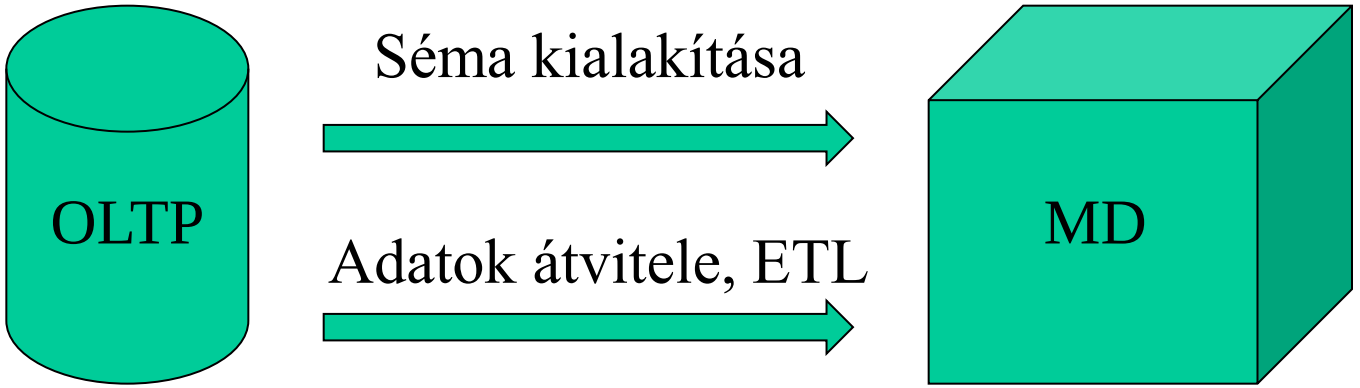
DW
6. előadás
MD séma tervezése, sémaintegráció
ETL

DW rendszer felépítésének fázisai



Séma kialakítása:

- új séma létrehozás (pull technology, kliens vezérelt),
- séma integráció (push technology, forrás vezérelt).

ETL-folyamat:

- kiolvasás,
- konverzió,
- ellenőrzés, tisztítás,
- ütemezés,
- beépítés, integrálás.

Általános integrációs szintek

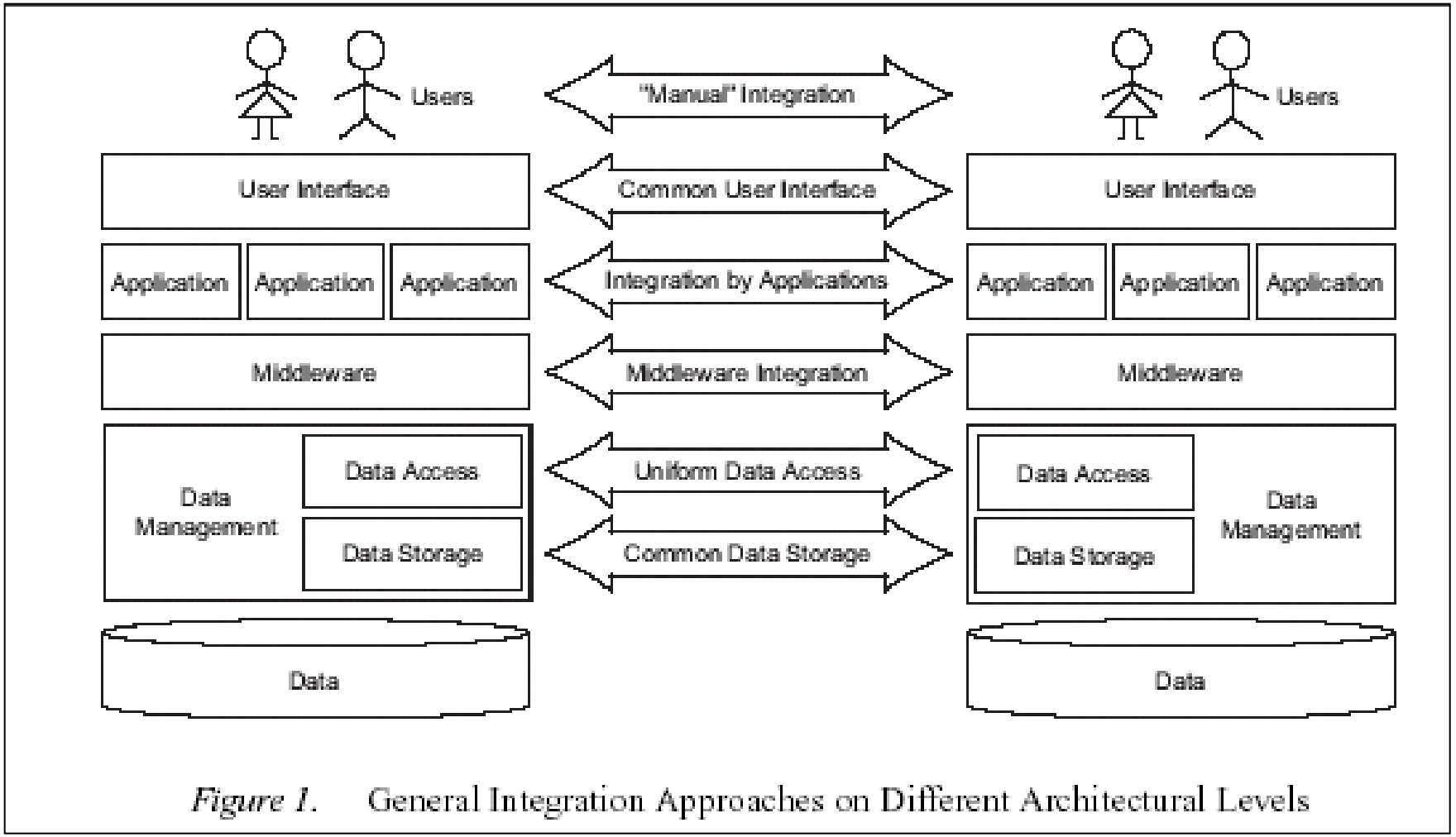
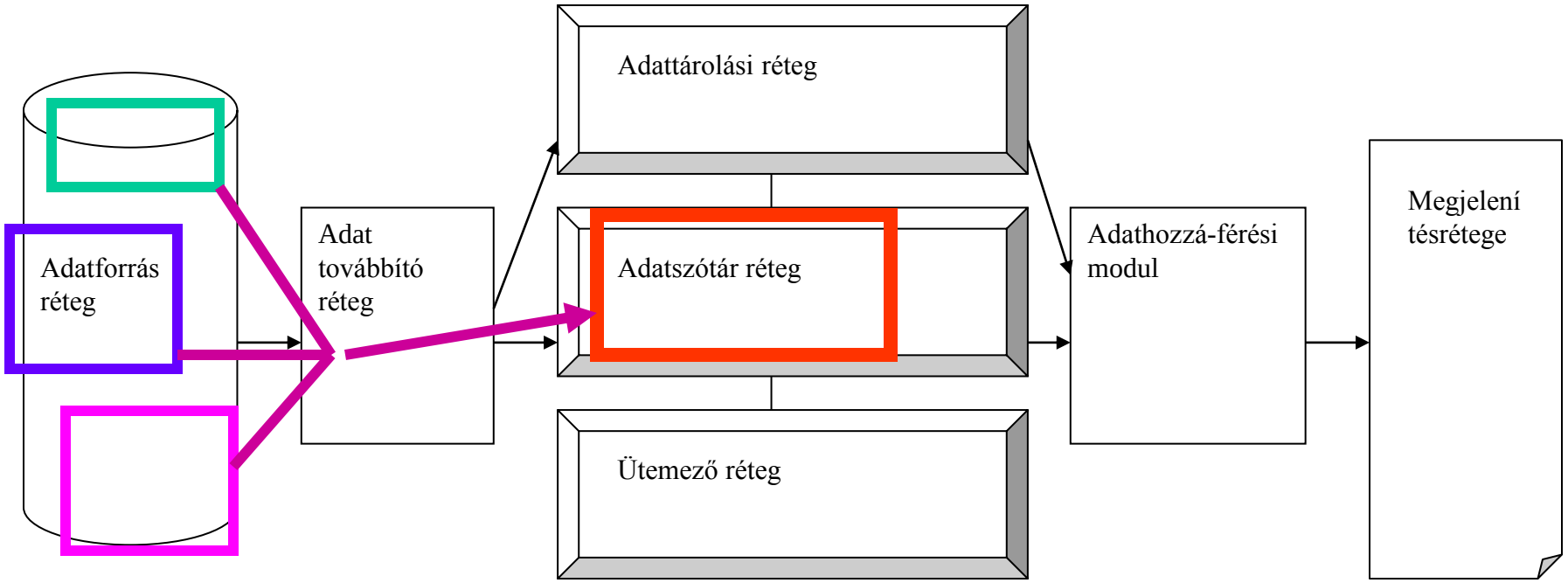
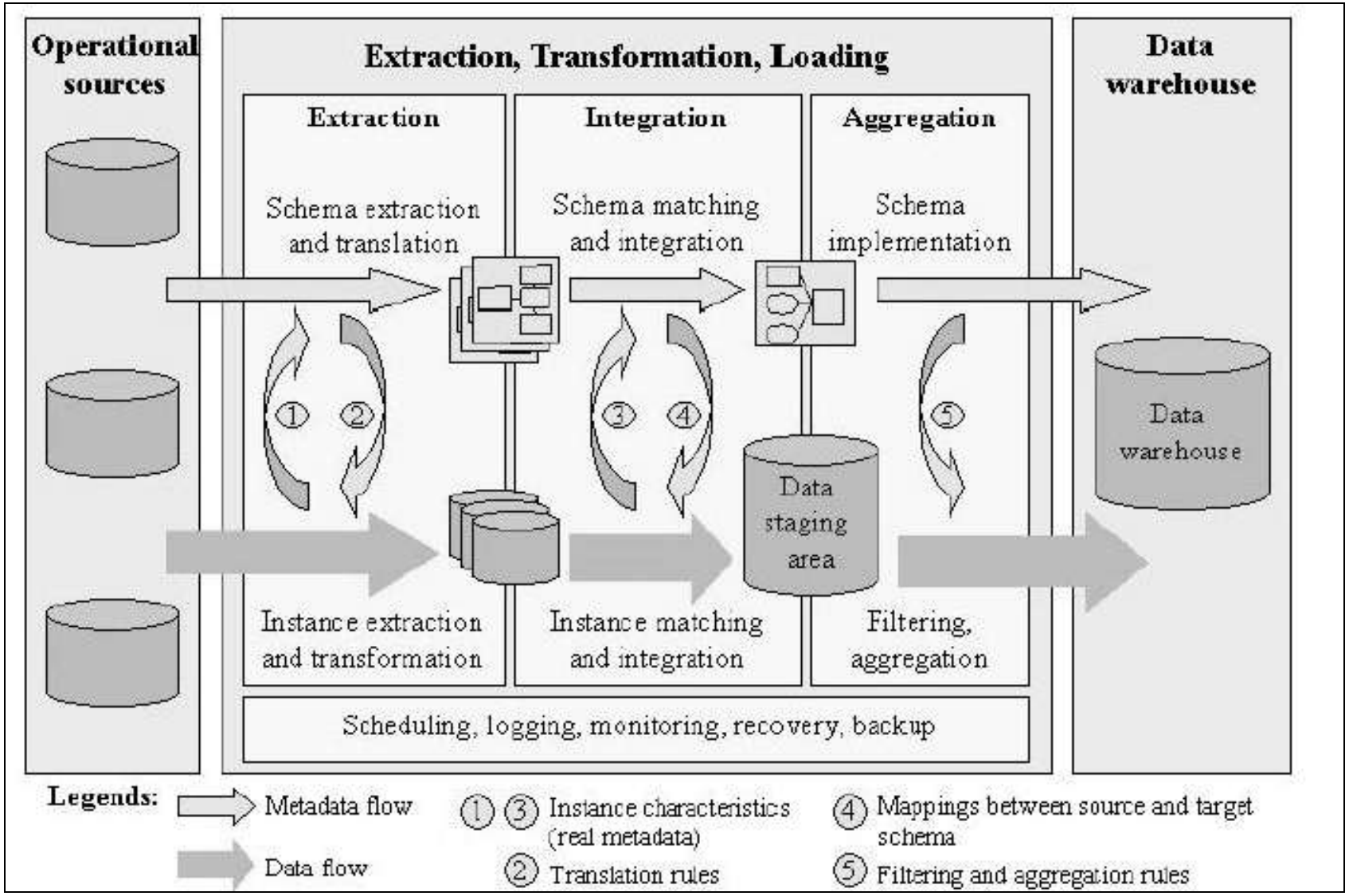


Figure 1. General Integration Approaches on Different Architectural Levels



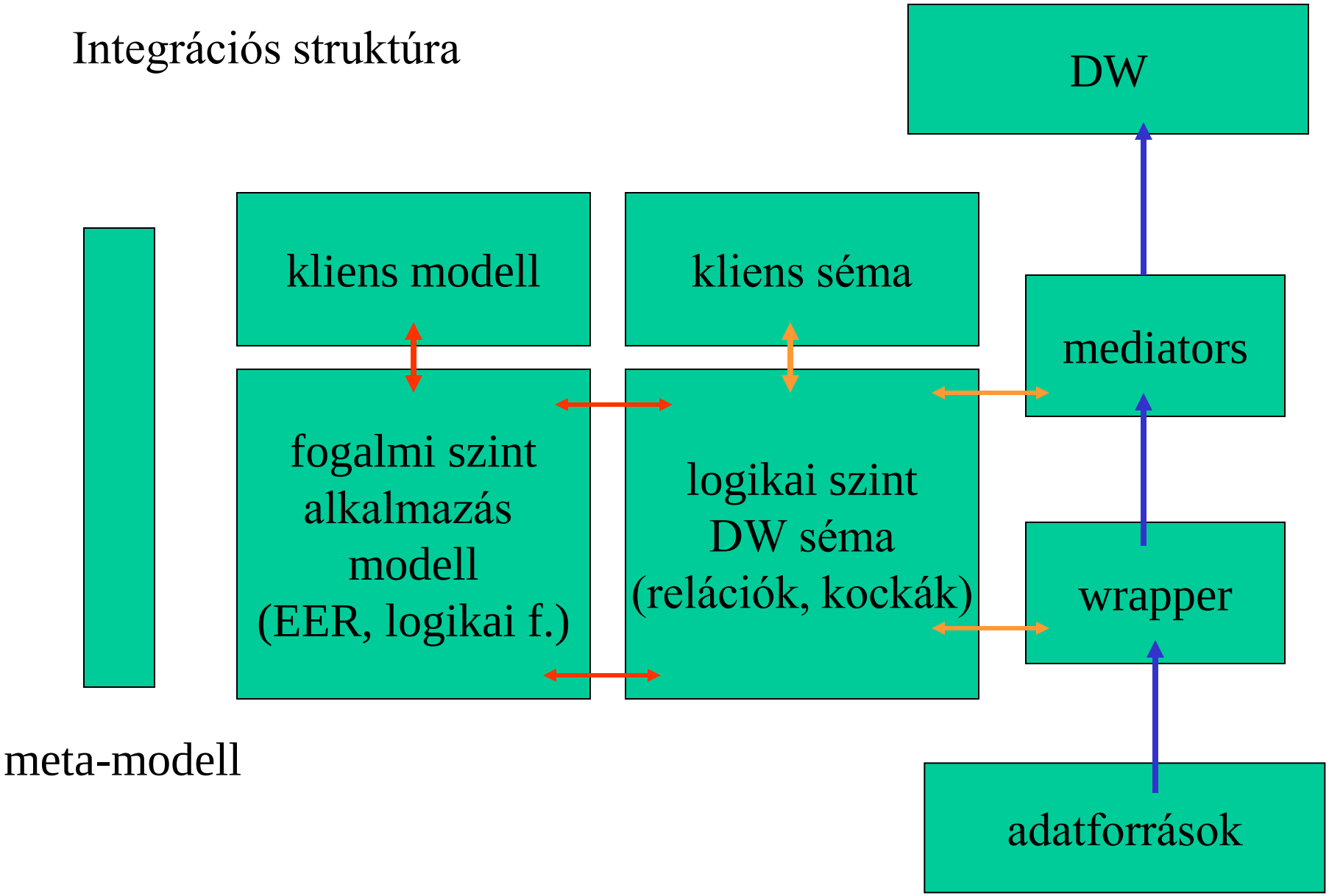
A források integrálása a DW rendszerek legfontosabb eleme

Elemei: - séma integráció,
- adat integráció:
-- virtuális,
-- valós.



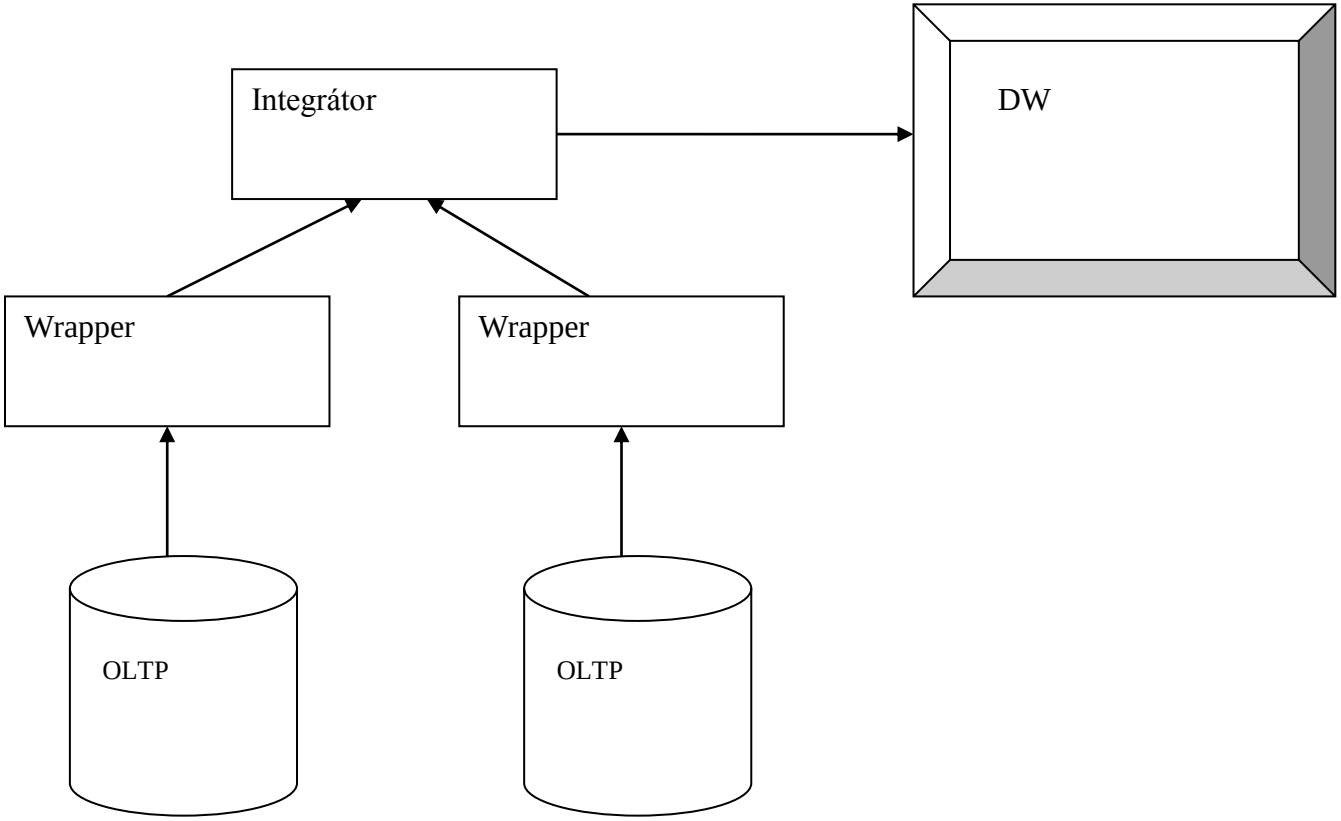
ETL-folyamatok

Integrációs struktúra



A modellben a wrapper komponens végzi az egyedi adatforrásból a **közös formátumra** történő **konverziót**.

Betöltési modul struktúrája



Integráció tervezési módszerek

Egyszintű (one-shot)
- csak egy célséma van.

Inkrementális
- modulok,
- független parciális sémák,
- inter-séma megkötések, szabályok.

Forrás vezérelt tervezés (push, séma integráció)
- vállalati szintű modell kialakítása a források alapján,
- a meglévő adatok határozzák meg az integrált modellt.

Kliens vezérelt tervezés (pull, új séma)
- a felhasználói igények kielégítése a cél,
- az igények határozzák meg az integrált modellt.

Adat integrációs lépések

- adat illesztés:
  - -- formátum,
  - -- kódolás,
  - -- érték,
- adatszűrés
  - (közös integritási feltételek),
- adat ellenőrzés
  - (inkonzisztencia feloldása).

A séma integrációban megadott leképzés (mapping) alapján működik

speciális feladatok:

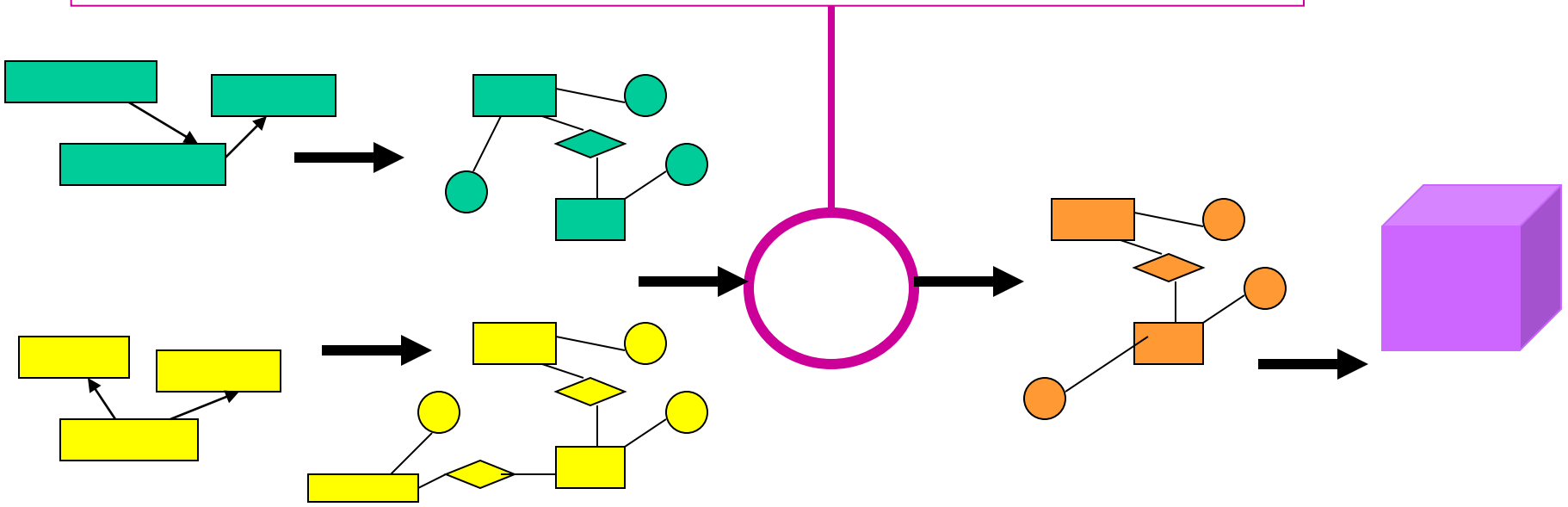
- adattisztítás,
- adatillesztés.

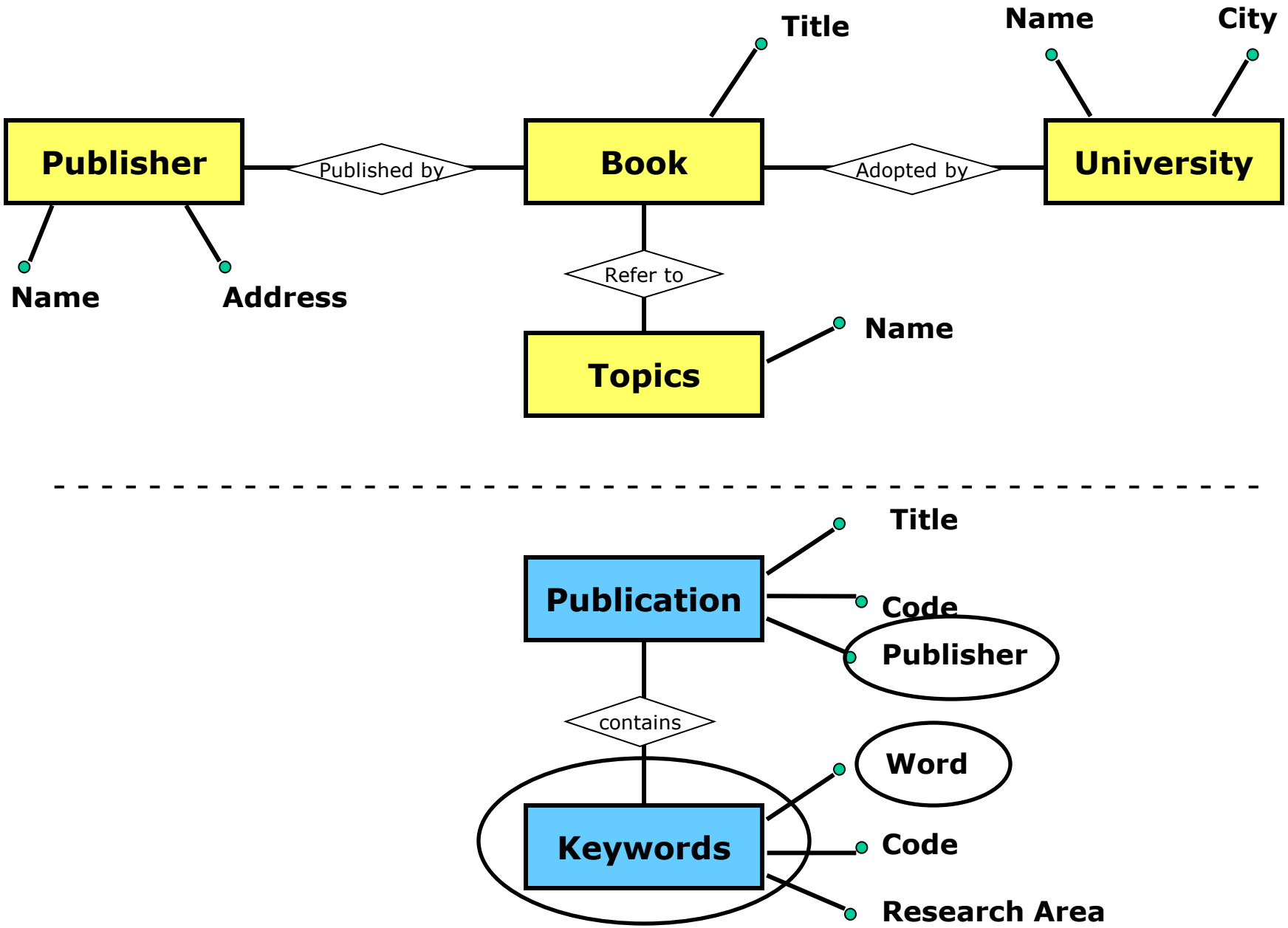
Séma integráció

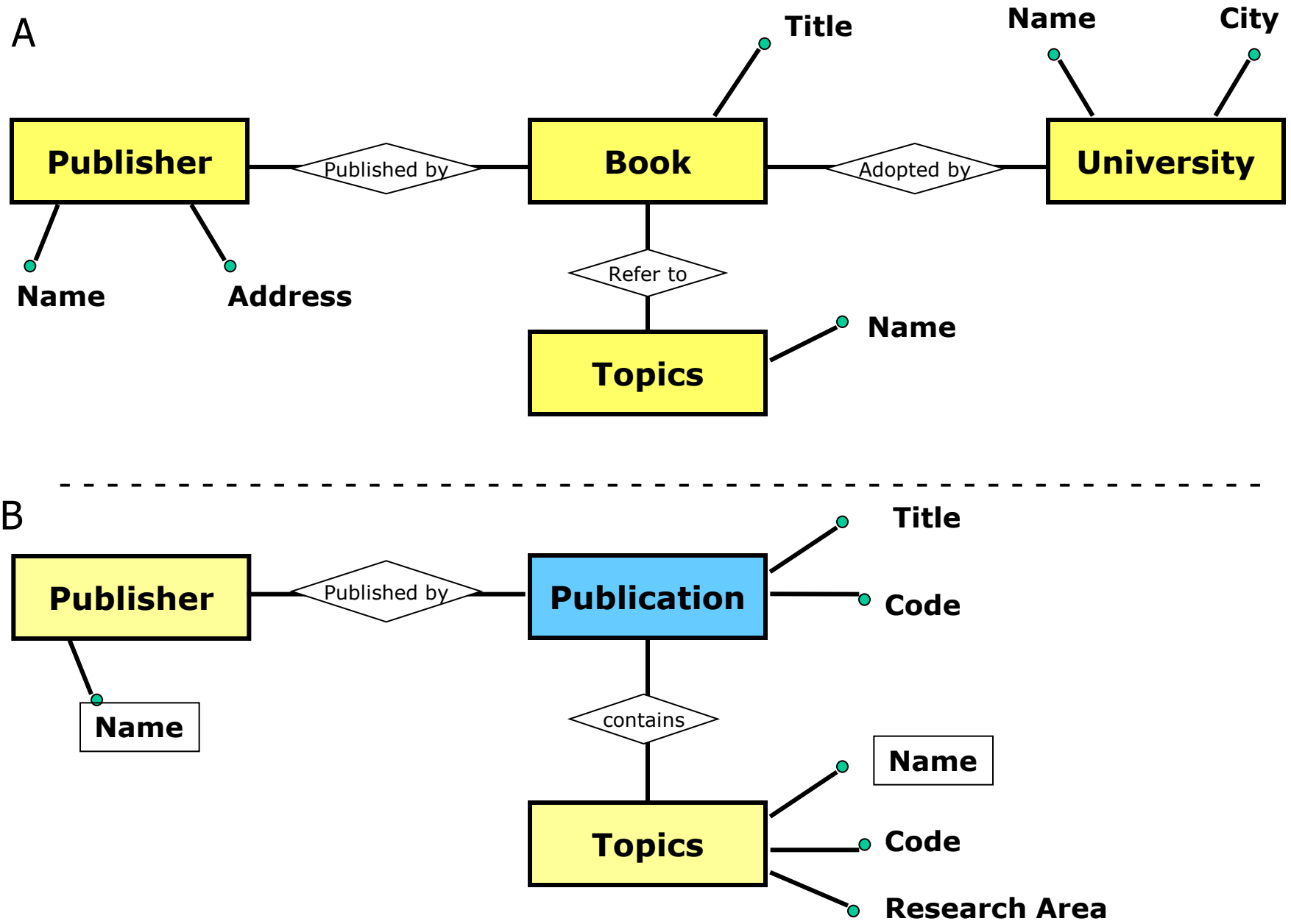
Célja: homogén, konzisztens közös séma előállítása.

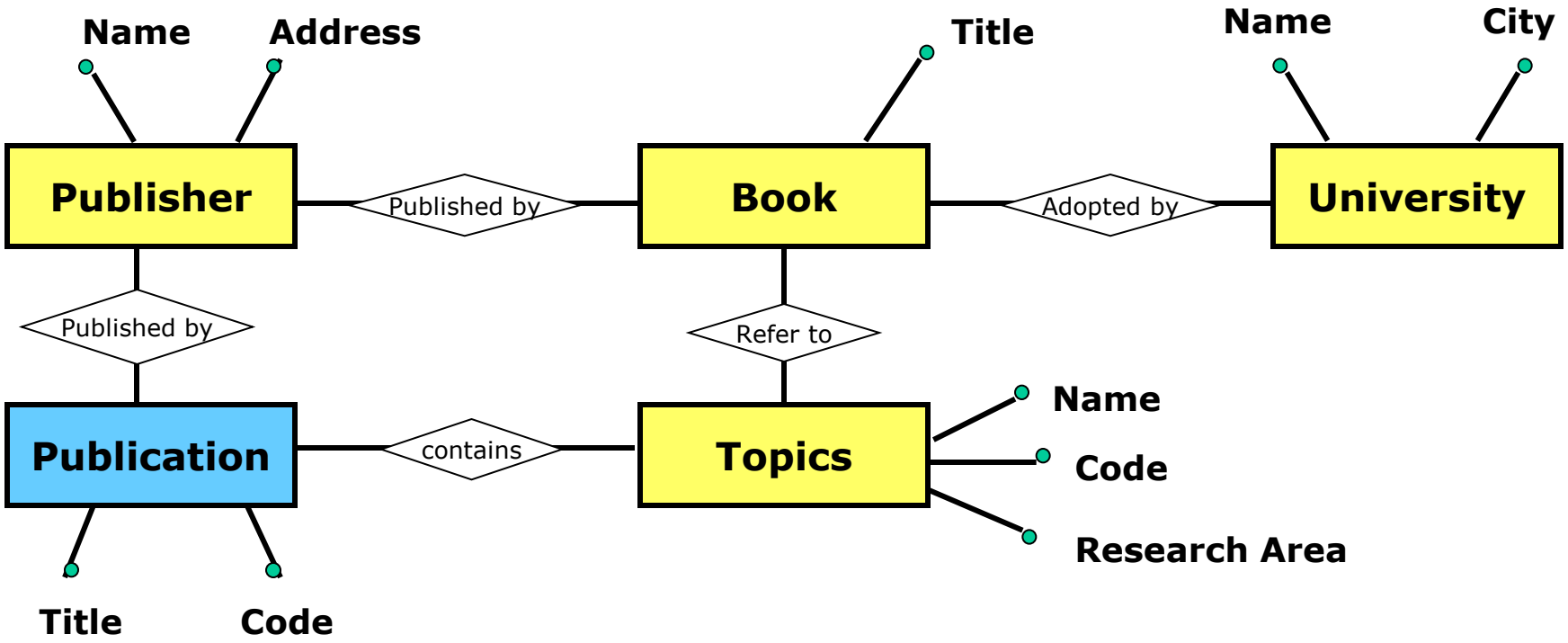
Integráció főbb lépései:

- előintegráció (preintegration),
- séma összehasonlítása (schema comparison),
- séma illesztése (schema conforming),
- séma összevonása (schema merging).



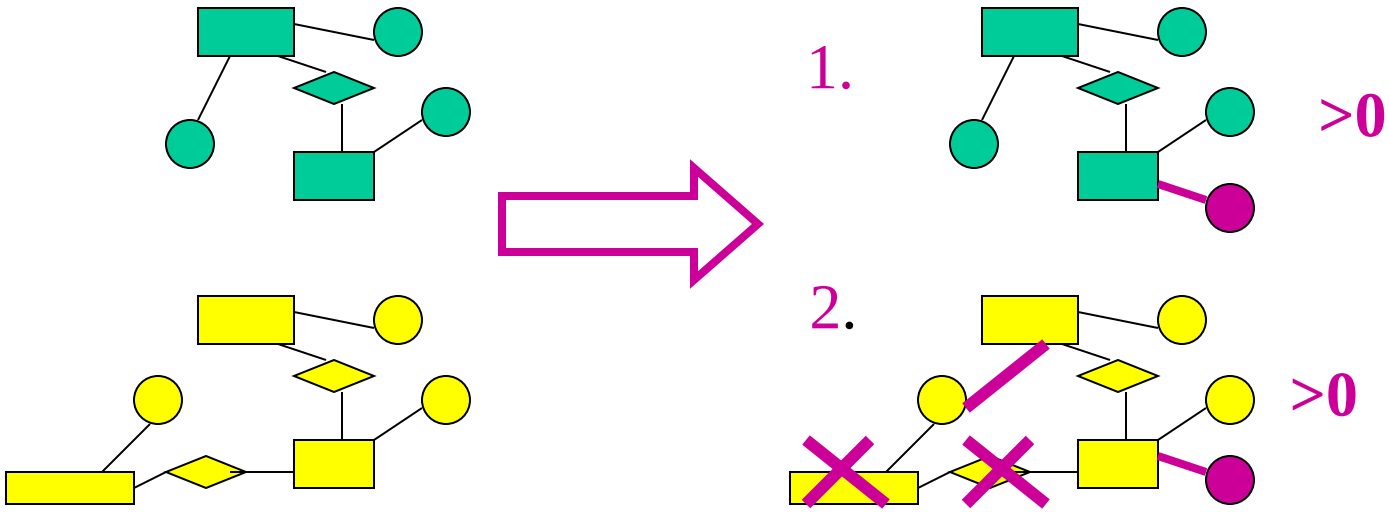

Publisher
Published by
Book
Adopted by
University
Title
Name
City
Name
Address
Refer to
Topics
Name
Publication
Title
Code
Publisher
contains
Keywords
Word
Code
Research Area


A
Title
Name
City
Publisher
Published by
Book
Adopted by
University
Name
Address
Refer to
Name
Topics
B
Title
Publisher
Published by
Publication
Code
Name
contains
Name
Topics
Code
Research Area


Name
Address
Publisher
Published by
Book
Title
Adopted by
Name
City
University
Published by
Refer to
Publication
contains
Topics
Name
Code
Research Area
Title
Code

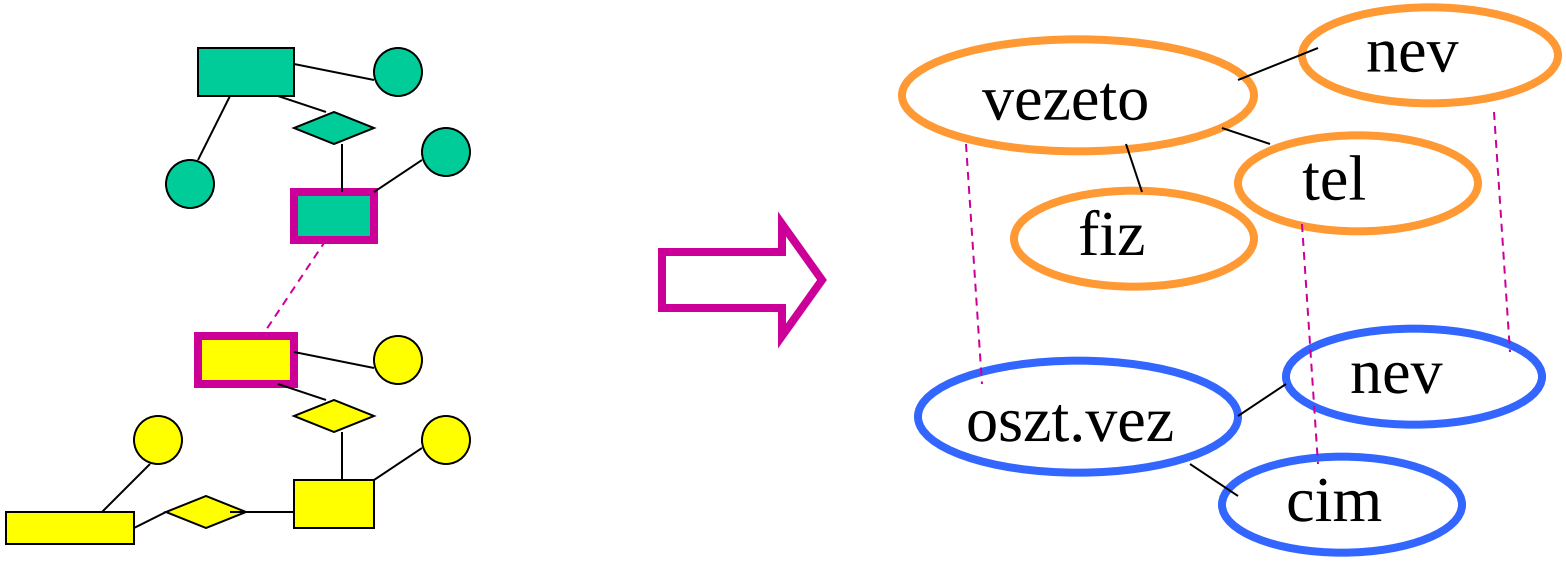
Előintegráció lépései

- az egyes sémák elemzése,
- integrációba bevonandó elemek kiválasztása,
- integrációs sorrend meghatározása,
- integritási elvek összegyűjtése,
- szemantikai kibővítés,
- közös szemantikai modellre alakítás
  (EER, ODL, formális logikai nyelv,...),
- adatszótár létrehozás.



Séma összehasonlítás lépései

- a különböző sémák elemei közötti kapcsolatok meghatározása,
- séma struktúra hasonlóság vizsgálata,
- modell heterogenitási konfliktusok feloldása,
- elnevezési konfliktusok feloldása
  (homonima/hasonlónevű/azonosalakú, szinonima/rokonértelmű),
- szemantikai konfliktusok feloldása,
- strukturális konfliktusok feloldása.

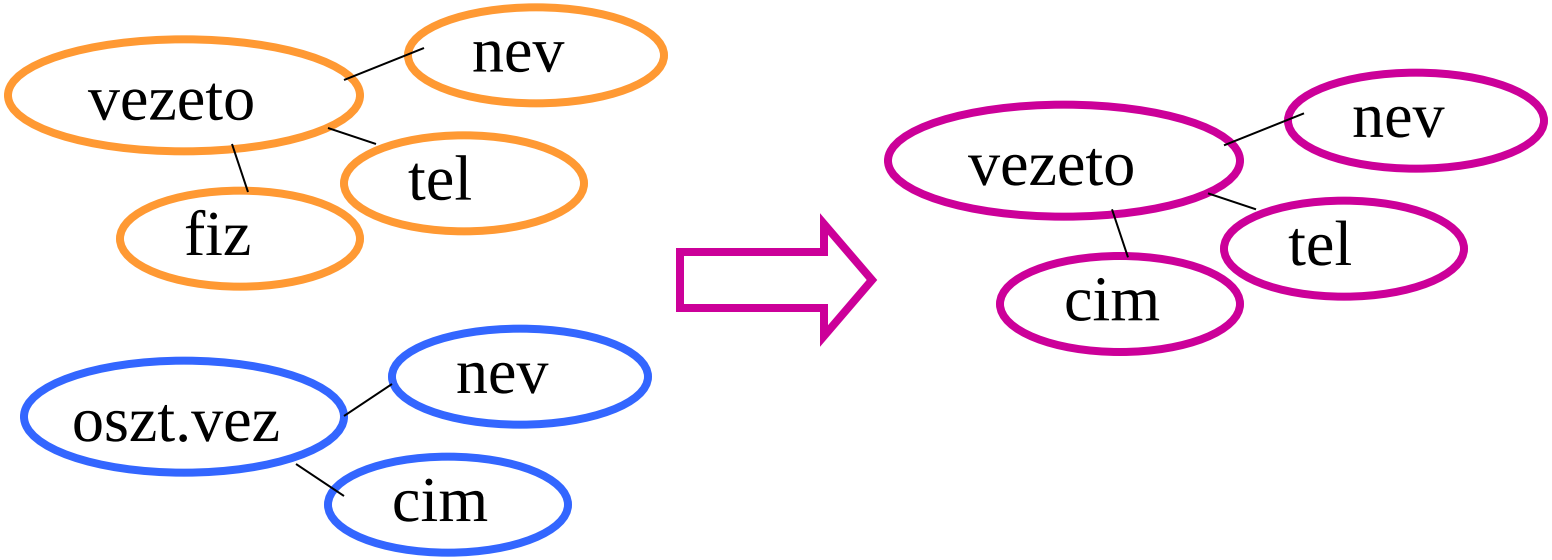


Séma illesztés elemei:

- elnevezés:
  - -- szinonimák,
  - -- általánosítás,
  - -- elírás,
- struktúra:
  - -- kapcsolatok,
  - -- szerkezet,
  - -- viselkedés.

Séma illesztés és összevonás lépései

- konfliktusok számbavétele,
- konfliktusok feloldása,
- sémák kombinálása,
- közös séma átalakítása
  (séma hasonlóság alapú vizsgálat),
- séma optimalizálása,
- teljesség, helyesség, minimalitás ellenőrzése.

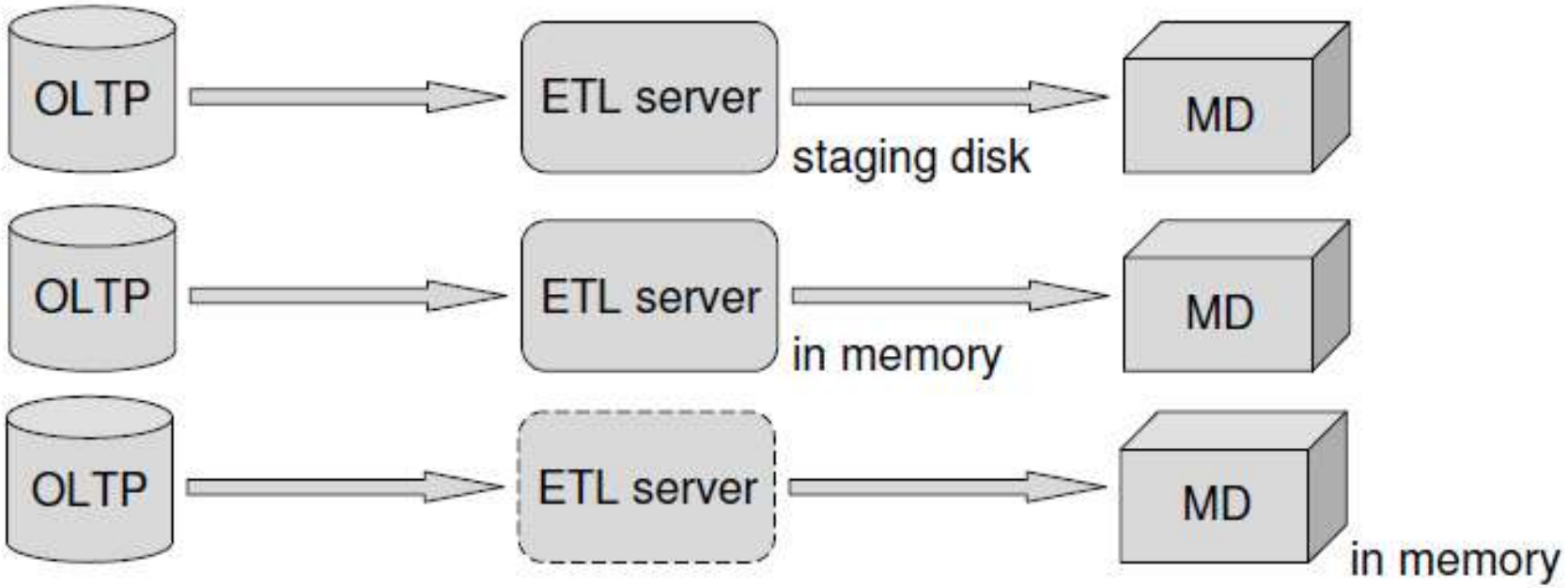


ETL-folyamatok áttekintése

Általános tervezési szempontok:
- ne terheljük le az OLTP rendszert,
- ne hozzunk át egyszerre túl sok adatot,
- ne legyen túl gyakori az adatátemelés,
- ne zavarjuk az OLTP rendszert,
- tiszta adatok kerüljenek be,
- kontrollált adatátvitel legyen
  (adatvesztés elleni védelem: leakage).

ETL-folyamatok áttekintése

Tipikus konfigurációk



Lehet külön ETL szerver vagy részleges ETL szerver

A transzformációt végezheti az MD rendszer is

ETL-folyamatok áttekintése

Ütemezési változatok:
- Az ETL szerver ütemezetten kezdeményezi és végzi el az olvasást,
- az OLTP DM triggereli és végzi el az adatok kiírását a köztes területre,
- az OLTP DM ütemezetten végzi el az adatok kiírását a köztes területre,
(DM – data modeller)
- az OLTP triggereli az eseménynapló átvitelét az OLAP rendszerbe.

ETL modul elhelyezése:
- middleware elem,
- OLAP DW elem,
- OLTP DB elem.

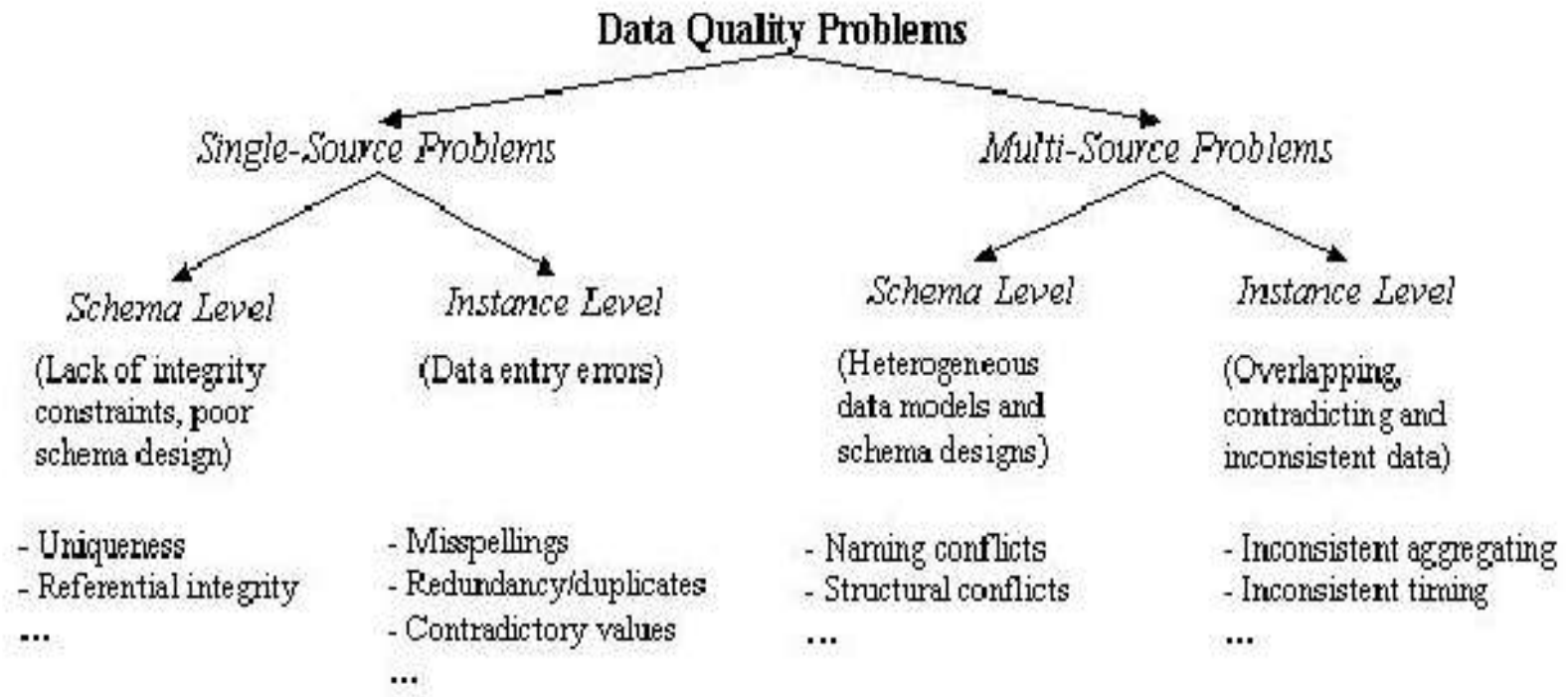
Megoldandó problémák:
- OLTP elérése és rendelkezésre állása a kívánt időben,
- engedély az olvasáshoz,
- konverzió elvégzése,
- tartalom és forma tisztítás elvégzése.

ETL-folyamatok áttekintése

Adatkiolvasás megvalósítása:
- teljes tábla (kis adatmennyiség),
- inkrementális:
    - időbélyeg alapú,
    - trigger alapú,
    - ID alapú,
- tartomány alapú (ROWID).

Minden ETL-ben szükséges utólagos ellenőrzés az átemeléshez.

Adattisztítás (data cleaning)



DSS: „garbage in garbage out”

Adattisztítás (data cleaning)

tipikus betöltési inkonzisztenciák:

- hiányos séma elem,
- hiányos adatelőfordulás,
- hibásan bevitt érték,
- téves számítások,
- duplikációk,
- eltérő formátum,
- eltérő kódolás,
- átfedő kódolás,
- integritási szabályok hiánya,
- nem összetartozó adatok,
- hiányzó kapcsolat,
- elnevezés-konfliktus,
- strukturális konfliktus.

Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Illegal values	bdate=30.13.70	values outside of domain range
Record	Violated attribute dependencies	age=22, bdate=12.02.70	age = current year - birth year should hold
Record type	Uniqueness violation	emp_1=(name="John Smith", SSN="123456"); emp_2=(name="Peter Miller", SSN="123456")	uniqueness for SSN (social security number) violated
Source	Referential integrity violation	emp=(name="John Smith", deptno=127)	referenced department (127) not defined

Table 1: Examples for single-source problems at schema level (violated integrity constraints)

Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Missing values	phone=9999-999999	unavailable values during data entry (dummy values or null)
	Misspellings	city="Liipzig"	usually typos, phonetic errors
	Cryptic values, Abbreviations	experience="B"; occupation="DB Prog."	
	Embedded values	name="J. Smith 12.02.70 New York"	multiple values entered in one attribute (e.g. in a free-form field)
	Misfielded values	city="Germany"	
Record	Violated attribute dependencies	city="Redmond", zip=77777	city and zip code should correspond
Record type	Word transpositions	$name_1$= "J. Smith", $name_2$="Miller P."	usually in a free-form field
	Duplicated records	emp_1=(name="John Smith",...); emp_2=(name="J. Smith",...)	same employee represented twice due to some data entry errors
	Contradicting records	emp_1=(name="John Smith", bdate=12.02.70); emp_2=(name="John Smith", bdate=12.12.70)	the same real world entity is described by different values
Source	Wrong references	emp=(name="John Smith", deptno=17)	referenced department (17) is defined but wrong

Table 2: Examples for single-source problems at instance level

Customer (source 1)

CID	*Name*	*Street*	*City*	*Sex*
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Client (source 2)

Cno	*LastName*	*FirstName*	*Gender*	*Address*	*Phone/Fax*
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

Customers (integrated target with cleaned data)

No	*LName*	*FName*	*Gender*	*Street*	*City*	*State*	*ZIP*	*Phone*	*Fax*	*CID*	*Cno*
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	Smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

Figure 3: Examples of multi-source problems at schema and instance level

Adattisztítási módszerek

A DW rendszer egyik legnehezebb feladata,

a séma/adat integrációval együtt hajtódik végre.

Fázisai:

- adatelemzés a lehetséges hibák felderítésére,
- transzformációs, leképzési metódusok elkészítése,
- algoritmusok ellenőrzése, validálás,
- adatok módosítása,
- tisztított adatok beépítése.

Problems	Metadata	Examples/Heuristics
Illegal values	cardinality	e.g., cardinality (gender) > 2 indicates problem
	max, min	max, min should not be outside of permissible range
	variance, deviation	variance, deviation of statistical values should not be higher than threshold
Misspellings	attribute values	sorting on values often brings misspelled values next to correct values
Missing values	null values	percentage/number of null values
	attribute values + default values	presence of default value may indicate real value is missing
Varying value representations	attribute values	comparing attribute value set of a column of one table against that of a column of another table
Duplicates	cardinality + uniqueness	attribute cardinality = # rows should hold
	attribute values	sorting values by number of occurrences; more than 1 occurrence indicates duplicates

Table 3: Examples for the use of reengineered metadata to address data quality problems

Adatelemzés két fő áramlata:

- data profiling,
- data mining.

A transzformáció
általános formátuma: SQL

Elírási hibák felderítése:

- n-gram módszer:
  gyors,
  pontatlan,

- szótár alapú:
  hash (hasító fv.),

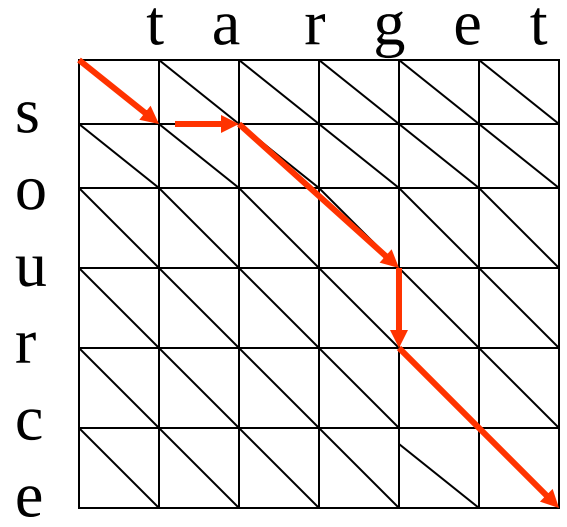
- editálási távolság (Vlagyimir Levenshtein, 1965):
  dinamikus programozás,
  lassú,
  pontos.

Minimális költség kiszámítása az editálási távolságnál

alapműveletek : insertion, deletion, substitution

$$d(s^n,t^m) = \min \begin{cases} c(s_n,t_m) + d(s^{n-1},t^{m-1}) \\ c(s_n,0) + d(s^{n-1},t^m) \\ c(0,t_m) + d(s^n,t^{m-1}) \end{cases}$$

Átalakítási mátrix



$O(n \cdot m)$

$O(n \cdot m / \log n)$

Hiányzó érték pótlása:

nem pontos, statisztikai alapú,

a többi attribútum alapján vett legvalószínűbb érték megadása.

1. attribútum-párok közötti korreláció számítása

$$korr = szumma(x_i y_i) / (szumma(x_i)\ szumma(y_i))$$

2. legszorosabb kapcsolatú attribútumok kiválasztása

3. értékek közelítése

$$d = szumma\ ((y_i - x_i)^2)$$

d → szélsőérték

Rekord illesztési módszerek:

más helyről származó rekordok illesztése (pl. biztosítottak),

nem egyeznek meg a kapcsolódó kulcsok (hiány, elírás).

módszerek:

- egy index : pontatlan, lassú,

- több index : ablak technika,

- valószínűségi : pozitív és negatív minták vizsgálatával megbecsüli az illeszkedési valószínűséget, maradnak bizonytalan esetek.