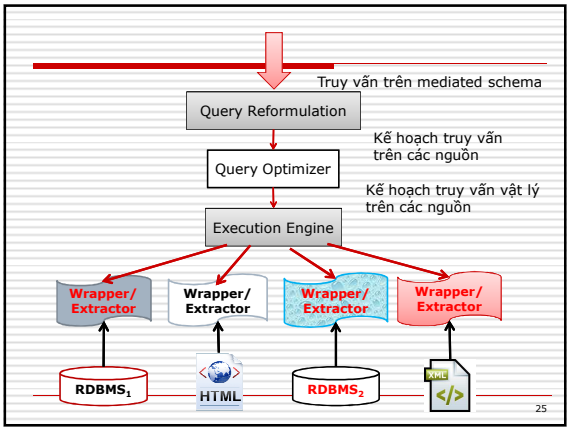
22

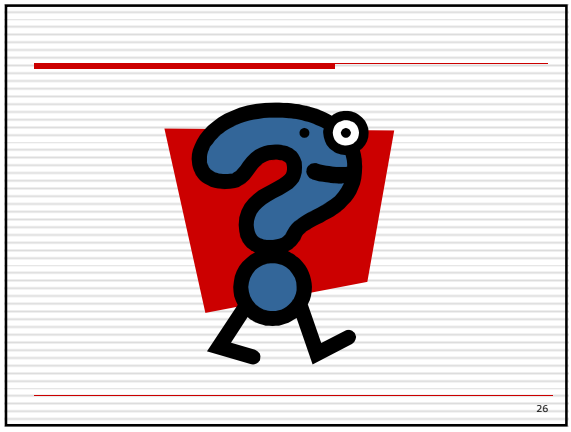
- Phân hệ tối ưu hóa truy vấn (query optimization)
  - Đầu vào là kế hoạch truy vấn logic
  - Đầu ra là kế hoạch truy vấn vật lý (physical query plan), xác định chính xác trình tự truy cập các nguồn; khi các kết quả được kết hợp, thuật giải nào được sử dụng để thực hiện các thao tác trên dữ liệu (kết nối giữa các nguồn) và lượng tài nguyên phân phối cho mỗi thao tác.
  - Hệ thống cũng phải kiểm soát các nguy cơ bắt nguồn từ tính phân tán của hệ THDL

23

- Phân hệ thực thi truy vấn (query execution)
  - Mô tơ thực thi chịu trách nhiệm cho việc thực hiện kế hoạch truy vấn vật lý
  - Mô tơ thực thi tách các truy vấn vào các nguồn dữ liệu cụ thể thông qua wrapper và tổng hợp kết quả theo kế hoạch truy vấn.
  - Mô tơ thực thi cũng có thể yêu cầu bộ tối ưu xem xét lại kế hoạch của nó.

24




## Lời hay ý đẹp

"Sở dĩ người ta đau khổ chính vì mãi đeo đuổi những thứ sai lầm"

*Phật giáo*

27