

Data Integration

Vu Tuyền Trinh

1

Outline

➤ Introduction

- Examples of data integration applications
- Schema heterogeneity
- Goal of data integration
- Data integration architectures
- Review of basic database concepts

2

Data Integration

- Databases are great ?
 - Assuming you've put it all into your schema.
 - Data sets are often created independently
 - Only to discover later that they need to combine their data!
 - Data in different systems, different schemata and limited interfaces to data.
- Data integration: tie together different sources, controlled by many people, under a common schema.

3

DBMS: it's all about abstraction

- *Logical vs. Physical; What vs. How.*

Students:

SSN	Name	Category
123-45-6789	Charles	undergrad
234-56-7890	Dan	grad

Takes:

SSN	CID
123-45-6789	CSE444
123-45-6789	CSE444
234-56-7890	CSE142
	...

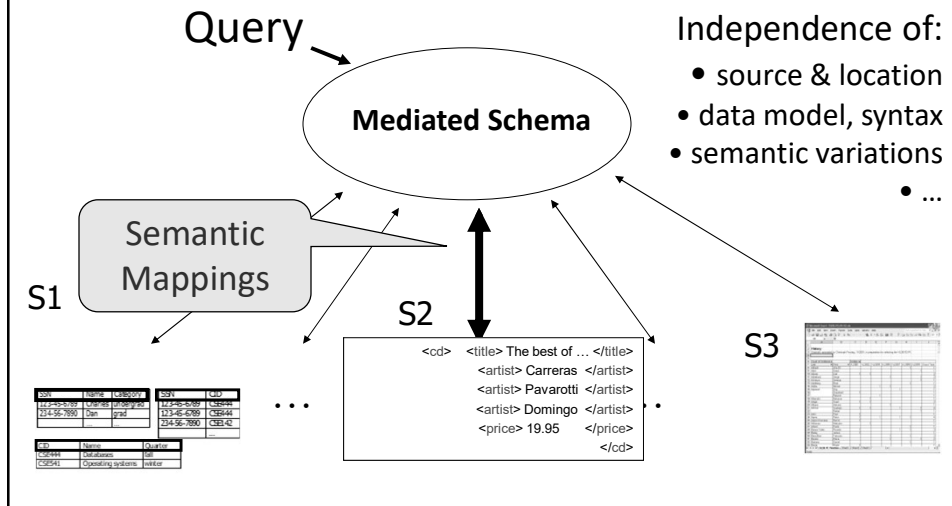
Courses:

CID	Name	Quarter
CSE444	Databases	fall
CSE541	Operating systems	winter

```
SELECT C.name
FROM Students S, Takes T, Courses C
WHERE S.name="Mary" and
      S.ssn = T.ssn and T.cid = C.cid
```

4

Data Integration: A Higher-level Abstraction



5

Outline

- ✓ Introduction: data integration as a new abstraction
- Examples of data integration applications
 - Schema heterogeneity
 - Goal of data integration, why it's a hard problem
 - Data integration architectures
 - Review of basic database concepts

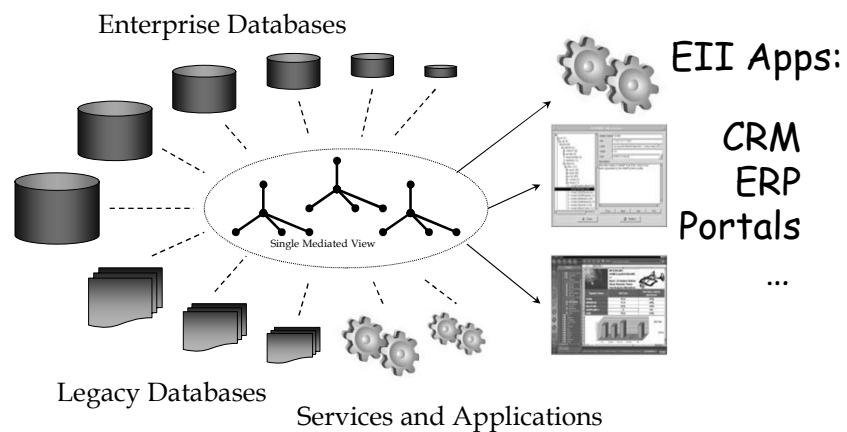
6

Applications of Data Integration

- Business
- Science
- Government
- The Web
- Pretty much everywhere

7

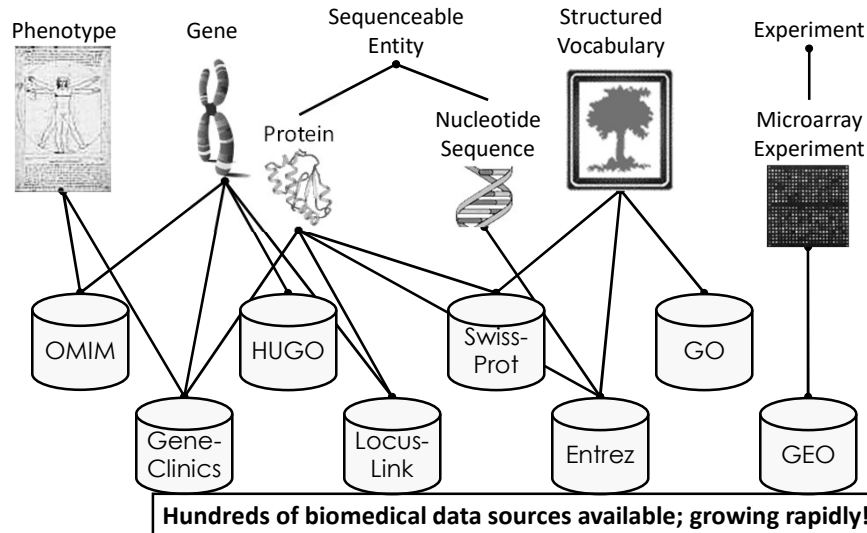
Application Area 1: Business



50% of all IT \$\$\$ spent here!

8

Application Area 2: Science



9

Application Area 3: The Web



10

The Presidents of the USA - EnchantedLearning.com - Mozilla Firefox

http://www.enchantedlearning.com/history/us/pres/list.shtml

As a thank-you bonus, site members have access to a banner-ad-free version of the site, with print-friendly pages.
(Already a member? [Click here.](#))

EnchantedLearning.com

US History

US Flags US Geography

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

African-Americans Artists Explorers of the US Inventors US Presidents US Symbols US States

EnchantedLearning.com

The Presidents of the United States of America

[In the order in which they served](#) [Alphabetical order](#) [Short table of Data](#)

President's Day Activities Abraham Lincoln

The President and Vice-President are elected every four years. They must be at least 35 years of age, they must be native-born citizens of the United States, and they must have been residents of the U.S. for at least 14 years. (Also, a person cannot be elected to a third term as President.)

President	Party	Term as President	Vice-President
1. George Washington (1732-1799)	None, Federalist	1789-1797	John Adams
2. John Adams (1735-1826)	Federalist	1797-1801	Thomas Jefferson
3. Thomas Jefferson (1743-1826)	Democratic-Republican	1801-1809	Aaron Burr, George Clinton
4. James Madison (1751-1836)	Democratic-Republican	1809-1817	George Clinton, Elbridge Gerry
5. James Monroe (1758-1831)	Democratic-Republican	1817-1825	Daniel Tompkins
6. John Quincy Adams (1767-1848)	Democratic-Republican	1825-1829	John Calhoun
7. Andrew Jackson (1767-1845)	Democrat	1829-1837	John Calhoun, Martin van Buren
8. Martin van Buren (1781-1862)	Democrat	1837-1841	Richard Mentor Johnson
9. William Henry Harrison (1773-1841)	Whig	1841	Richard Mentor Johnson
10. John Tyler (1790-1862)	Whig	1841-1845	Richard Mentor Johnson
11. James K. Polk (1795-1846)	Democrat	1845-1849	George M. Dallas
12. Zachary Taylor (1784-1850)	Whig	1849-1850	Millard Fillmore
13. Millard Fillmore (1811-1874)	Whig	1850-1853	William A. R. King
14. Franklin Pierce (1804-1869)	Democrat	1853-1857	William A. R. King
15. James Buchanan (1791-1868)	Democrat	1857-1861	John Breckinridge

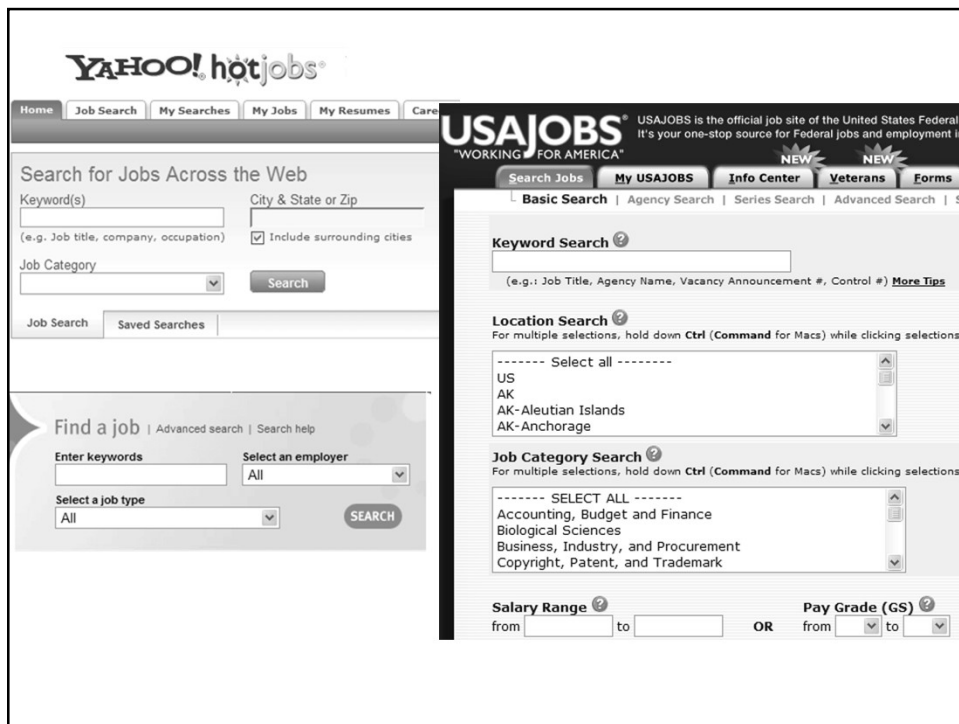
Hundreds of millions of high-quality tables on the Web

11

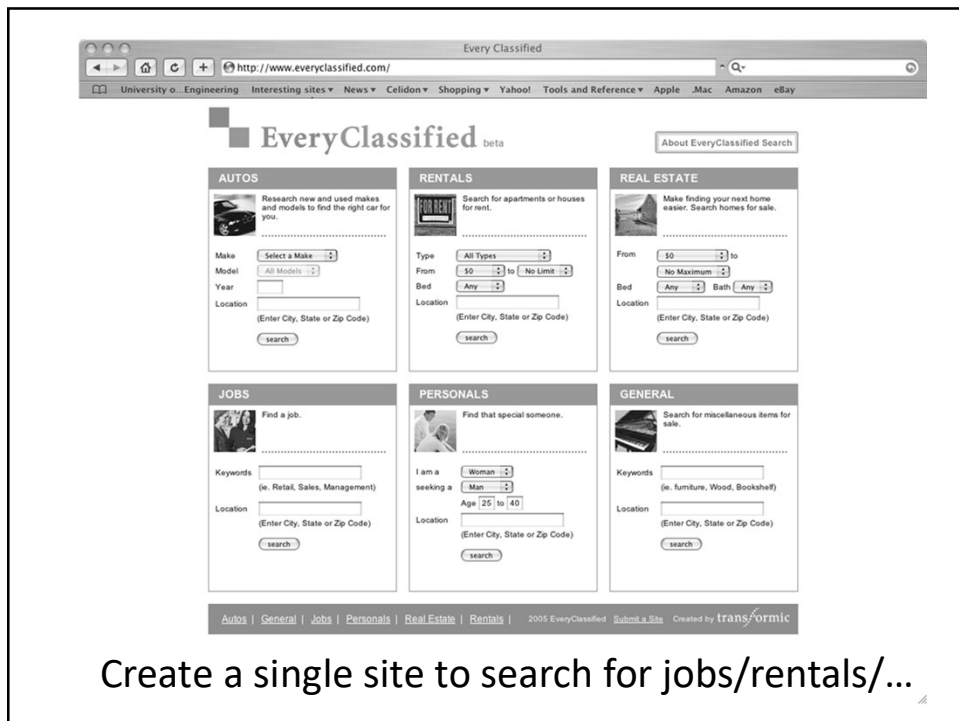
The Deep Web

- Millions of high quality HTML forms out there
- Each form has its own special interface
 - Hard to explore data across sites.
- Goal (for some domains):
 - A single interface into a multitude of deep-web sources.

12



13



Create a single site to search for jobs/rentals/...

14

EveryClassified

Home | Autos | General | Jobs | Personals | Real Estate | Rentals

AUTOS

Yahoo! Autos
Cars.com
MercuryNews
ContraCosta Times
Monterey County Herald
Modesto Bee
Craiglist - Bay Area
CarsDirect
Autobytel.com
Motorway
InsideBayArea
Valley Classifieds
SF Gate
Main Independent Journal
SF Weekly
Santa Cruz Sentinel
Automotive Search
East Bay Express
Palo Alto Online
Recordnet.com
 backpage.com

More Classifieds>

About This Search

MercuryNews.com: News | Business | Sports | Entertainment | Living | Travel | Shopping | Classifieds | Jobs | Cars | Homes

AutoWest Acura of Stevens Creek

cars.com Home Buy Sell Research Shopping Advice The Mercury News

Results: 1-36 | [Revise Search](#) [Print](#) [See Saved Vehicles](#)

Year	Vehicle	Price ↓	Mileage	Photo	Seller	Body	Color	Distance	Save
2001	Acura Integra LS	\$18,888	40,547		Mike Harvey Honda	Sedan	Red	17 mi.	<input type="checkbox"/>
2001	Acura Integra GS	\$17,725	35,962		Mike Harvey Acura	Hatch	Green	17 mi.	<input type="checkbox"/>
2001	Acura Integra LS	\$15,865	33,409		Mike Harvey Acura	Hatch	Silver	17 mi.	<input type="checkbox"/>
2001	Acura Integra LS	\$15,505	41,115		Mike Harvey Acura	Hatch		17 mi.	<input type="checkbox"/>
2001	Acura Integra LS	\$14,600	31,000		Stevens Creek Toyota	Hatch	Silver	12 mi.	<input type="checkbox"/>
2000	Acura Integra LS	\$14,335	59,868		Mike Harvey Acura	Hatch	Black	17 mi.	<input type="checkbox"/>
2001	Acura Integra LS	\$12,875	46,672		Burlingame European	Hatch	Silver	16 mi.	<input type="checkbox"/>
1999	Acura Integra GS-R	\$12,500	79,688		Individual Seller	Coupe	Black	18 mi.	<input type="checkbox"/>
2000	Acura Integra LS	\$11,999	35,000		Carlsen Subaru	Sedan		8 mi.	<input type="checkbox"/>
2000	Acura Integra LS	\$11,988	60,871		Putnam Toyota	Hatch	Black	16 mi.	<input type="checkbox"/>
2000	Acura Integra LS	\$11,888	--		Classified Ad	Sedan		15 mi.	<input type="checkbox"/>
2000	Acura Integra	\$10,999	83		Dealer	Hatch	Silver	12 mi.	<input type="checkbox"/>
1999	Acura Integra LS	\$9,999	78,000		Dealer	Hatch	Silver	12 mi.	<input type="checkbox"/>

AutoWest Acura of Stevens Creek
#1 Volume Acura Store in Northern California for 2004

Easily traverse between the site by clicking its name

15

Phòng chống dịch nCoV - BAOMOI

baomoi.com/phong-chong-dich-ncov/top/328.epi

Apps Google Drive Blockchain DBLP Deadline Solid Blockchain technol... Other Bookmarks

BAOMOI
trang thông tin điện tử

NHẬP NỘI DUNG TÌM KIẾM

NÓNG MỚI VIDEO CHỦ ĐỀ # Năng lượng tích cực # Khám phá Việt Nam # Phòng chống dịch nCoV # Xây dựng Đảng

PHÒNG CHỐNG DỊCH NCOV

Yêu cầu khai báo trung thực về việc tiếp xúc với người trên chuyến bay VN0054
Tin Tức TTXVN 2 giờ 48 liên quan

Người dân cần bình tĩnh, không hoang mang
Tổ Quốc 2 giờ 6634 liên quan

VIDEO

V NEWS

Đảm bảo an toàn cho các c...
tham dự AEM
VNEWS 27 phút 40 liên quan

Baomoi.com

16

Outline

- ✓ Introduction: data integration as a new abstraction
- ✓ Examples of data integration applications
 - Schema heterogeneity
 - Goal of data integration, why it's a hard problem
 - Data integration architectures
 - Review of basic database concepts

17

Enterprise Data Integration:

FullServe Corporation

Employees

FullTimeEmp
Hire
TempEmployees

Training

Courses
Enrollments

Sales

Products
Sales

Resumes

Interview
CV

Services

Services
Customers
Contracts

HelpLine

Calls

18

EuroCard Corporation

Employees

Employees
Hire

Resumes

Interview

Credit Cards

Customer
CustDetail

HelpLine

Calls

19

Examples of Heterogeneity

FullServe

FullTimeEmp

ssn, empId, firstName
middleName, lastName

Hire

empId, hireDate, recruiter

TempEmployees

ssn, hireStart, hireEnd

EuroCard

Employees

ID, firstNameMiddleInitial,
lastName

Hire

ID, hireDate, recruiter

Find all employees (making over \$100K)

20

Customer Call Center

Agents should have a full view of customer when they call in.

Sales

Products
Sales

Credit Cards

Customer
CustDetail

Services

Services
Customers
Contracts

21

Other Reasons to Integrate Data

- Create a (useful) web site for tracking services
- Collaborate with third parties
 - E.g., create branded services
- Comply with government regulations
 - Find “risky” employees
- Business intelligence
 - What’s really wrong with our products?

22

Outline

- ✓ Introduction: data integration as a new abstraction
- ✓ Examples of data integration applications
- ✓ Schema heterogeneity
- Goal of data integration, why it's a hard problem
 - Data integration architectures
 - Review of basic database concepts

23

Goal of Data Integration

- Uniform query access to a set of data sources
- Handle:
 - Scale of sources: from tens to millions
 - Heterogeneity
 - Semi-structure

24

Why is it Hard?

- Systems-level reasons:
 - Managing different platforms
 - SQL across multiple systems is not so simple
 - Distributed query processing
- Logical reasons:
 - Schema (and data) heterogeneity
- ‘Social’ reasons:
 - Locating and capturing relevant data in the enterprise.
 - Convincing people to share (data fiefdoms)
 - Security, privacy and performance implications.

25

Data Integration Smorgasbord

Something for everyone:

- **Theory** of modeling data sources
- **Systems** aspects of data integration
- **Architectural** issues: e.g., P2P data sharing
- **AI @ work**: automated schema matching
- **Web**: latest on data integration & web
- **Commercial** products: BEA, IBM
- **Semantic Web**: what does it have to offer?
- New trends in DBMS: **uncertainty, dataspace**s

26

Outline

- ✓ Introduction: data integration as a new abstraction
- ✓ Examples of data integration applications
- ✓ Schema heterogeneity
- ✓ Goal of data integration, why it's a hard problem
- Data integration architectures
- Review of basic database concepts

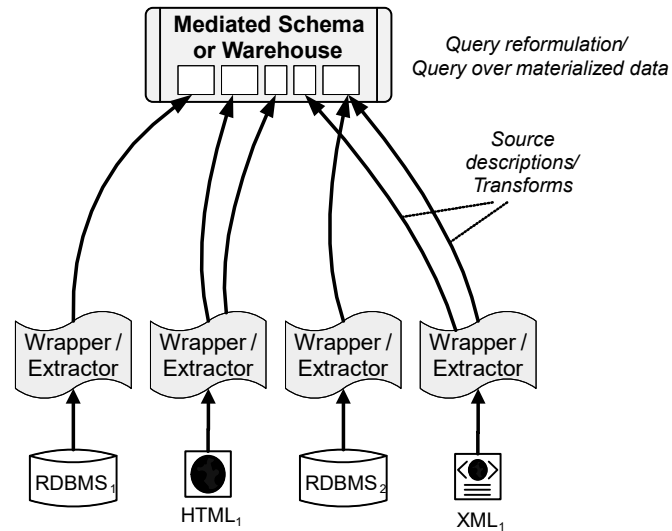
27

Virtual, Warehousing and in Between

- Virtual data integration: leave the data at the sources and access it at query time.
- Data warehousing: integrate by bringing the data into a single physical warehouse
- ❖ semantic heterogeneity
- ❖ Numerous intermediate architectures.

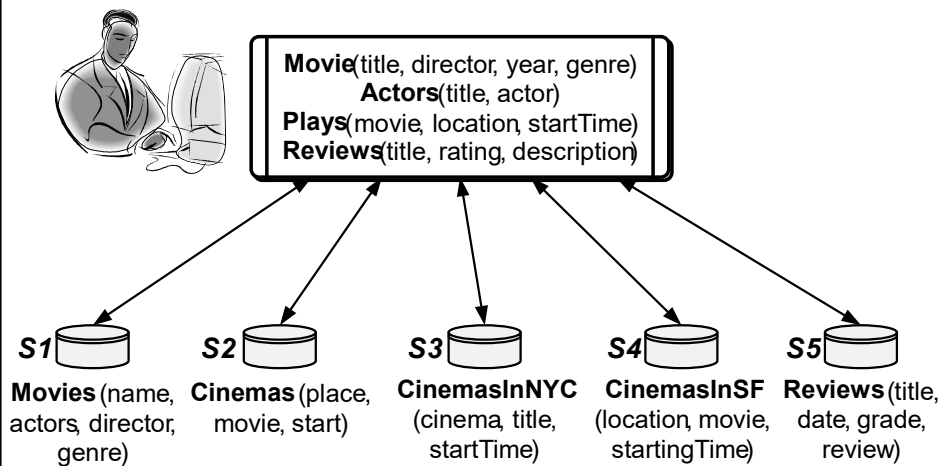
28

Virtual Data Integration Architecture











29

Example



30

Wrappers

2.		The Best of the Three Tenors (Audio CD) ~ by Luciano Pavarotti, Placido Domingo, Jose Carreras Avg. Customer Rating:  (View recommendations)
Usually ships in 24 hours		
List Price: \$18.98		Used & new from \$8.95
Buy new: \$14.99		
3.		The Three Tenors In Concert 1994 (Audio CD) ~ by Jules Massenet, Federico Moreno Torroba, Richard Rodgers Avg. Customer Rating:  (View recommendations)
Usually ships in 24 hours		
List Price: \$11.98		Used & new from \$1.79
Buy new: \$10.99		Club price: \$8.49
4.		Trombonastics (Audio CD) ~ by Joseph Alessi Avg. Customer Rating:  (Rate this item)
Usually ships in 24 hours		
List Price: \$18.98		Used & new from \$14.23
Buy new: \$14.99		
5.		The Three Tenors Christmas (Audio CD) ~ by Carreras, Domingo, Pavarotti Avg. Customer Rating:  (View recommendations)
Usually ships in 3 to 4 days		
List Price: \$13.98		Used & new from \$1.89
Buy new: \$13.98		

```
<cd>  <title> The best of ... </title>
      <artist> Abiteboul </artist>
      <artist> Pavarotti </artist>
      <artist> Domingo </artist>
      <price> 19.95  </price>

</cd>
...
```

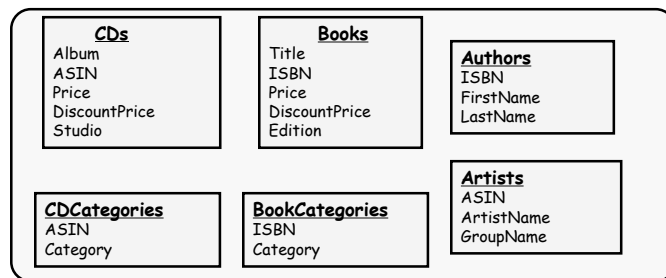
Send queries to data sources
and transform answers into
tuples (or other internal data
model).

31

Mediation Languages

Describe
relationships
between mediated
schema and data
sources

Mediated Schema
CD: ASIN, Title, Genre,...
Artist: ASIN, name, ...



32

Woody Allen Comedies in NY

Mediated schema:

Movie: Title, director, year, genre
Actors: title, actor
Plays: movie, location, startTime
Reviews: title, rating, description

```
select title, startTime
from Movie, Plays
where Movie.title=Plays.movie AND
      location="New York" AND
      director="Woody Allen"
```

33

Movie: Title, director, year, genre
Actors: title, actor
Plays: movie, location, startTime
Reviews: title, rating, description

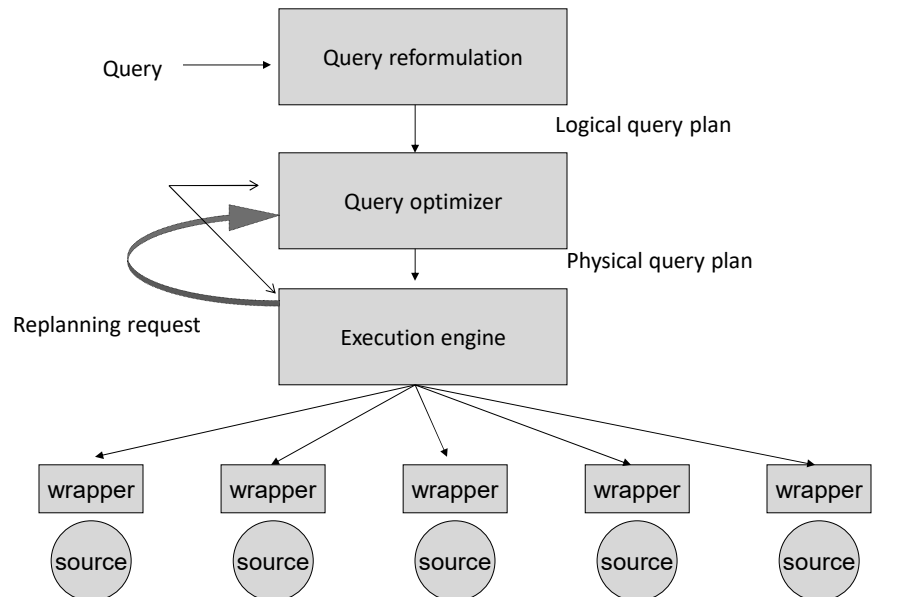
```
select title, startTime
from Movie, Plays
where Movie.title=Plays.movie AND
      location="New York" AND
      director="Woody Allen"
```

Sources S1 and S3 are relevant, sources S4 and S5 are irrelevant, and source S2 is relevant but possibly redundant.

S1	S2	S3	S4	S5
Movies: name, actors, director, genre	Cinemas: place, movie, start	Cinemas in NYC: cinema, title, startTime	Cinemas in SF: location, movie, startingTime	Reviews: title, date grade, review

34

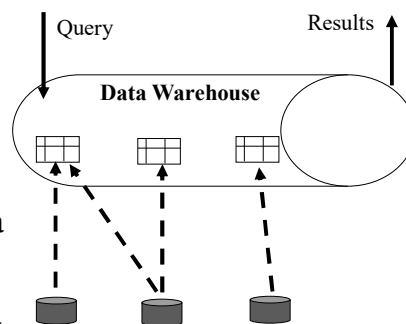
Query Processing



35

Data Warehouses – Offline Replication

- Determine physical schema
- Define a database with this schema
- Define procedural *mappings* in an “ETL tool” to import the data and clean it.
- Periodically copy all of the data from the data sources
 - Note that the sources and the warehouse are basically independent at this point



36

36

Pros and Cons of Data Warehouses

- ✖ Need to spend time to design the physical database layout, as well as logical
 - ✖ This actually takes a lot of effort!
- ✖ Data is generally not up-to-date (lazy or offline refresh)
- ✓ Queries over the warehouse don't disrupt the data sources
- ✓ Can run very heavy-duty computations, including data mining and cleaning

37

37

Outline

- ✓ Introduction: data integration as a new abstraction
- ✓ Examples of data integration applications
- ✓ Schema heterogeneity
- ✓ Goal of data integration, why it's a hard problem
- ✓ Data integration architectures
- Review of basic database concepts

38

Basic Database Concepts

- Relational data model
- Integrity constraints
- Queries and answers
- Conjunctive queries
- Datalog

39

Relational Terminology

Relational schemas

- Tables, attributes

Relation instances

- Sets (or multi-sets) of tuples

Integrity constraints

- Keys, foreign keys, inclusion dependencies

40

Product

PName	Price	Category	Manufacturer
Gizmo	\$19.99	Gadgets	GizmoWorks
Powergizmo	\$29.99	Gadgets	GizmoWorks
SingleTouch	\$149.99	Photography	Canon
MultiTouch	\$203.99	Household	Hitachi

41

SQL (very basic)

Interview:

candidate, date, recruiter, hireDecision, grade

EmployeePerf:

empID, name, reviewQuarter, grade, reviewer

```
select recruiter, candidate
from Interview, EmployeePerf
where recruiter=name AND
      grade < 2.5
```

42

Query Answers

- $Q(D)$: the set (or multi-set) of rows resulting from applying the query Q on the database D .
- Unless otherwise stated, we will consider sets rather than multi-sets.

43

SQL (w/aggregation)

EmployeePerf:

emplID, name, reviewQuarter, grade, reviewer

```
select reviewer, Avg(grade)
from EmployeePerf
where reviewQuarter="1/2007"
```

44

Integrity Constraints (Keys)

- A key is a set of columns that uniquely determine a row in the database:
 - There do not exist two tuples, t_1 and t_2 such that $t_1 \neq t_2$ and t_1 and t_2 have the same values for the key columns.
 - (EmpID, reviewQuarter) is a key for **EmployeePerf**

45

Integrity Constraints (Functional Dependencies)

- A set of attribute **A** functionally determines a set of attributes **B** if: whenever, t_1 and t_2 agree on the values of **A**, they must also agree on the values of **B**.
- For example, (EmpID, reviewQuarter) functionally determine (grade).
- Note: a key dependency is a functional dependency where the key determines all the other columns.

46

Integrity Constraints (Foreign Keys)

- Given table **T** with key B and table **S** with key A : A is a foreign key of B in **T** if whenever a **S** has a row where the value of A is v , then **T** must have a row where the value of B is v .
- Example: the empID attribute of **EmployeePerf** is a foreign key for attribute emp of **Employee**.

47

General Integrity Constraints

Tuple generating dependencies (TGD's)

$$(\forall \bar{X}) s_1(\bar{X}_1), \dots, s_m(\bar{X}_m) \rightarrow (\exists \bar{Y}) t_1(\bar{Y}_1), \dots, t_l(\bar{Y}_l)$$

Equality generating dependencies (EGD's): right hand side contains only equalities.

$$(\forall \bar{X}) s_1(\bar{X}_1), \dots, s_m(\bar{X}_m) \rightarrow Y_1^1 = Y_2^1, \dots, Y_1^k = Y_2^k$$

Exercise: express the previous constraints using general integrity constraints.

48

Conjunctive Queries

Q(X,T) :-

Interview(X,D,Y,H,F), EmployeePerf(E,Y,T,W,Z),
W < 2.5.

Joins are expressed with multiple
occurrences of the same variable

```
select recruiter, candidate
from Interview, EmployeePerf
where recruiter=name AND
      grade < 2.5
```

49

Conjunctive Queries (interpreted predicates)

Q(X,T) :-

Interview(X,D,Y,H,F), EmployeePerf(E,Y,T,W,Z),
W < 2.5.

Interpreted (or comparison) predicates.

Variables must also appear in regular atoms.

```
select recruiter, candidate
from Interview, EmployeePerf
where recruiter=name AND
      grade < 2.5
```

50

Conjunctive Queries (negated subgoals)

Q(X,T) :-

Interview(X,D,Y,H,F), EmployeePerf(E,Y,T,W,Z),
¬OfferMade(X, date).

Safety: every head variable must appear in a positive subgoal.

51

Unions of Conjunctive Queries

Multiple rules with the same head predicate express a union

Q(R,C) :-

Interview(X,D,Y,H,F), EmployeePerf(E,Y,T,W,Z),
W < 2.5.

Q(R,C) :-

Interview(X,D,Y,H,F), EmployeePerf(E,Y,T,W,Z),
Manager(y), W > 3.9.

52

Summary

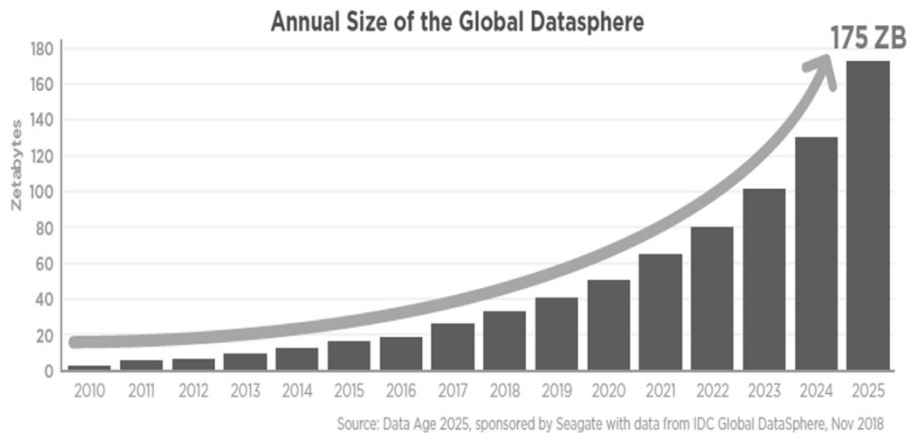
- Data integration: abstract away the fact that data comes from multiple sources in varying schemata.
- Problem occurs everywhere: it's key to business, science, Web and government.
- Goal: reduce the effort involved in integrating.
- Regardless of the architecture, heterogeneity is a key issue.
- Architectures range from warehousing to virtual integration.

53

Big Data Integration

54

How big is big data?



55

55

Big Data Growth Statistics

- An internet user generates ~ 1.7 megabytes (MB) of data / second.
- 2022: ~ 97 zettabytes - the estimated volume of data created worldwide
- 2023:
 - ~ 2/3 world population be online
 - internet users generate nearly 3 times the volume of data generated in 2019.
- 2025:
 - people will create more than 181 ZB of data. That's 181, followed by 21 zeros.
 - there will be 55.7 billion connected IoT devices. These IoT devices alone will generate almost 80 ZB by 2025.
- An internet user would need more than 180 million years to download all the data from the web.
- Nearly 80% of companies estimate that 50%-90% of their data is unstructured. Think text, video, audio, web server logs, or social media activities.

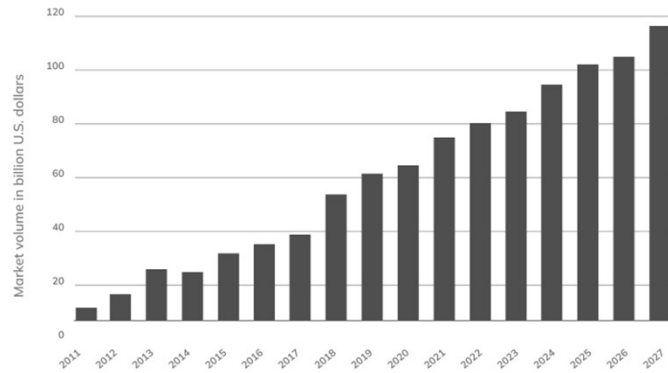
<https://www.brimco.io/analytics/big-data-analytics-statistics/>

56

Big Data Market

Big data market size revenue forecast worldwide from 2011 to 2027

(in billion U.S. dollars)



<https://innowise.com/blog/big-data-trends-2024/>

Source: Statista

57

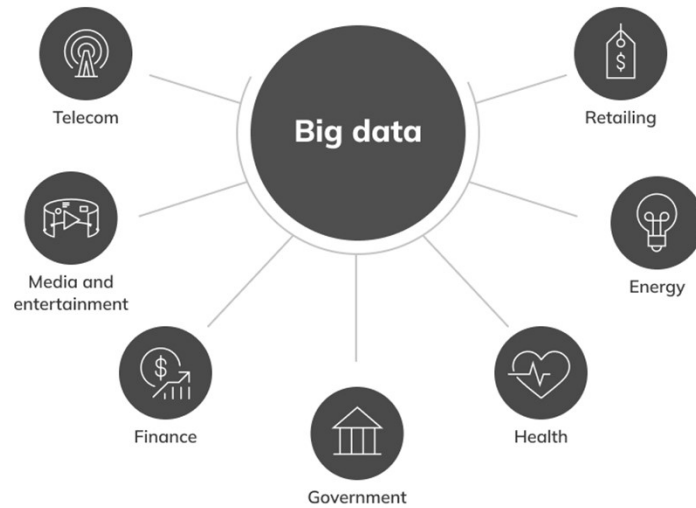
Big data analytics trends



<https://innowise.com/blog/big-data-trends-2024/>

58

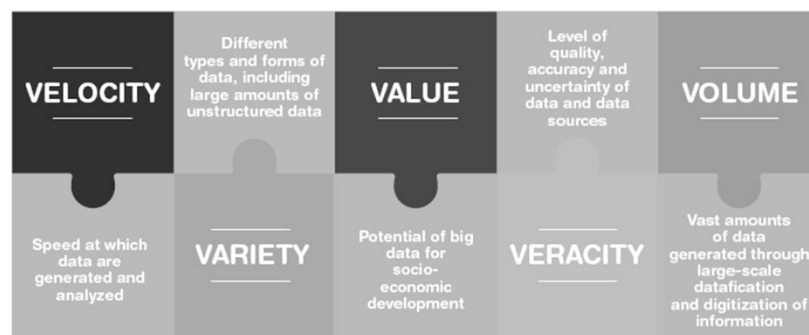
Industry-specific Solutions



<https://innowise.com/blog/big-data-trends-2024/>

59

Big data 5'V

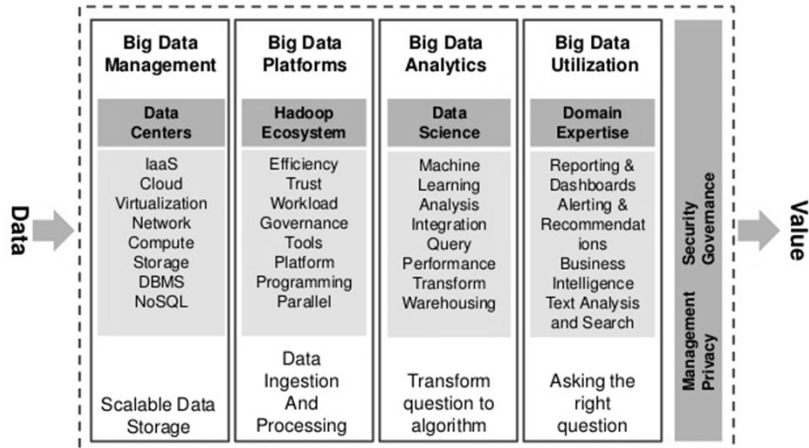


Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them (wikipedia)

60

60

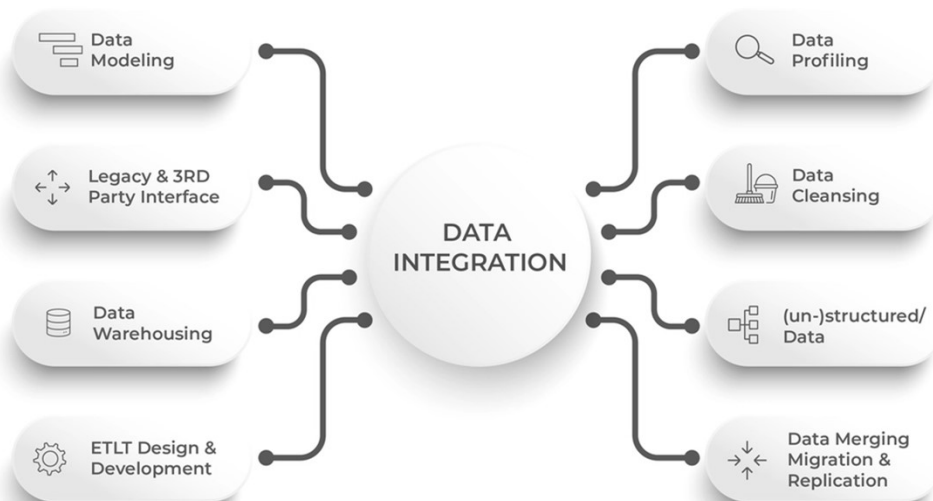
Big data technology stack



61

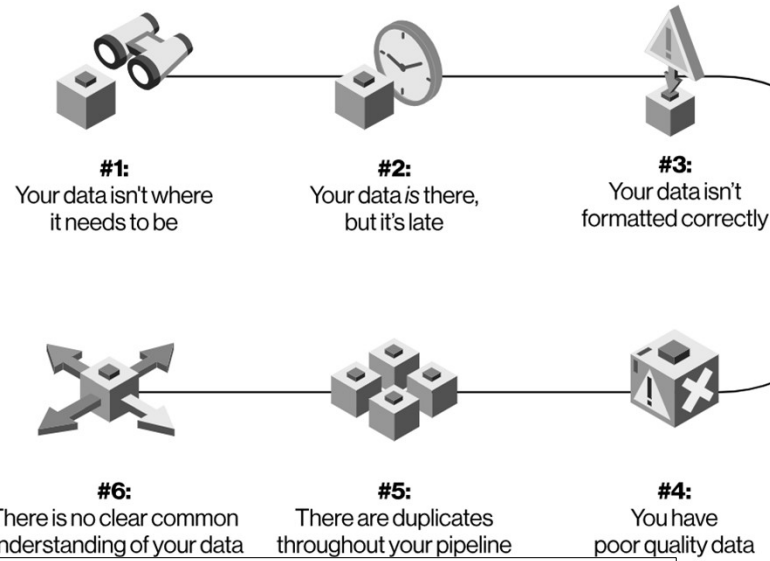
61

Data Integration



62

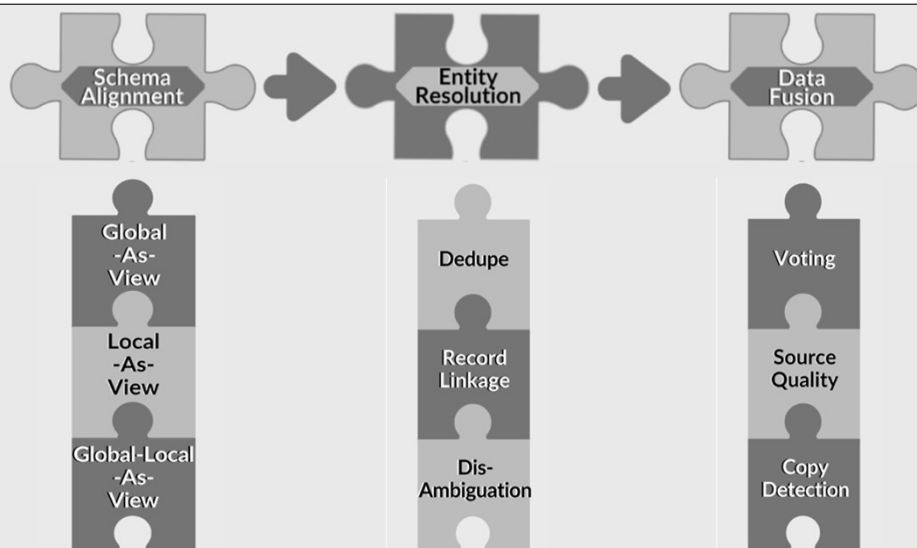
Challenges



<https://www.cloverdx.com/blog/biggest-data-integration-challenges>

63

Task for Big Data Integration



<https://www.selecthub.com/big-data-analytics/big-data-integration/>

64