Data integration
An overview

NGUYEN Hong-Phuong
Email: phuongnh@soict.hust.edu.vn
Site: http://is.hust.edu.vn/~phuongnh
Department of Information Systems
School of Information and Communication Technology
Hanoi University of Science and Technology

1

Contents

2

1. The need of data integration

- FullServe
- Website
- Other fields

3

FullServe

- provide internet access services, computing infrastructure such as modem, wireless router, voice-over-IP phone
- FullServe expanse its market to Europe, merge with EuroCard, which specializes in providing credit cards.
- FullServe: hundreds of tables
- EuroCard: dozens of tables

4

FullServe

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)
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)

5

FullServe

- EuroCard DB

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)

6

Website

- http://www.monster.com/
- http://www.careerbuilder.com/

7

Other fields

- Biology, ecology, water resources management
- Scientists collect data independently and they want to collaborate with others.

8

2. Goal of data integration system

- Provide an uniform access to a set of heterogeneous and autonomous data sources.

9

3. Why is it a hard problem?

- System reasons
  - platform, standard
  - distributed database
  - ability to process queries on the data sources
- Logic reasons
  - Data are organized in data source via schema. Schemes are often different
  - Data in different sources are also represented in different ways

10

3. Why is it a hard problem?

- Social and management reasons
  - Are data stored on electronic devices?
  - Is it easy to access to the data sources?
  - data integration system access and use data of organizations can add load to the system of the organizations.
  - security issues

11

4. Expectations

- Construct tools reduce the effort when integrating data sources.
- Improve the ability to answer questions in the uncertain environment of the system.

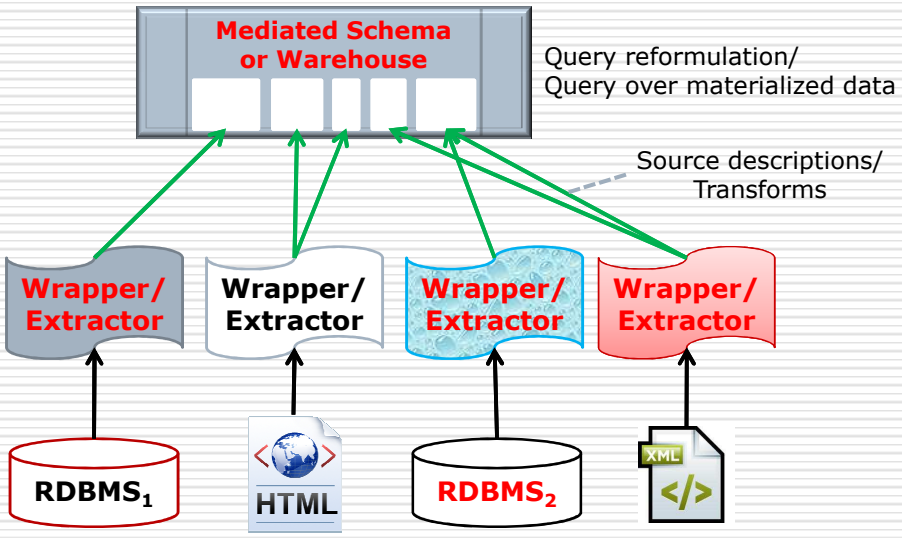
12

5. Architecture of Data Integration systems

- warehousing
  - data from separate sources must be loaded into a physical database (called the warehouse - data warehousing), and query answering is performed on this database.
- virtual integration
  - data remains in the sources, and is accessed when needed at query processing time.

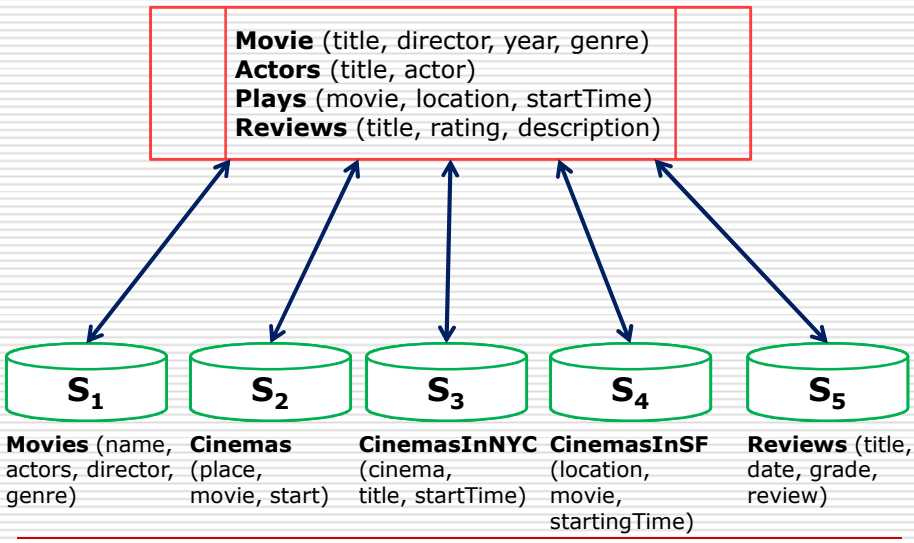
13

Architecture



14

6. An example



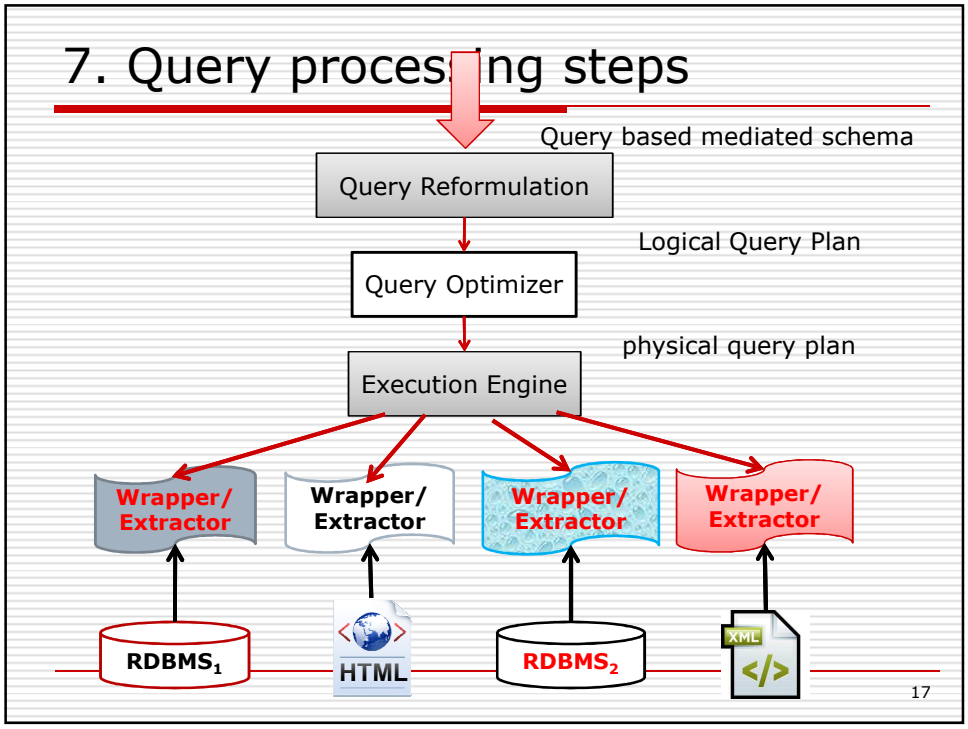
15

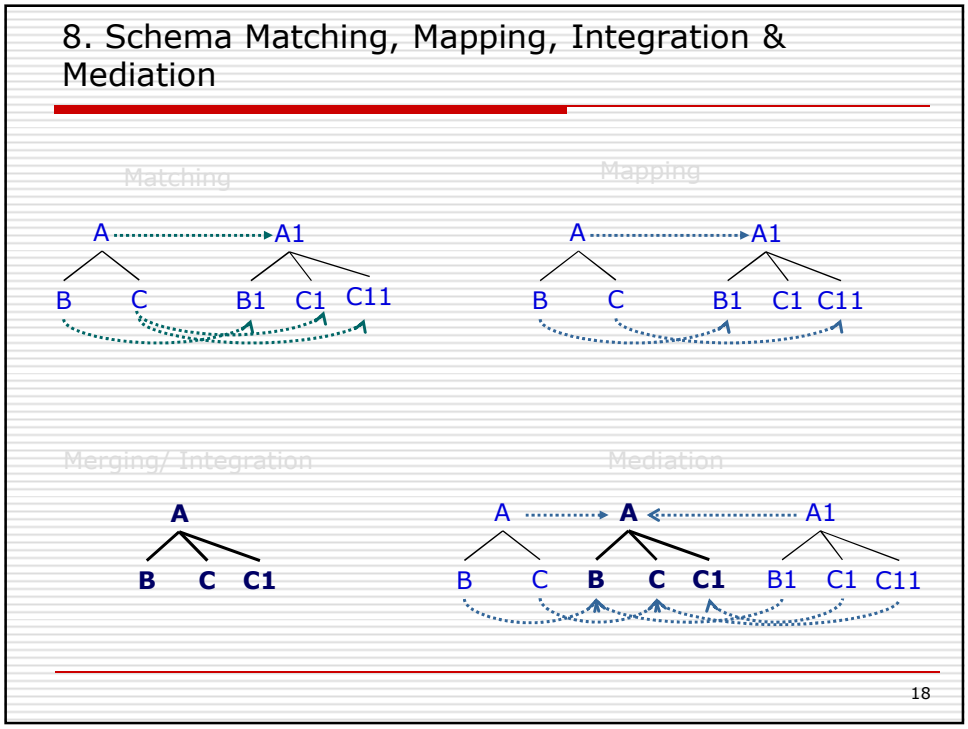
Query

- Suppose that an user want to find the time for a film in New York directed by Woody Allen

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

16


7. Query processing steps
Query based mediated schema
Query Reformulation
Logical Query Plan
Query Optimizer
physical query plan
Execution Engine
Wrapper/ Extractor
Wrapper/ Extractor
Wrapper/ Extractor
Wrapper/ Extractor
RDBMS1
HTML
RDBMS2
XML
17


8. Schema Matching, Mapping, Integration & Mediation
Matching
A
A1
B
C
B1
C1
C11
Mapping
A
A1
B
C
B1
C1
C11
Merging/ Integration
A
B
C
C1
Mediation
A
A
A1
B
C
B
C
C1
B1
C1
C11
18

Future directions

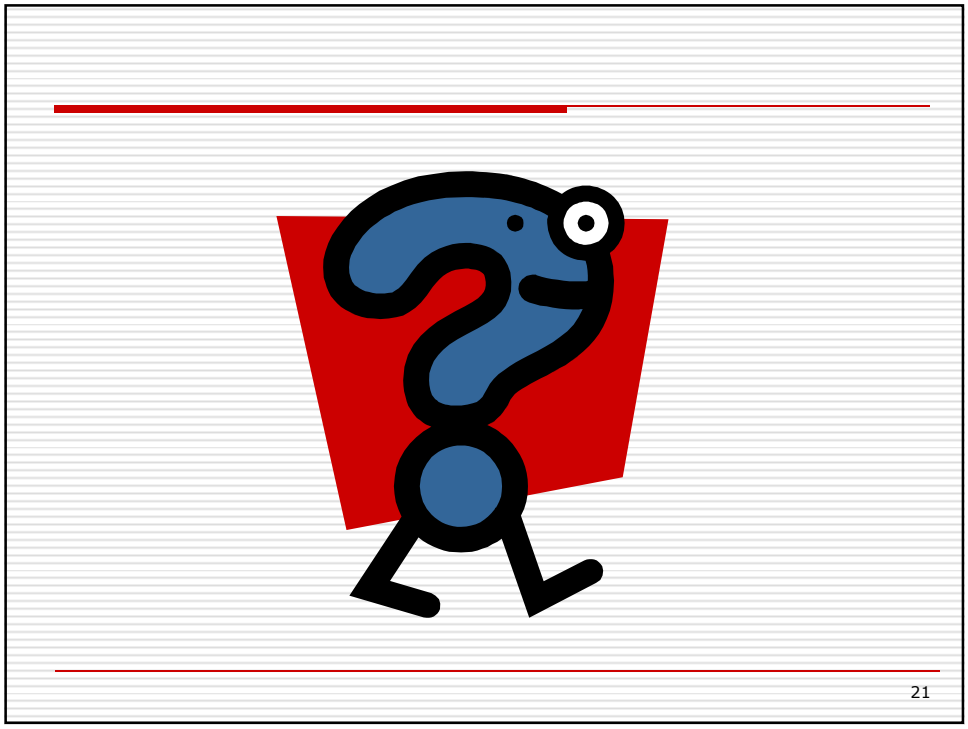
- Large scale schema matching and integration as web service.
- Merging web services at large scale
- Large scale schema matching and integration in schema based P2P database systems.
- Application of matching research in domain specific large scale social network environments.
- Large scale schema matching and integration using parallel computing techniques.

19

References

- AnHai Doan, Alon Halevy, Zachary Ives, "Principles of Data Integration", MK-Elsevier, 2012. ISBN: 978-0-12-416044-6
- Do Hong Hai, "Schema Matching and Mapping-based Data Integration, Architecture, Approaches and Evaluation", VDM Verlag Dr. Muller, 2006. ISBN-10: 3-86550-997-5
- Zohra Bellahsene, Angela Bonifati, Erhard Rahm, "Schema Matching and Mapping", Springer, 2011. ISBN: 978-3-642-16517-7
- Avigdor Gal, "Uncertain Schema Matching", Synthesis Lectures on Data Management, Morgan & Claypool, 2011. ISBN: 9781608454334

20



21