Data Integration

Vu Tuyêt 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
	...

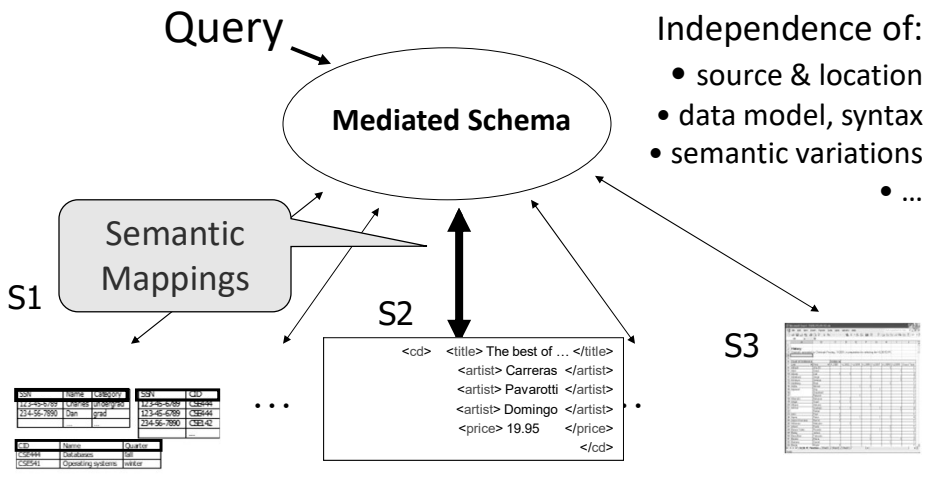
Courses:

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

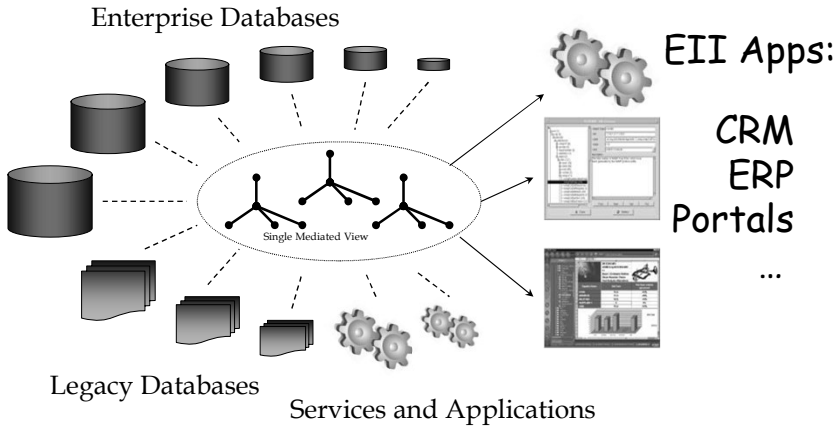
6

Applications of Data Integration

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

7

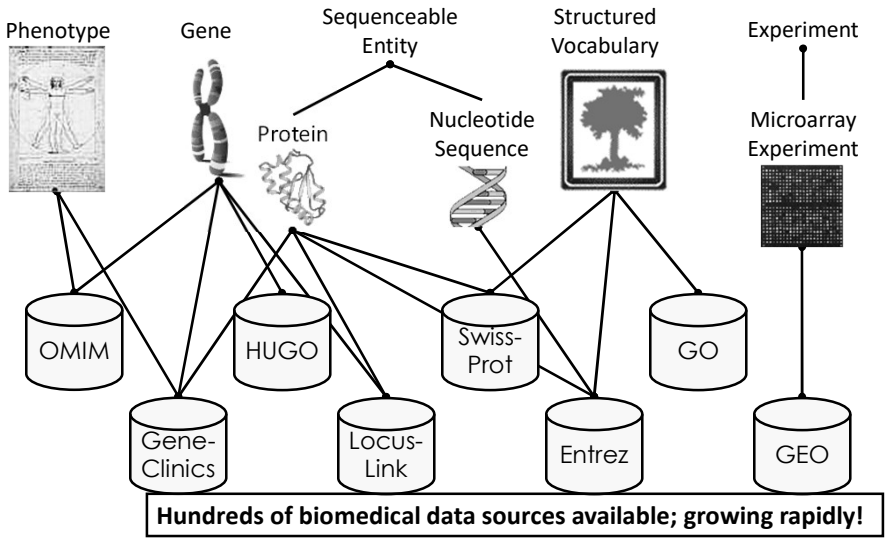
Application Area 1: Business



50% of all IT $$$ spent here!

8

Application Area 2: Science



9

Application Area 3: The Web

10

11

The Deep Web

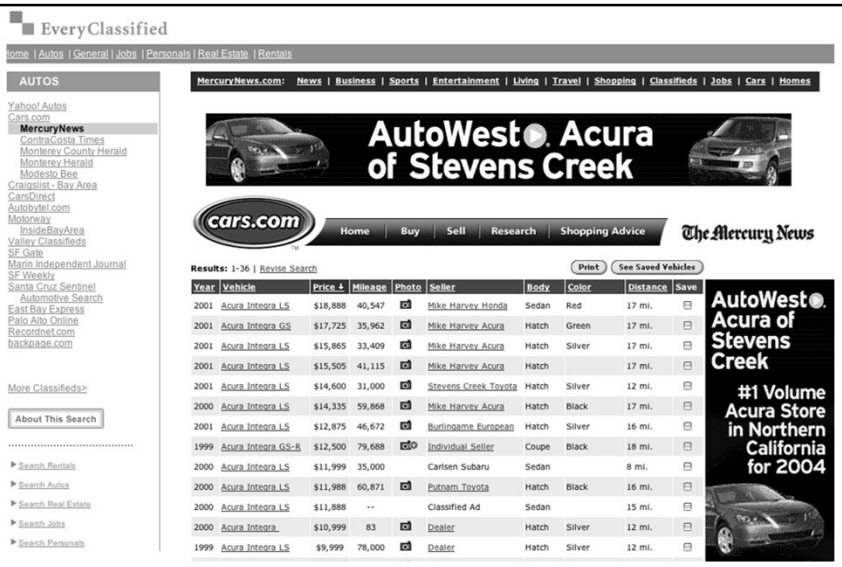
- Millions of high quality HTML forms out there
- Each form has its own special interface
  - Hard to explore data across sites.
- Goal (for some domains):
  - A single interface into a multitude of deep-web sources.

12

13

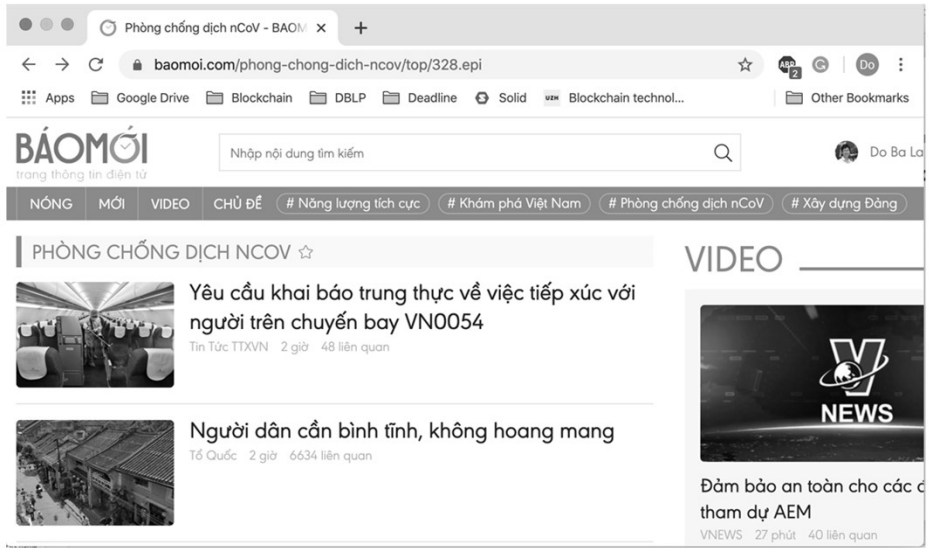
Create a single site to search for jobs/rentals/…

14



Easily traverse between the site by clicking its name

15



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

Services

- Services
- Customers
- Contracts

Credit Cards

- Customer
- CustDetail

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, dataspaces**

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

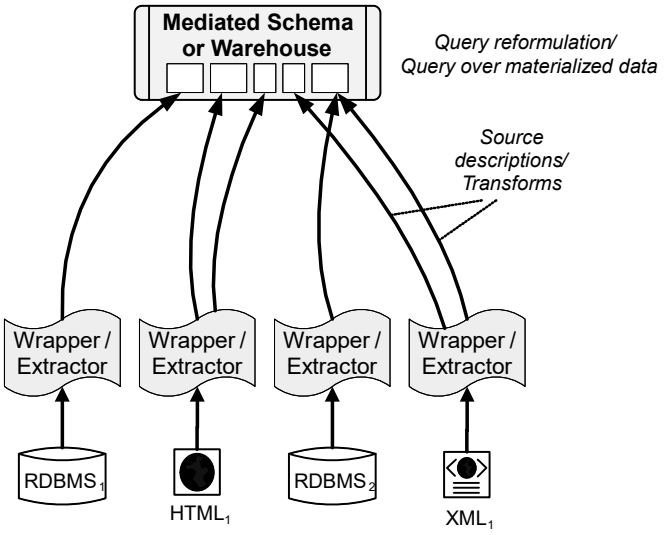
Virtual, Warehousing and in Between

• 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

Movie(title, director, year, genre)
Actors(title, actor)
Plays(movie, location, startTime)
Reviews(title, rating, description)

S1 **Movies** (name, actors, director, genre)

S2 **Cinemas** (place, movie, start)

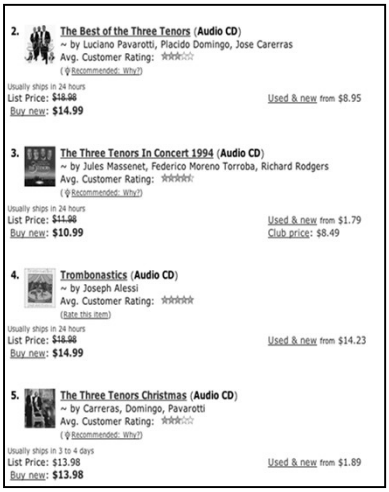
S3 **CinemasInNYC** (cinema, title, startTime)

S4 **CinemasInSF** (location, movie, startingTime)

S5 **Reviews** (title, date, grade, review)

30

Wrappers




```
<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, …

logic

CDs
Album
ASIN
Price
DiscountPrice
Studio

Books
Title
ISBN
Price
DiscountPrice
Edition

Authors
ISBN
FirstName
LastName

Artists
ASIN
ArtistName
GroupName

CDCategories
ASIN
Category

BookCategories
ISBN
Category

32

Woody Allen Comedies in NY

Mediated schema:

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

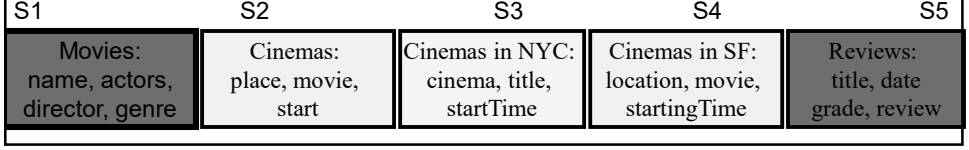
```
select title, startTime
from Movie, Plays
where Movie.title=Plays.movie AND
      location="New York" AND
      director="Woody Allen"
```

33

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

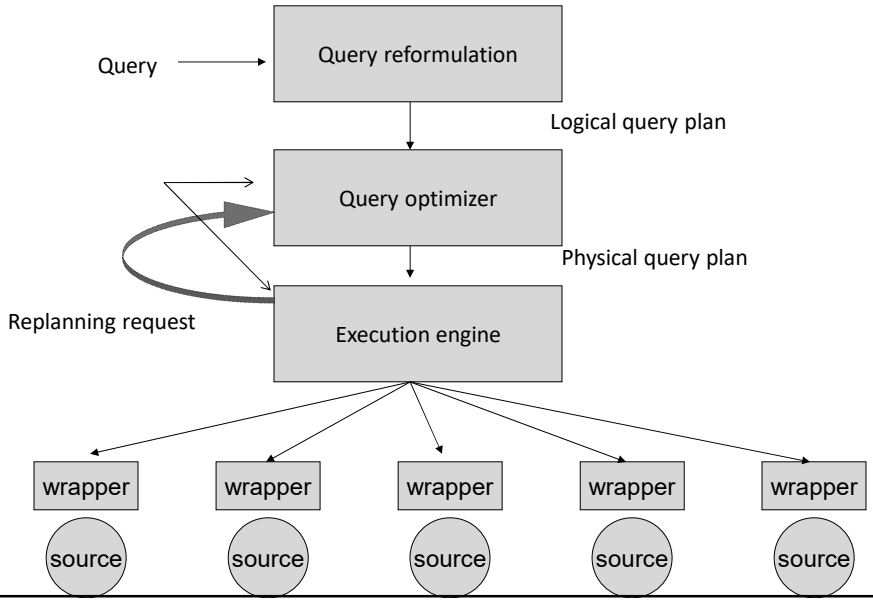
```
select title, startTime
from Movie, Plays
where Movie.title=Plays.movie AND
      location="New York" AND
      director="Woody Allen"
```

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



34

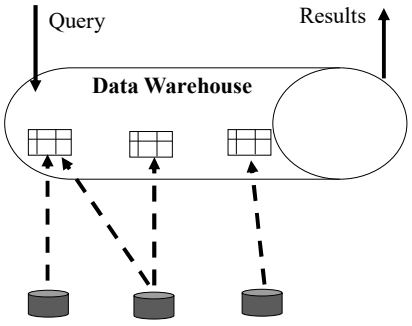
Query Processing



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

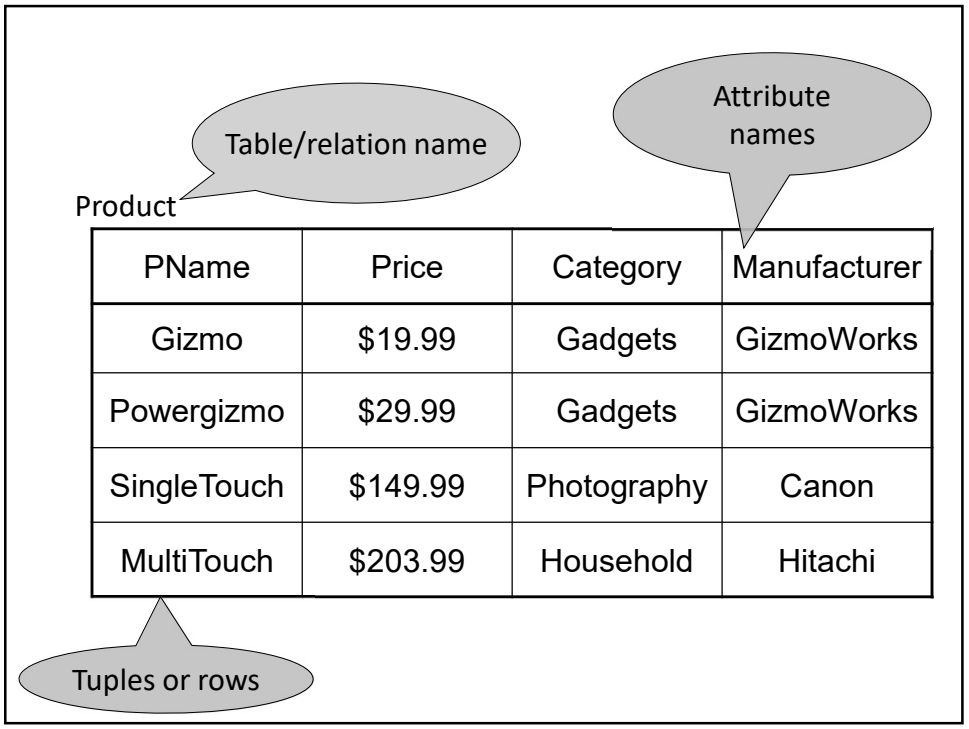
Relational Terminology

Relational schemas
- Tables, attributes

Relation instances
- Sets (or multi-sets) of tuples

Integrity constraints
- Keys, foreign keys, inclusion dependencies

40



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:

empID, 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)

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

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

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

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

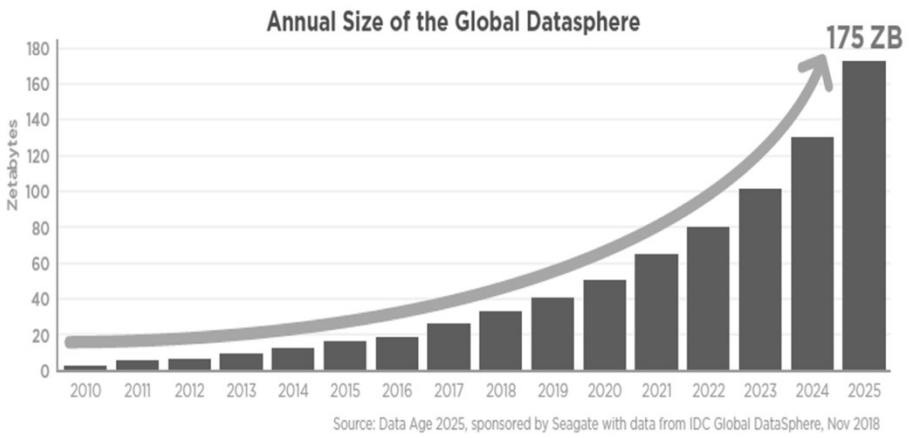
- Data integration: abstract away the fact that data comes from multiple sources in varying schemata.
- Problem occurs everywhere: it's key to business, science, Web and government.
- Goal: reduce the effort involved in integrating.
- Regardless of the architecture, heterogeneity is a key issue.
- Architectures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

53

Big Data …. Integration

54

How big is big data?



55

55

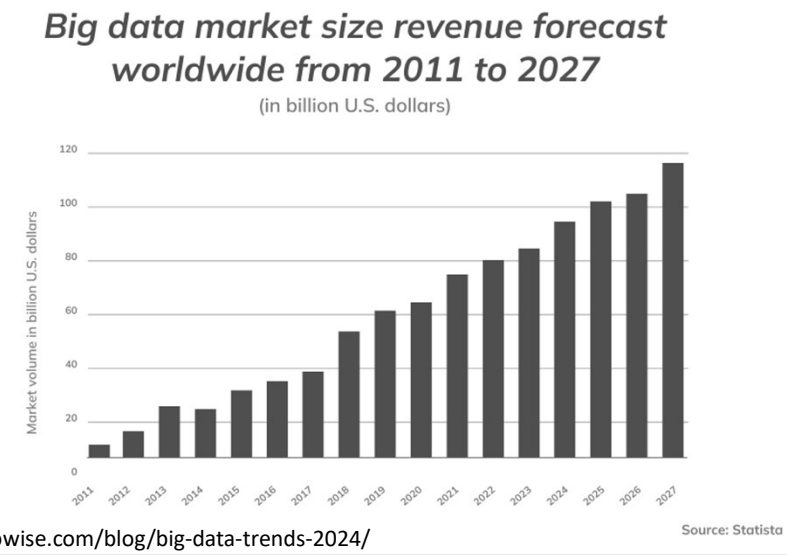
Big Data Growth Statistics

- 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



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

57

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

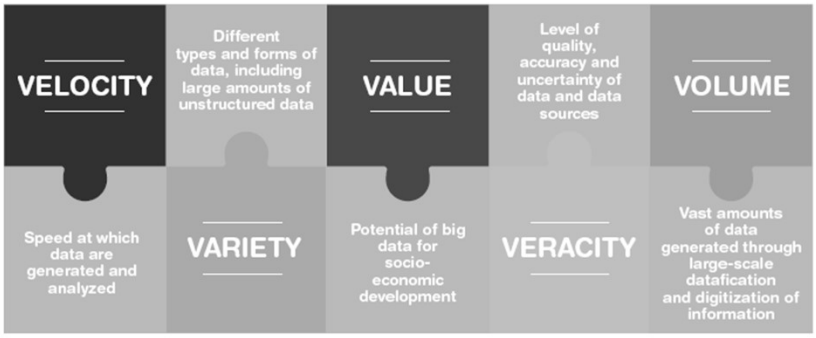
58

Industry-specific Solutions

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

59

Big data 5'V

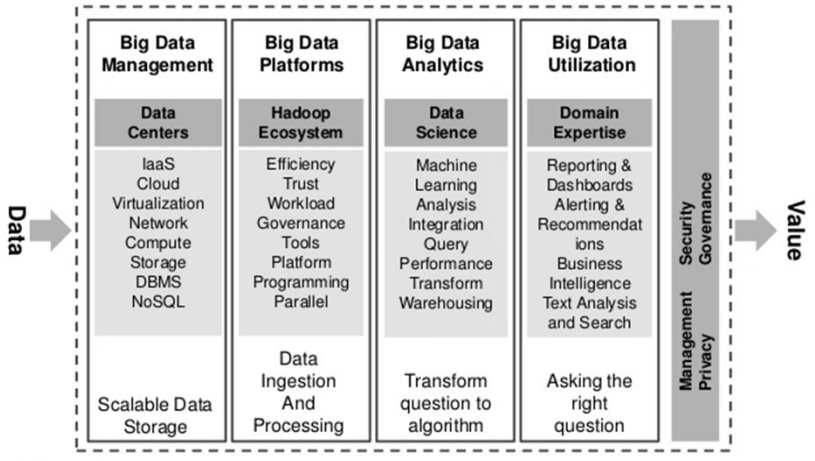


Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them (wikipedia)

60

60

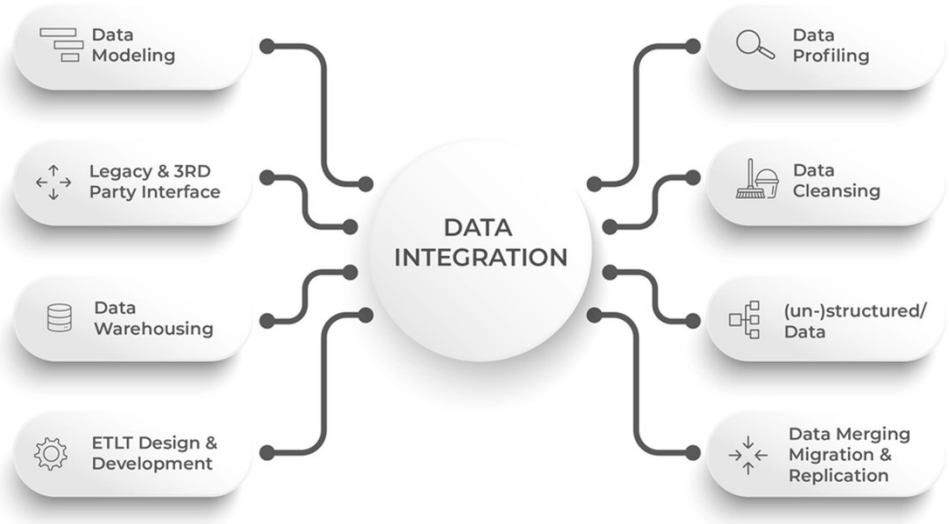
Big data technology stack



61

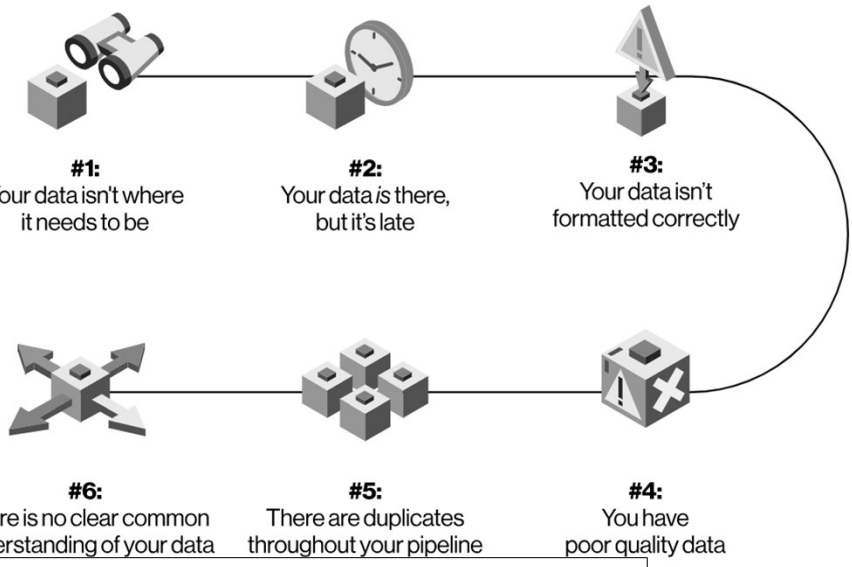
61

Data Integration



62

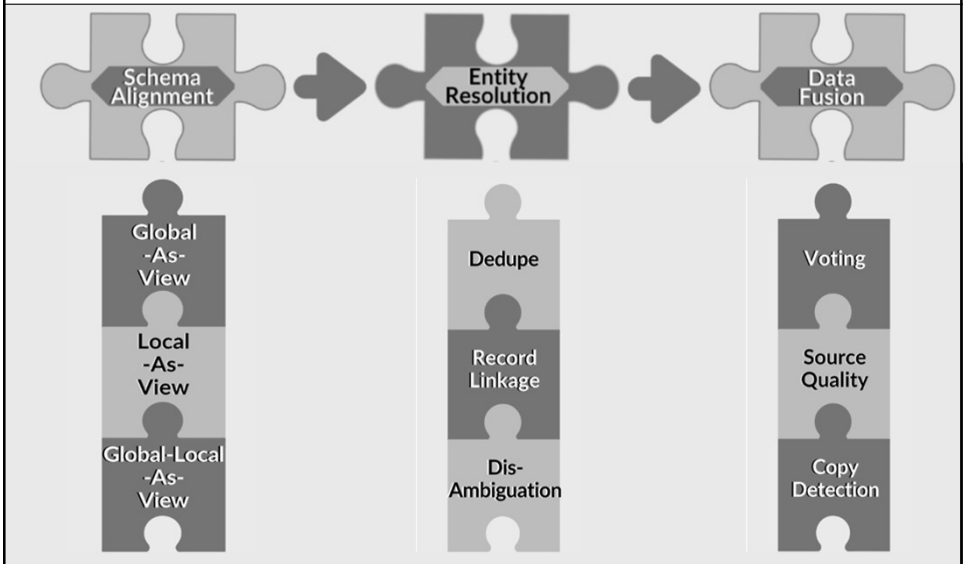
Challenges



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