

data integration

introducció

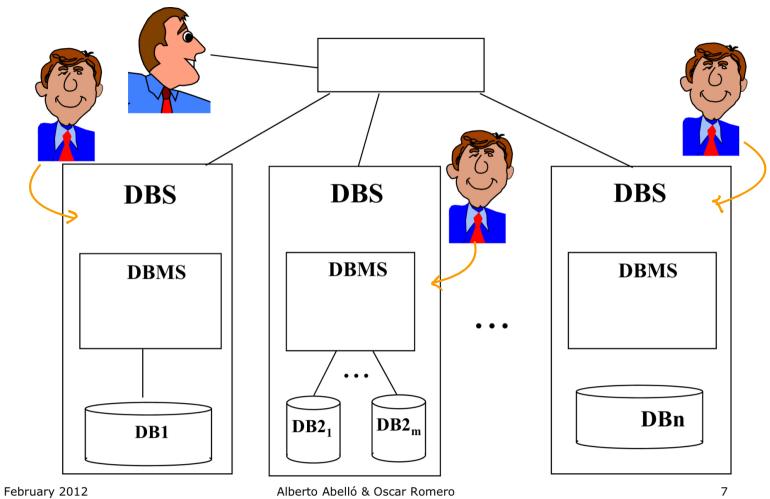
- problema ve quan les bases de dades estan definides per diferents persones -> apareixen heterogenitats:
 - system: de hardware/software
 - semantic: significat (difícil de resoldre)
 - e.g. un cas de covid a què equival?
 - els tenim a nivell:
 - d'instàncies
 - de classes
 - com ho resolem? wrapper-mediator

com ve el problema?

- una persona vol fer una consulta, i, per tant, vol obtenir una sola resposta (no moltes) -> necessitem que hi hagi un sol lloc on buscar-ho
- no parlem de connectivitat en les bases de dades, ni d'intercanvi de fitxers, ni d'accés remot, ni de multiclient/multiserver, ni de distribució de la base de dades
- el problema és: 1 consulta = 1 resposta (amb el mateix significat -> sense heterogenitats)

solutions possibles

- superman -> solució poc realista
 - un usuari ho sap tot
 - les BDs disponibles
 - els models de dades
 - els llengüatges de consulta
 - descomposar la query
 - i recompor-lo en una sola resposta
- crear una BD contenint totes les dades necessàries
 - e.g. ERP (SAP)
 - et dissenyen la bd
 - has de moure les teves dades al ERP
 - s'ha de modificar totes les apps perquè siguin compatibles
 - testejar-ho tot
- crear una capa de software a la part superior de les bases de dades que divideixi automàticament les consultes i integri les respostes
 - afegir una nova capa de software que defineixi dos nivells d'accés
 - processar automàticament les consultes
 - bd's segueixen igual però tenim un nivell per sobre



realment el que definim és una base de dades distribuïda

"Una base de dades distribuïda (DDB) és una col·lecció de múltiples bases de dades **lògicament interrelacionades** (conegeudes com a nodes o llocs) **distribuïdes** a través d'una xarxa informàtica.

Un sistema de gestió de bases de dades distribuïdes (DDBMS) és, per tant, el sistema de programari que permet la gestió de la base de dades distribuïda i fa que la distribució sigui transparent per als usuaris".

característiques importants de les que la definició anterior no parla (interdependents)

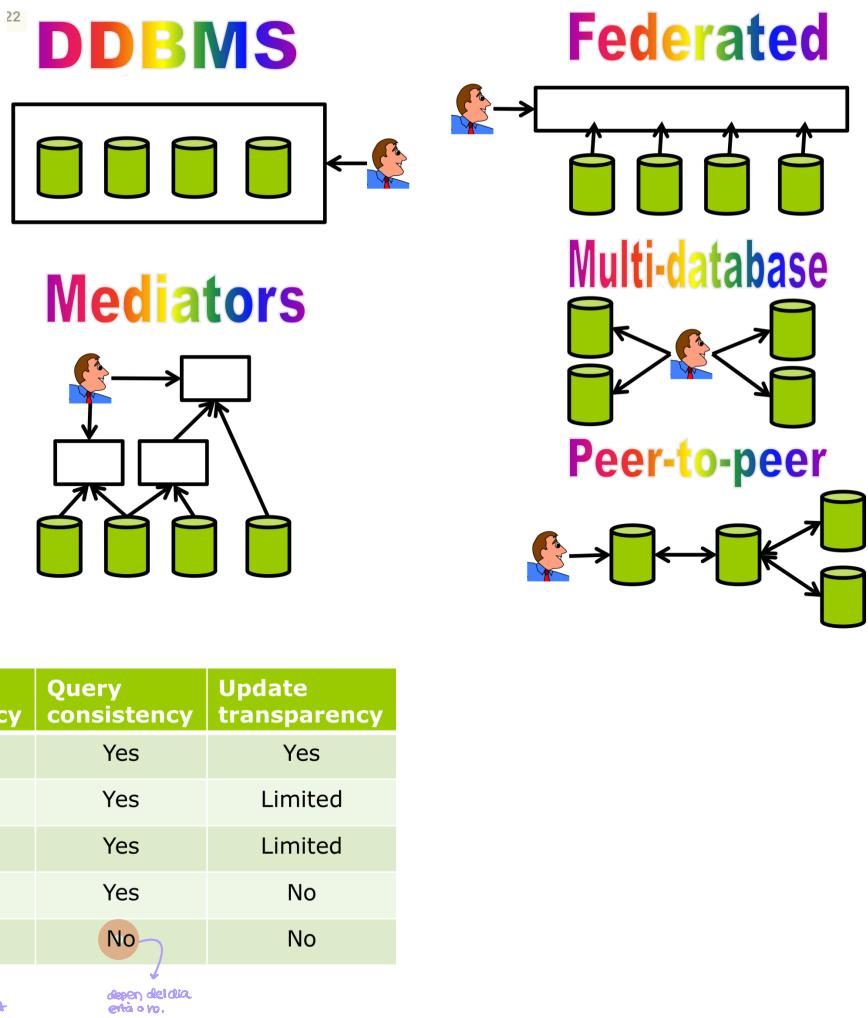


- autonimia
 - disseny: definicó de cada cosa
 - execució: decidir què executes i què no (e.g. confidencialitat de dades)
 - associació: tu pots formar part del sistema global o no
- heterogeneitat
 - sistema
 - semàntica
- polyglot persistence

Polyglot persistence, Martin Fowler

- ddbms
 - fàcil, sistema homogeni
- federated
 - driver amb més autonomia
- mediators
 - igual que l'anterior però més general
 - amb diferents capes d'integració / nivells
 - cada caixeta = 1 driver
- multi-database
 - opció superman d'abans (poc realista)
- peer-to-peer
 - torrent
 - bitcoin
 - ...

comparació:

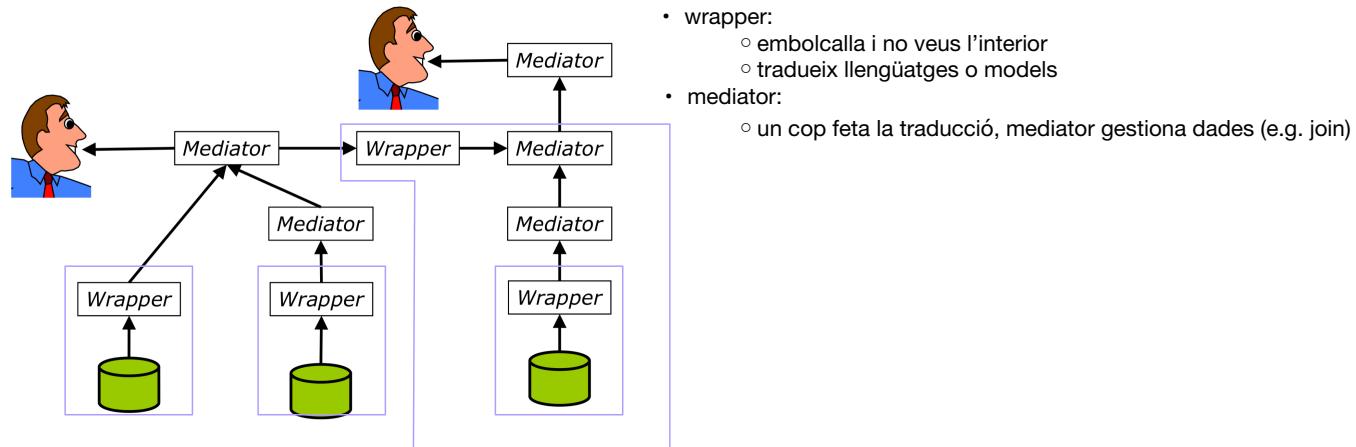


	Autonomy	Central schema	Query transparency	Query consistency	Update transparency
DDBMS	No	Yes	Yes	Yes	Yes
Federated	Yes	Yes	Yes	Yes	Limited
Mediators	Yes	No	Yes	Yes	Limited
Multi-database	Yes	No	No	Yes	No
Peer-to-Peer	Yes	No	Yes	No	No

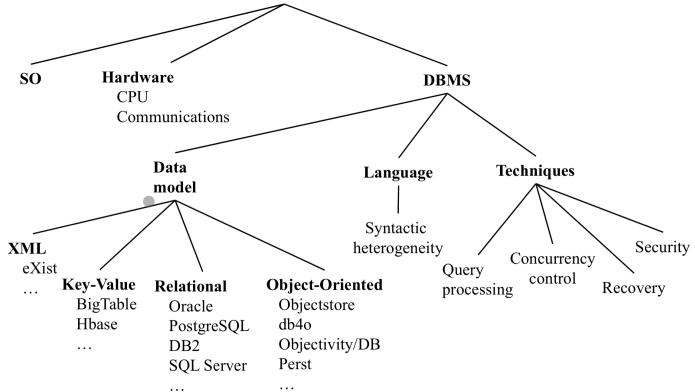
Annotations in the table:

- A red circle labeled "No" is placed over the "Query transparency" cell for Multi-database.
- A red circle labeled "No" is placed over the "Update transparency" cell for Peer-to-Peer.
- A blue arrow points from the "No" in the Multi-database row to the "No" in the Peer-to-Peer row with the text "tu no contestes tot".
- A blue arrow points from the "No" in the Peer-to-Peer row to the "No" in the Multi-database row with the text "dependrà dia a dia si es o no".

arquitectura wrapper-mediator



heterogenitats de sistema



heterogenitats semàntiques

> instàncies

- presència/absència
- nombre de valors (multi/mono-valued)
- existència de valors nulls
- valor

> classes

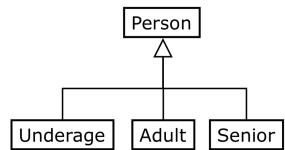
- extensió (e.g. coding colors)
- nom
- atributs/mètodes
 - presència/absència
 - arity
 - constants d'integritat (considerar que edat màxima és 100, quan pot ser més en una altra bd)
- domini
 - claus
 - tipus de dades (string vs integer per un número de telèfon)
 - dimensió (e.g. volum o pes)
 - unitats de mesura (inch vs meter)
 - escala (litres vs m^3)
- constraints

> estructura

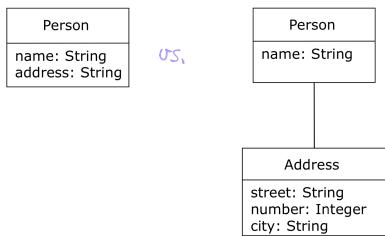
- generalització / especialització



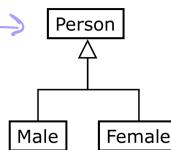
vs.



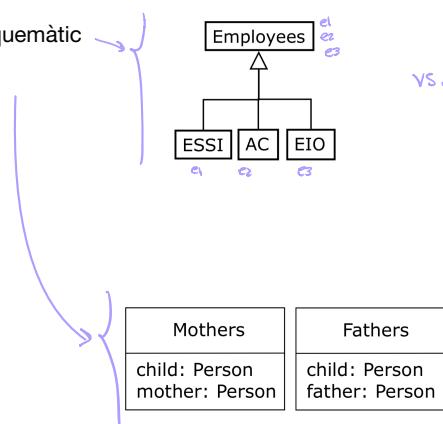
- agregació



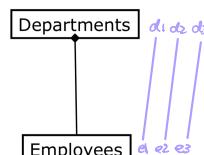
vs.



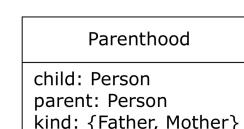
- esquemàtic



vs.



vs.



- generalització / especialització
 - criteri (p. ex., Sexe vs feina)
 - grau i caracterització (per exemple, diferents grups d'edat)
 - tipus (és a dir, complet o no, disjunt o superposat)
 - restriccions d'integritat (per exemple, efecte de supressió)
- agregació / descomposició
 - tipus d'agregació (és a dir, composició o no) classes Classes participants
 - especialització en la classe agregada (per exemple, pare contra pare)
 - col·lecció a la classe agregada (per exemple, projectes contra subprojectes)
 - composició a la classe agregada (per exemple, adreça vs carrer + número + ciutat)
 - tipus de col·lecció de particions (és a dir, completa o no, disjunta o superposada)
 - classe de components de la col·lecció (per exemple, col·lecció de comtats versus col·lecció d'estats)
- esquema
 - especialització vs Composició
 - dades vs metadades

Data Quality

Introducció

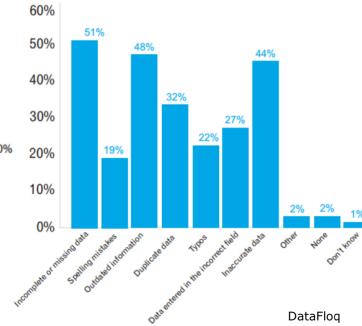
- els problemes poden passar amb només un esquema o amb variacions (com més pitjor)
- exemple: pagar pensió si no saps a quina setmana s'ha mort (exemple de classe, setmana 99)

Coses que es milloren si tens data quality

Reason for maintaining high quality data



Most common data errors



In Data We (Don't) Trust

25% of Critical Data in the World's Top Companies is Flawed



Nearly 40% of all company data is found to be inaccurate



92% OF BUSINESSES ADMIT THEIR CONTACT DATA IS NOT ACCURATE!

66% OF ORGANISATIONS BELIEVE THEY'RE NEGATIVELY AFFECTED BY INACCURATE DATA

Data Cleansing Helps Businesses

The implementation of a data quality initiative can lead to:

REDUCTIONS

Costs, budgets and overhead, oh my!



INCREASES

All the stuff you want more of



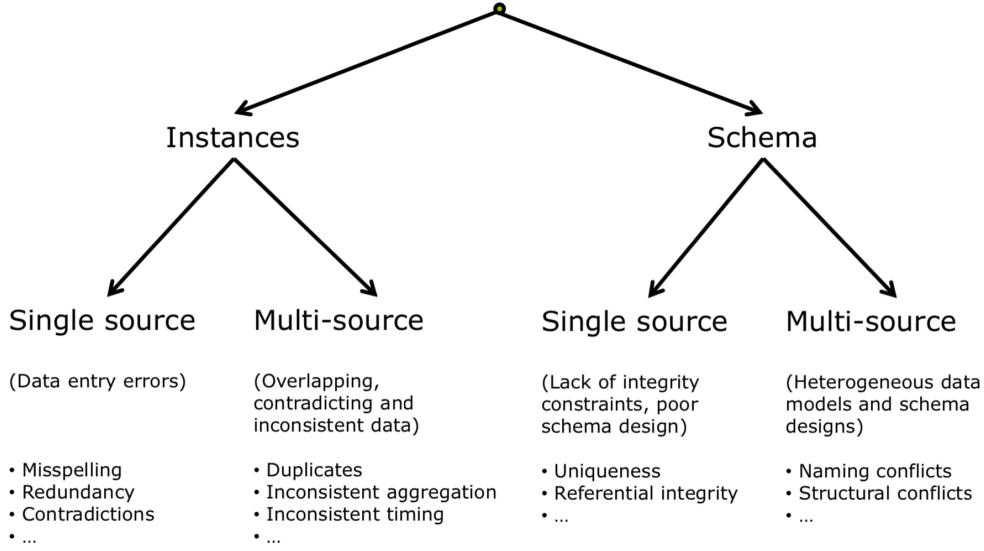
Aptitud per al seu ús

"Un usuari només pot avaluar el nivell de qualitat d'un conjunt de dades per a una tasca particular que s'executarà en un **context específic**, d'acord amb un conjunt de criteris, determinant així si aquestes dades es poden utilitzar o no **amb aquesta finalitat**".

- error pot ser més o menys greu dependent del seu ús:

- si tens @ de correu erroni
 - client que li envies newsletters: no passa res
 - client que li envies factures: perds diners!

conflictes de les dades



mesures

- completeness
- accuracy
- consistency
- timeliness
- relevance
- response time
- latency

completeness

“The degree to which a given collection of data describes the corresponding set of real-world objects.”

□ Missing entities → falta alumne

□ Missing values → falta foto d'alumne → + fàcil

$$Q_{Cm}(A_i) = \frac{|\{R(\text{NotNull}(A_i))\}|}{|R|} = \frac{3}{5} = 60\%$$

$$Q_{Cm}(R) = \frac{|\{R(\bigwedge_{A_i \in R} \text{NotNull}(A_i))\}|}{|R|} = \frac{4}{5} = 80\%$$

and

n	x	
v	y	
x	n	
y	η	
z	z'	

accuracy

“The extent to which data are correct, reliable and certified error free.”

□ Free of typing errors

□ Appropriate precision

relevance

response time

latency

$$e_A = |v_A - v_{RealWorld}|$$

$$Q_A(A_i) = |\{R(e_{A_i} \leq \epsilon)\}| / |R|$$

$$Q_A(R) = |\{R(\bigwedge_{A_i \in R} e_{A_i} \leq \epsilon)\}| / |R|$$

↑ valor absolut

↑ cardinalitat

consistency

“The degree of violation of semantic rules defined over a set of data items.”

□ Integrity constraints

- Entity **PK**
- Domain **type**
- Referential **FK**
- User-defined **checks/assertions**

□ Coincidence of copies

- Temporal
- Permanent

$$Q_{Cn}(R, B) = |\{R(\bigwedge_{rule \in B} rule(A_1, \dots, A_n))\}| / |R|$$

↳ regular desactivades

timeliness

“How old the stored value of an attribute is with regard to the current value in the real world.”

$$age(v) = now - TransactionTime = 20 \text{ anys}$$

$$f_u(v) = \text{updates per time unit} = 0$$

$$Q_T(v) = (1 + f_u(v) \cdot age(v))^{-1} = 1/1 = 1 \approx 100\%$$

$$Q_T(A_i) = \text{Avg}_{v \in R[A_i]} Q_T(v)$$

$$Q_T(R) = \text{Avg}_{A_i \in R} Q_T(A_i)$$

$$\frac{\lambda}{1 + f_u(v) \cdot age(v)}$$