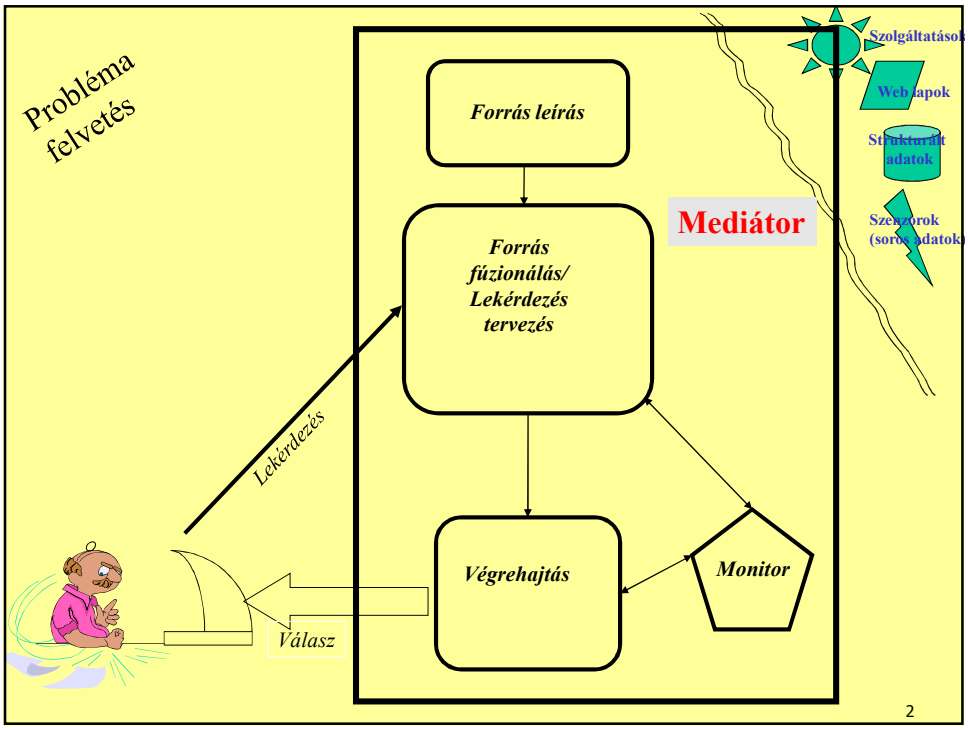
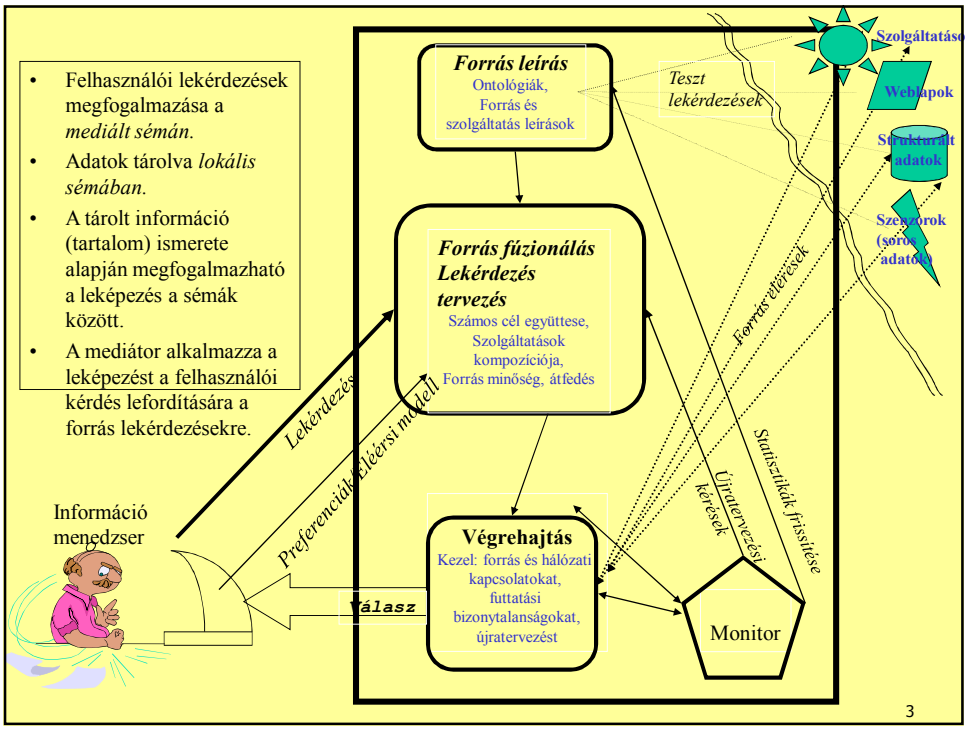
Információ integráció (Szemantikus Web megközelítés a másik irányból)

1





Miért van szükségünk ilyesmire? (Alkalmazások)

- WWW:
  - Összehasonlítás alapú vásárlás
  - Portál építések több adatforrás felhasználásával
  - B2B, elektronikus piacterek
- Tudomány és kultúra:
  - Genetika: gén információk integrálása
  - Asztrofizika: égi jelenségek gyűjtése.
  - Kultúra: kulturális információs adatbázisok egységes elérése országhatárokon túl
- Vállalati adatintegráció
  - Egy átlagos KNV 49 adatbázist alkalmaz és IT költségvetésének 30%-át az adatintegrációra költi (US)

4

Szkeptikusoknak

Csak szöveg volna a weben?

- A web jelentős része valójában strukturált…
  - A legtöbb web szerver mögött adatbázisok állnak
  - Dinamikusan konvertálják az adatokat olvasható nyelvi formára
    - <India, New Delhi> => The capital of India is New Delhi.
    - Ha vissza tudnánk konvertálni lenne strukturált adatunk!
      - » (ki)csomagolók, csomagolók tanulása, stb…
    - Dinamikus lapokat is fel tudunk deríteni...
- Félig-strukturált web (kialakulóban)
  - Legtöbb lap részben strukturált (pl. XML)
  - XML a szabvány a szintaktikára, ismert problémák az értelmezéssel
- Szolgáltatások
  - Utazási szolgáltatások, vásárlások támogatása
- Érzékelők

  Tőzsdei árfolyamok, hőmérsékletek, jegyárak…

5

Szkeptikusoknak

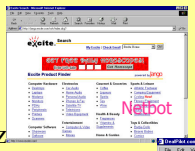
Miért nem elég:

- Keresőgépek szövegalapú keresést végeznek
  - Jól működik egyedi dokumentumokon
    - Nem tudnak integrálni több dokumentumból származó információkat
    - Nem képesek hatékony általánosításra
    - Nem tudnak dokumentumokat és adatbázisokat összekapcsolni
- Az információ integráció célja strukturált és félig-strukturált információforrások együttes kezelése

6

Összehasonlítás alapú vásárlás?

- Hasonló
- De:
  - Szélesebb fókusz
    - Szélesebb spektruma az adatbázisoknak
    - Szolgáltatások
  - Új kihívás
    - "adattár" nem működik
    - Kézi forrásleírás, kezelés korlátai

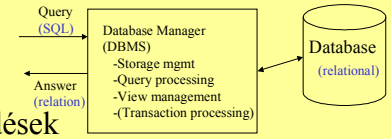


7

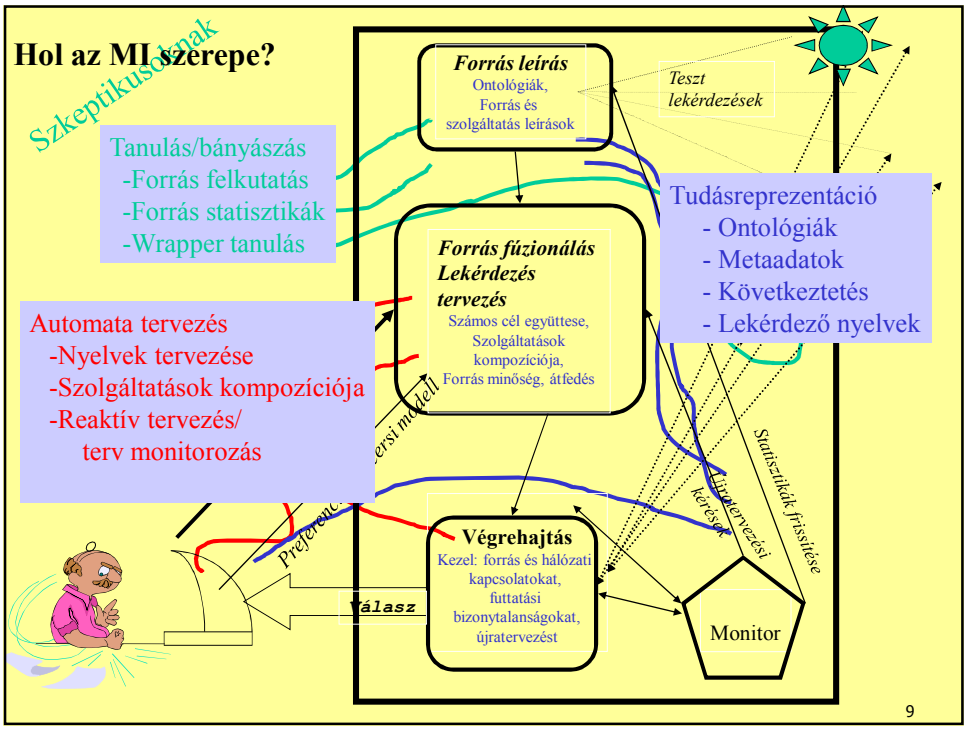
Szkeptikusoknak

Miért nem csak adatbázisok elosztott adatbázisok

- Közös séma hiánya
  - Források heterogén sémákkal (és fogalmakkal, ontológiákkal) rendelkeznek
  - Félig-strukturált források
- Régi források
  - Nem relációs sémák
  - Eltérő elérési módok
- Független források
  - Nincs közös adminisztráció
  - Nem kezelt forrás tartalmi átfedések
- Nehezen előrejelezhető viselkedés
  - Lekérdezés végrehajtás bonyolult
- Általában csak olvashatóak
  - Ez lehet szerencsés is
  - Bár terjednek a tranzakció kezelési megoldások a weben

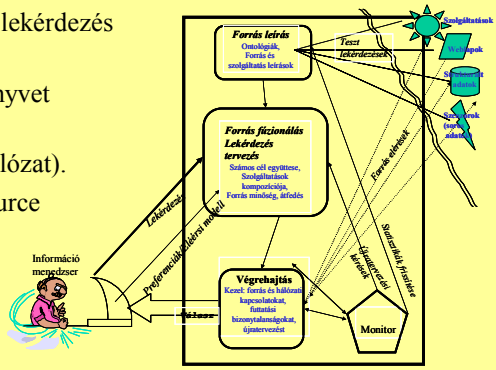


8



Forrás leírások

- Minden meta-adat információt tartalmaz
  - Forrás tartalom logikai leírása (könyvek, új autók).
  - Forrás képességek (pl. SQL lekérdezés feltehető)
  - Forrás teljesség (*minden* könyvet tartalmaz).
  - Fizikai jellemzők (forrás, hálózat).
  - Statisztikák az adatokról Source reliability
  - Tükör források
  - Frissítési frekvencia.



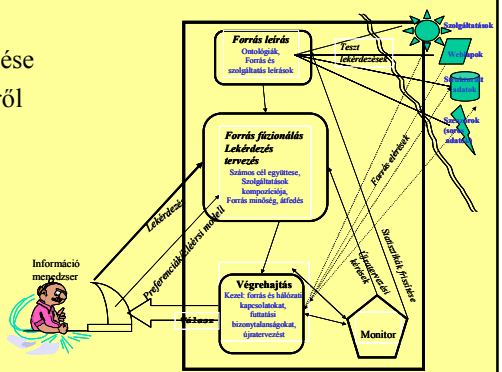
Forrás elérések

- Hogyan kapunk n-eseket
  - Számos forrás strukturálatlan adatokat ad
    - Néhány inherensen strukturálatlan, mások természetes nyelvi köntösben vannak
  - Vissza kell csomagolni az adatokat
  - Wrapper építés/információ kinyerés
  - Kézi munka/fél-automatikus

11

Forrás fúzió/ lekérdezés tervezés

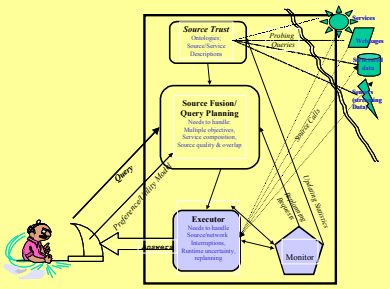
- Feldolgozza a felhasználói lekérdezést és előállítja a végrehajtási tervet
  - Költség és hatékonyság közti optimalizáció
  - Forrás elérési korlátok kezelése
  - Információ a forrásminőségről



12

Monitoring/ Végrehajtás

- Lekérdezési terv alapján elvégzi a feladatot a forrásokon
  - Forrás késleltetések kezelése
  - Hálózati, tranziens kimaradások
  - Forrás elérési korlátok
  - Szükséges lehet újratervezések elvégzése



Méretek figyelembe vétele

- Hány forrást kell elérni?
- Mennyire autonómok ezek?
- Van ismeretünk a forrásokról?
- Strukturáltak az adatok?
- Csak lekérdezés lehetséges vagy módosítás is?
- Követelmények: pontosság, teljesség, teljesítmény, inkonzisztenciák kezelése
- Zárt vagy nyílt világ feltételezés?

Deduktív adatbázisok

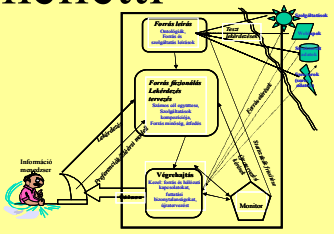
- Relációkat predikátumokkal írjuk le.
- Relációk közti relációkat datalog szabályokkal írjuk le
  - (Horn klózok, függvényszimbólumok nélkül)
  - Lekérdezések megfelelnek egy datalog programnak

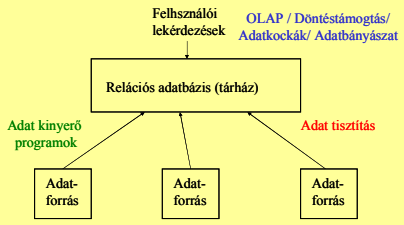
```
Emprelated(Name,Dname) :-  Empdep(Name,Dname)

Emprelated(Name,Dname) :- Empdep(Name,D1), Emprelated(D1,Dname)
```

15

Kis forrás szám melletti integráció

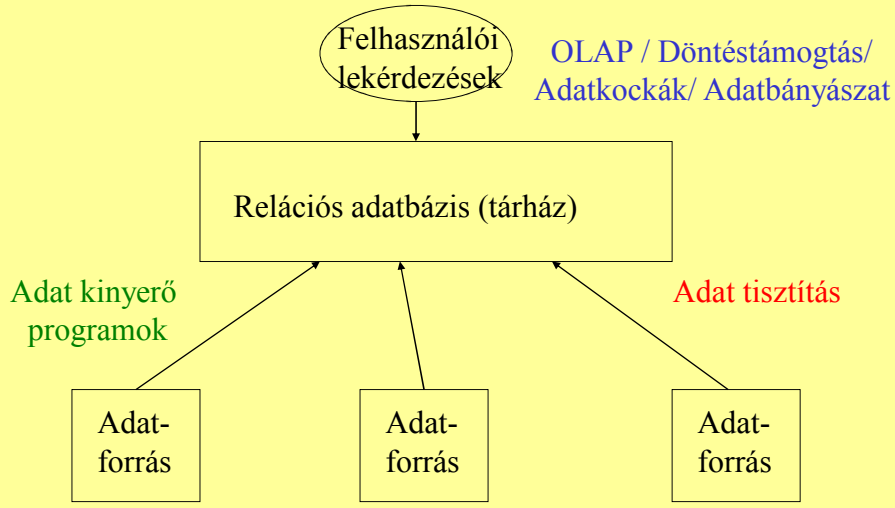
- Általában ad-hoc programozás: speciális eset megvalósítása minden esetre, sok konzultáció.
- Adattárházak: minden adat periódikus feltöltése az adattárházba.
  - 6-18 hónap bevezetési idő
  - Operációs és döntéstámogatási RDBMS elválasztás. (nem csak adatintegrációra megoldás).
  - Teljesítmény jó,
  - adat lehet, hogy nem friss;
  - Rendszeres adattisztítás szükséges.





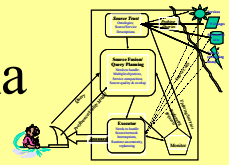
16

Integrátor séma

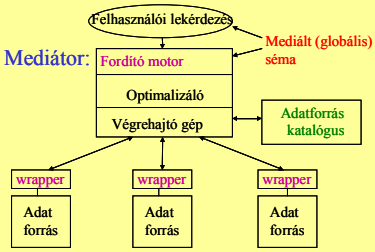


17

Virtuális integrációs séma



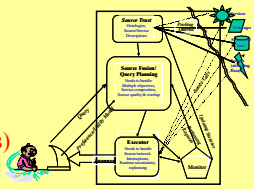
- Adatok a forrásokban maradnak
- Lekérdezés végrehajtásakor:
  - Releváns források meghatározása
  - Lekérdezés szétválasztása forrásokra vonatkozó lekérdezésekre.
  - Válaszok begyűjtése a forrásokból, és megfelelő kombinálása a válasz előállításához.
- Friss adatok
- A megoldás skálázható

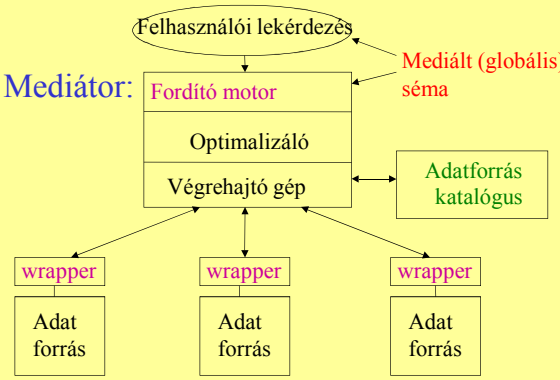


Garlic [IBM], Hermes[UMD];Tsimmis, InfoMaster[Stanford]; DISCO[INRIA]; Information Manifold [AT&T]; SIMS/Ariadne[USC];Emerac/Havasu[ASU]

18

Virtuális integrátor architektúra





Források: relációs adatbázisok, weblapok, szövegek.

19

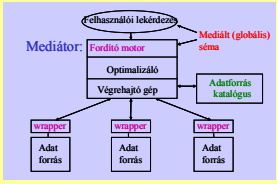
Projektek

- Garlic (IBM),
- Information Manifold (AT&T)
- Tsimmis, InfoMaster (Stanford)
- The Internet Softbot/Razor/Tukwila (UW)
- Hermes (Maryland)
- DISCO (INRIA, France)
- SIMS/Ariadne (USC/ISI)
- Emerac/Havasu (ASU)
- BibFinder (ASU)

20

Forrás-mediátor relációs sémával szembeni elvárások

- Kifejező erő: hasonló adattartalommal rendelkező források megkülönböztetése, irreleváns források felismerése.
- Egyszerű bővíthetőség: tegyük könnyűvé források hozzáadását.
- Fordítás/átalakítás: felhasználói lekérdezés lefordítása forrásokon értelmezett lekérdezésekre hatékonyan és eredményesen.
- Vesztességmentesség: minden lehetséges adatelérés biztosítása



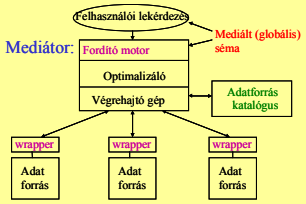
Lekérdezés átalakítás

- Adott:
  - Egy Q lekérdezés a mediátor sémára vonatkozóan
  - Adat források leírása
- Létrehozandó:
  - Egy Q' lekérdezés az adat forrásokra vonatkozóan, amely:
    - Q' csak *helyes válaszokat* ad a Q lekérdezéshez és
    - Q' minden lehetséges választ megtalál Q-hoz az elérhető forrásokból.

21

Fordítási/átfogalmazási probléma

- Adott:
  - Egy Q lekérdezés a mediátor sémára vonatkozóan
  - Adat források leírása
- Létrehozandó:
  - Egy Q' lekérdezés az adat forrásokra vonatkozóan, amely:
    - Q' csak *helyes válaszokat* ad a Q lekérdezéshez és
    - Q' minden lehetséges választ megtalál Q-hoz az elérhető forrásokból.



22

Forrás és felhasználói sémák reláció leírásának megközelítései

- Globális mediált sémák(Global-as-view, GAV): a mediált séma kifejezése a forrásokra vonatkozó nézetek relációjaként
- Lokális mediált sémák (Local-as-view, LAV): forrás relációk kifejezése a mediált sémákon értelmezett relációkkal.
- Módszerek kombinációja…?

"Nézet" frissítés

```
CREATE VIEW Seattle-view AS

    SELECT buyer, seller, product, store
    FROM   Person, Purchase
    WHERE  Person.city = "Seattle"    AND
           Person.name = Purchase.buyer

A nézet felhasználása:

    SELECT  name, store
    FROM    Seattle-view, Product
    WHERE   Seattle-view.product = Product.name  AND
            Product.category = "shoes"
```

Hasonlítsuk össze őket egy film adatbázis lekérdezésénél

23

Mintapélda

- Egy mediátor egy film adatbázishoz
  - Információk szolgáltatása filmekről, illetve mozi programról

24

Globális mediált nézet
GAV (Global-As-View)

Mediált séma kifejezése a forrásokra vonatkozó nézetek relációjaként.

Mediált/felhasználói séma:
Filmek(cím, rendező, év, típus),
Műsor(mozi, cím, idő).

```
Create View Filmek AS
  select *  from S1     [S1(cím, rendező, év, típus)]
  union
  select  * from S2     [S2(cím, rendező, év, típus)]
  union                 [S3(cím,rendező),
   S4(cím,év,típus)]
  select S3.cím, S3.rendező, S4.év, S4.típus
  from  S3, S4
  where S3.cím=S4.cím
```

25

GAV

Mediált séma kifejezése a forrásokra vonatkozó nézetek relációjaként.

Mediált/felhasználói séma:
Filmek(cím, rendező, év, típus),
Műsor(mozi, cím, idő).

```
Create View Filmek AS
  select *  from S1     [S1(cím, rendező, év, típus)]
  union
  select  * from S2     [S2(cím, rendező, év, típus)]
  union                 [S3(cím,rendező), S4(cím,év,típus)]
  select S3.cím, S3.rendező, S4.év, S4.típus
  from  S3, S4
  where S3.cím=S4.cím
```

A mediátor séma relációk virtuális nézetek a forrásrelációkon.

26

GAV: példa 2.

Mediált/felhasználói séma:
Filmek(cím, rendező, év, típus),
Műsor(mozi, cím, idő).

Mediált séma kifejezése a forrásokra vonatkozó nézetek relációjaként.

```
Create View Filmek AS
  select *  from S1                    [S1(cím,rendező,év)]
  select cím, rendező, év, NULL
  from S1
  union                                [S2(cím, rendező,típus)]
  select  cím, rendező, NULL, típus
  from S2
```

Null értékek

27

GAV: példa 2.

Mediált/felhasználói séma:
Filmek(cím, rendező, év, típus),
Műsor(mozi, cím, idő).

Mediált séma kifejezése a forrásokra vonatkozó nézetek relációjaként.

Forrás S4: S4(mozi, típus)

```
Create View Filmek AS
  select NULL, NULL, NULL, típus
  from S4
Create View Műsor AS
  select mozi, NULL, NULL
  from S4.
```

De mit lehetne tenni, ha minket a vígjátékokat játszó mozik érdekelnének?

"Veszteséges medáció"

28

LAV: példa 1

Mediált/felhasználói séma:

Filmek(cím, rendező, év, típus),
Műsor(mozi, cím, idő).

Forrás séma kifejezése a mediált nézeteken értelmezett relációkként.

```
Create Source S1 AS
  select * from Filmek
```

S1(cím, rendező, év, típus)

```
Create Source S3 AS
  select cím, rendező from Filmek
```

S3(cím, rendező)

```
Create Source S5 AS
  select cím, rendező, év
  from Filmek
  where év > 1960 AND típus="vígjáték"
```

S5(cím, rendező, év), év >1960

A források "materializált nézetek" a mediált sémák felett.

29

LAV: példa 1

Mediált/felhasználói séma:

Filmek(cím, rendező, év, típus),
Műsor(mozi, cím, idő).

Forrás séma kifejezése a mediált nézeteken értelmezett relációkként.

```
Create Source S4 AS
 select mozi, típus
   from Filmek m, Műsor s
   where m.cím=s.cím
```

S4(Mozi,Típus)

Van remény a vígjátékokat játszó mozik felderítésére!

30

GAV vs. LAV

Mediált séma:
Filmek(cím, rendező, év, típus),
Műsor(mozi, cím, idő).

Forrás S4: S4(mozi, típus)

```
Create View Filmek AS
  select NULL, NULL, NULL, típus
  from S4
Create View Műsor AS
  select mozi, NULL, NULL
  from S4.
```

De mit lehetne tenni, ha minket a vígjátékokat játszó mozik érdekelnének?

Veszteséges mediáció

```
Create Source S4 AS
 select mozi, típus
   from Filmek m, Műsor s
   where m.cím=s.cím
```

31

GAV vs. LAV

- Nem moduláris
  - Források hozzáadása módosítja a meglévő mediált séma definícióját
- Nehézkes lehet veszteségmentes mediátort készíteni.
- Lekérdezés átalakítás egyszerű
  - *Nézetek kibontását jelenti (polinomiális)*
  - Hierarchikus mediátor sémák létrehozása lehetséges
- Hatékony, ha
  - Kis számú, ritkán változó adatforrás van
  - Feladat teljesen ismert a mediátor tervezésekor (pl. vállalati adatintegráció)
    - Garlic, TSIMMIS, HERMES

- Moduláris—új forrás hozzáadása egyszerű
- Igen rugalmas – a lekérdező nyelv közvetlenül alkalmazható a források leírására
- Lekérdezés átalakítás bonyolult
  - Válaszokat a nézeteken keresztül kell előállítani (nem mindig megoldható)
- Hatékony, ha
  - Sok, kevéssé korrelált forrás
  - Források dinamikus hozzáadása és törlése
    - Information Manifold, InfoMaster, Emerac, Havasu

32