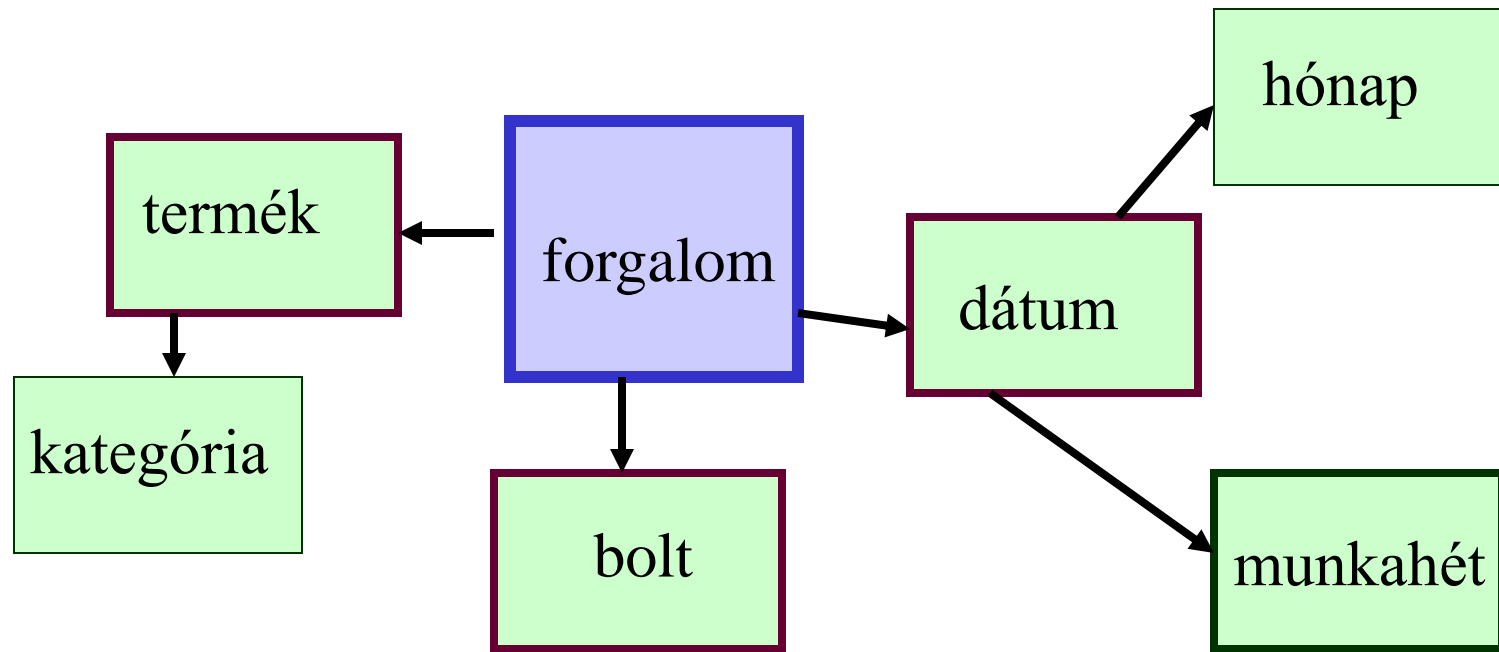


DW

5. előadás

MD adatmodell műveletei

# MD szemantikai séma modell



Hogyan modellezzük egzaktabb módon?

nincs egységes modell



# Cabibbo- Torlone szemantikai modell

A modell elemei:

- dimenzió-sémák,
- fogalom-sémák,
- adatkocka-sémák,
- adatbázis-sémák.

Grafikus jelölésrendszert is alkalmaz

Egyszerűsített megközelítés

ER-jellegű leírás

# Cabibbo-Torlone modell

## Grafikus ábrázolás



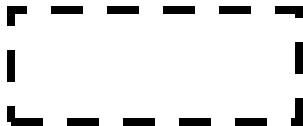
dimenzió



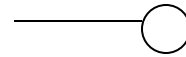
dimenzió-szint



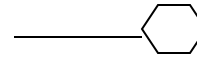
változó



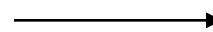
kocka



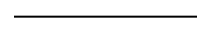
mező



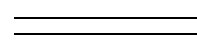
tulajdonság



dimenzió hierarchia



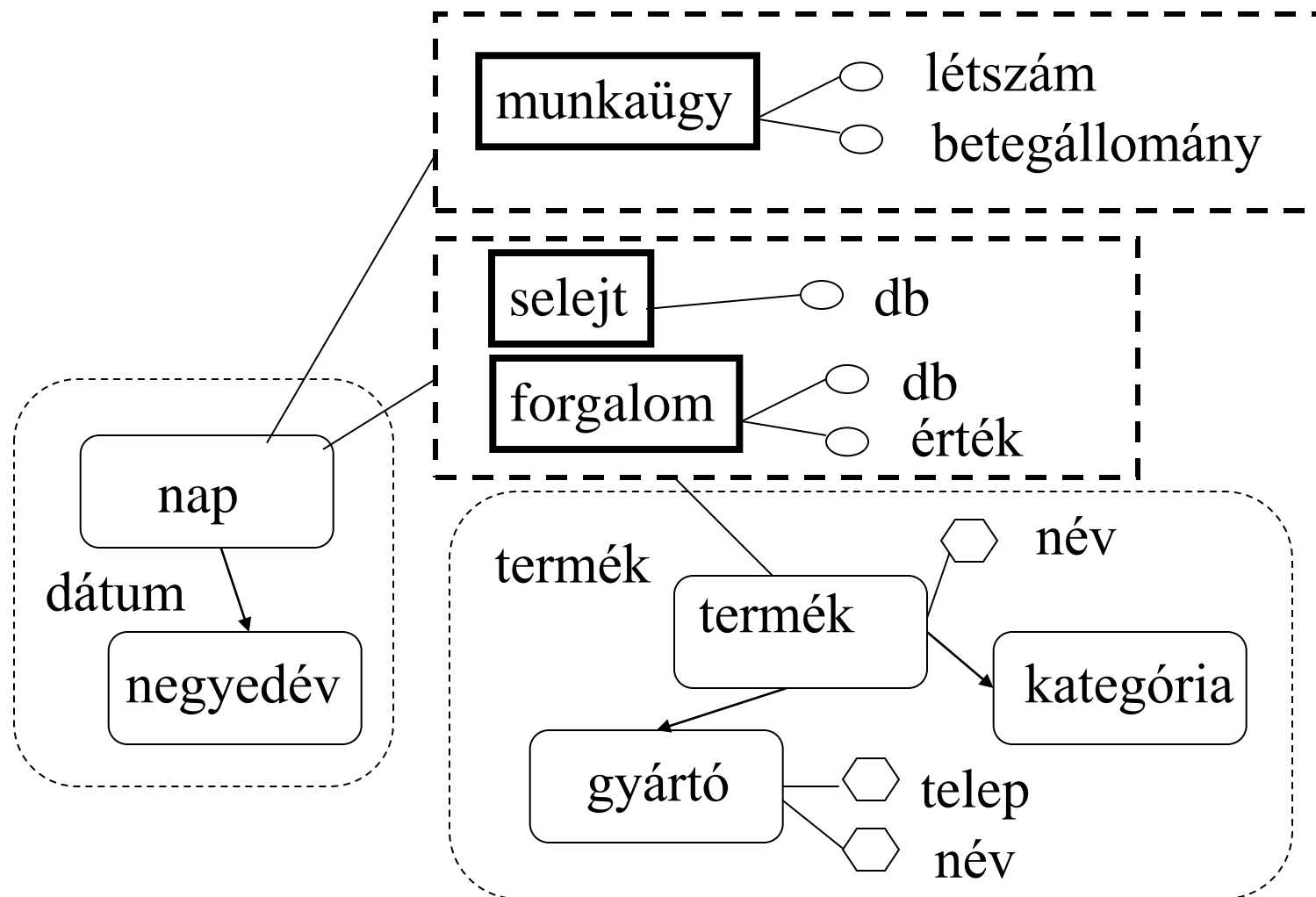
tény-dimenzió kapcsolat



többértékű dimenzió

nem OO alapú

# Cabibbo-Torlone modell



# Cabibbo-Torlone modell

univerzum :  $U$

fogalom-nevek halmaza:  $C = \{c\} \subseteq U$

típusok halmaza :  $T \subseteq U$

dimenzió-sémák halmaza:  $D = \{d, R\}$

dimenzió-séma:  $d = (L, <, P)$

$L$  dimenzió szintek ( $L \subseteq C$ )

$<$  : szint hierarchia ( $< \subseteq L \times L$ )

$P$ : tulajdonság ( $P : L \rightarrow 2^{(C \times T)}$ )

dimenzió-bázisok :  $R$

$r(d) = \{h \in L(d) \mid \nexists h' : h < h'\}$

$R = \{r(d)\}$

# Cabibbo-Torlone modell

ténytábla-sémák halmaza:  $F = \{f\}$

ténytábla-séma:  $f = \{(A_1:I_1, \dots, A_n:I_n), I_0\}$

$$f \in C$$

$$A_i \in C$$

$$I_i \in r(R)$$

$$I_0 \subseteq 2^{(C \times T)}$$

adatbázis-séma:  $DB =$

$(C, T,$

$D\{d(L, <, P), R\},$

$F\{f(A:I):I_0\})$

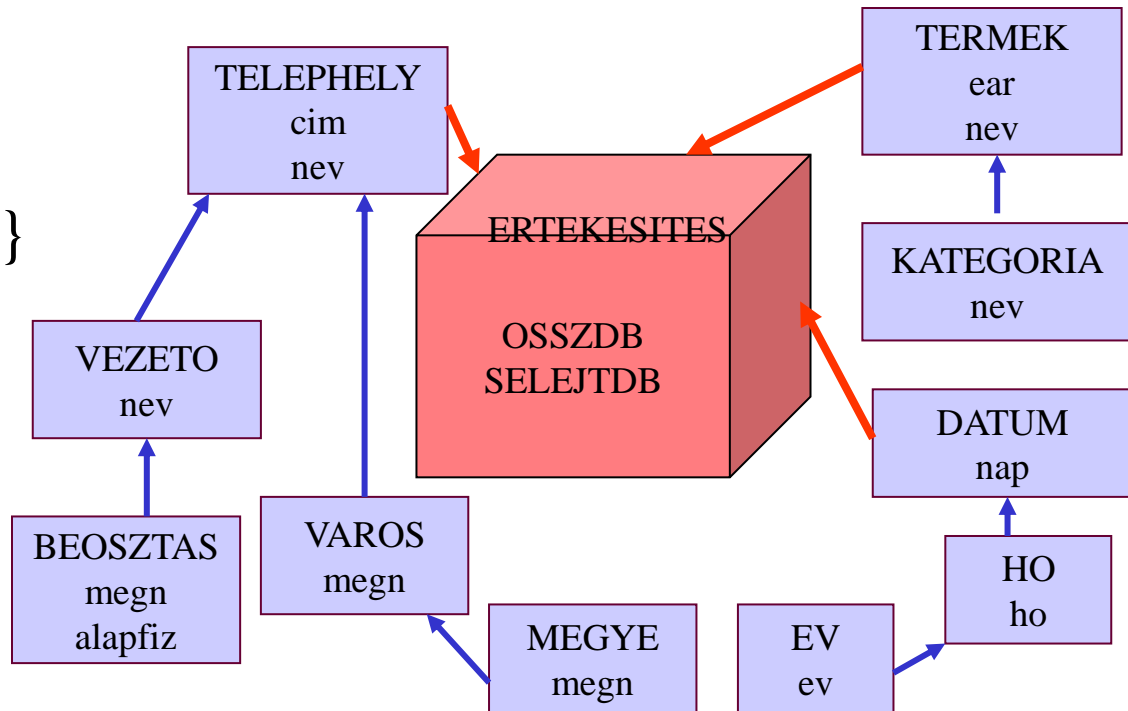
# Minta séma felírása

$C = \{\text{alapfiz, beosztas, ear, ertekesites, ev, cím, osszdb, selejtodb, telephely, termék, ...}\}$

$T = \{\text{int, char}\}$

$D = \{d_{\text{termek}}, d_{\text{datum}}, d_{\text{telep}}, R\}$

$d_{\text{termek}} = \{$   
   $\{\text{termek, kategoria}\},$   
   $\{\text{kategoria} < \text{termek}\},$   
   $\{\text{termek} \rightarrow (\text{ear, int}),$   
   $\text{termek} \rightarrow (\text{nev, char}),$   
   $\text{kategori} \rightarrow (\text{nev, char})\}$   
   $\}$



...



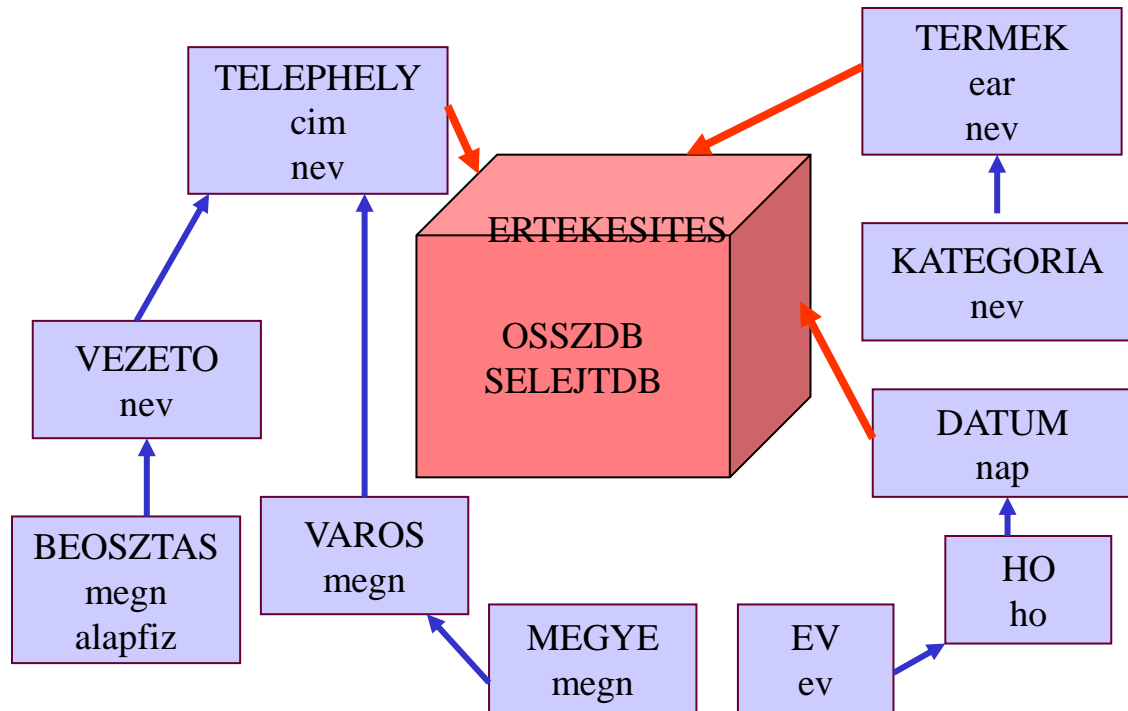
# Minta séma felírása

$R = \{\text{termek, datum, telephely}\}$

$F = \{\text{ertekesites}\}$

ertekesites = {  
  (termelte:telephely,  
  termek:termek,  
  datum:datum),  
  {(osszdb,int)  
  (selejtdb,int)}  
}

$DB = \{C,T,D,F\}$



# Cabibbo-Torlone modell

Formális felírás előnyei:

- egyértelműbb jelentés,
- könnyebben feldolgozható,
- tömörebb leírás,
- könnyebben konvertálható,
- alapot adhat további egyértelmű bővítésekre.

néhány további modell:

- Li-Wang(1996),
- Agrawal-Gupta-Sarawagi(1997),
- ADAPT,
- ...

# MD műveletek

Felhasználó igény:

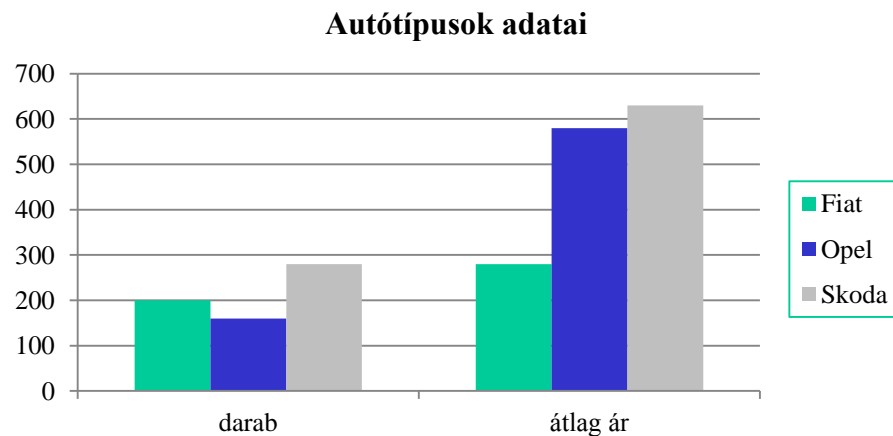
- ad-hoc lekérdezések a kocka különböző szeleteire,
- táblázatos megjelenítés (többdimenziós eredmény),
- áttekinthető megjelenítés.

‘termelés alakulása a keleti régióra vonatkozóan az elmúlt három hónapra vonatkoztatva..’



termékek

dátumok

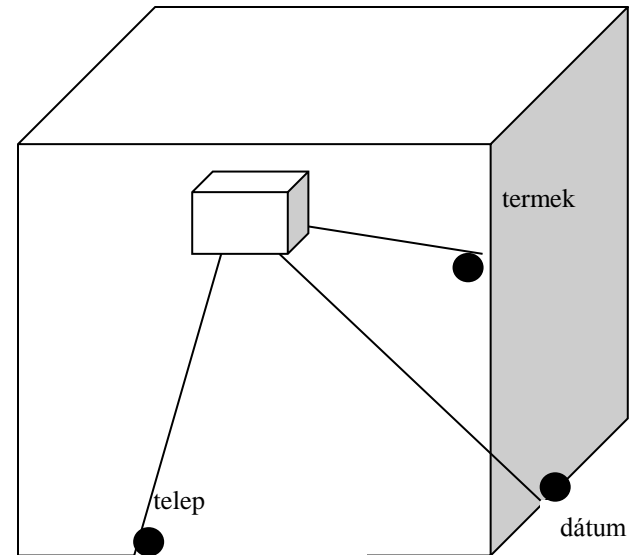



# MD műveletek

Itt sincs egységes modell

gyakorlatban elterjedt műveletek:

- szűrés,
- aggregálás,
- szintváltás,
- összekapcsolás,
- kibontás,
- ...



mintaként a relációs algebra jöhet szóba

adatkockán értelmezett, adatkockát előállító operátorok

# MD műveletek

## Szelekció (szűkítés, slice and dice) - szigma

- változó szelekció  $\sigma_{f(v)}(F)$

A feltételnek eleget tévő cellák maradnak meg,  
a többi cella NULL értékű lesz.

- attributum szelekció  $\sigma_{f(D.a)}(F)$

A feltételnek eleget tévő dimenzió értékek maradnak  
meg, a többi kikerül a kockából.

$\sigma_{\text{selejtdb} > \text{osszdb} * 0.2}$  (ertekekesites)

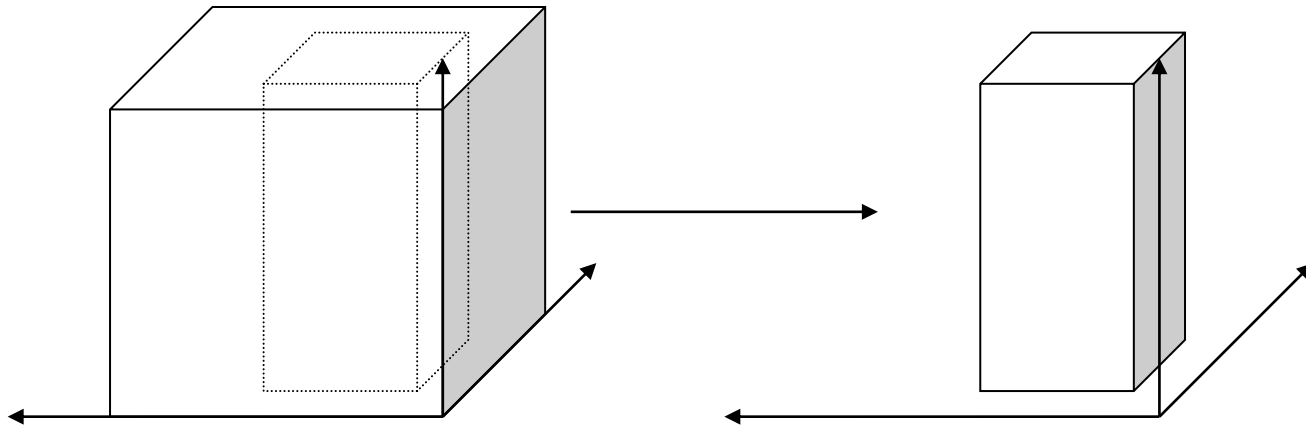
$\sigma_{\text{selejtdb} > \text{osszdb} * 0.2}$  (ertekekesites)

## Szűkítés, slice and dice

A vásárlások adatkockán több dimenziót is érintő feltételt adunk meg.

Az alkalmazott szelekciós feltétel:

a termék legyen labda és a dátum pedig a nyári hónapokat ölelje át.



# MD műveletek

## Projekció (slice and dice) - pí

- változó projekció  $\pi_v(F)$

A kijelölt változók maradnak meg a cellában.

- attributum projekció  $\pi_{D.a}(F)$

A kijelölt attributumok maradnak meg a dimenziónál.

$\pi_{\text{selejt db}}(\text{ertesites})$

$\pi_{\text{selejt db}}(\sigma_{\text{selejt db} > \text{ossz db} * 0.2}(\text{ertesites}))$

## Slicing

Néha a **szeletelés** alatt egy dimenzió mentén egy érték kiválasztását értenek, így eggyel kevesebb számú dimenziós kockát létrehozva. (Attribútum szelekció egy értéket kiválasztva.)  
Például a termékek közül csak a labdát kiválasztva, de a dátum és boltok dimenziók változatlanok maradnak.

## Dicing

A **kockára vágás** alatt néha az adatkockán több dimenziót is érintő szelekciót értenek. Tehát a hagyományos szelekció.



## MD műveletek

### **Dimenzió összevonása (roll up, drill up) - nő**

- attributum szintű

$v_{D1.D2}(F)$

A megadott dimenzióból a megadott (durvább) dimenzióba való áttérés a dimenzió hierarchia mentén, vagy egy dimenzió attribútum kivételét a táblázatból az aktuális dimenzió szintjén maradva.

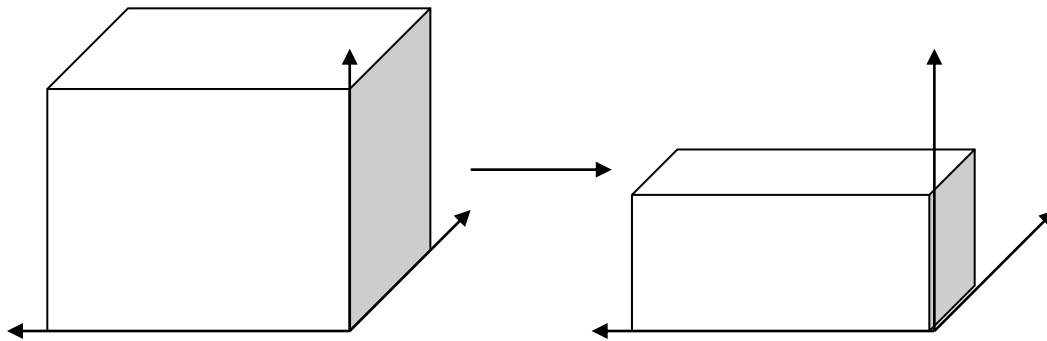
aggregáltabb adatokra való áttérés  
megváltozik a szint a megadott dimenziónál  
durvább felbontást kapunk

$v_{\text{termek.kategoria}}$  (ertekekesites)

## Összevonás, drill up

A vásárlások adatkockán eddig napi bontásban és boltonként láthattuk a forgalom adatokat, akkor az összevonás révén át lehet térni havi idő bontásra, ahol az adatkockában az adott hónap dimenzióértéknél az oda tartozó napi adatok összesített értéke fog megjelenni.

Hasonlóan elérhető az is, hogy a vevő dimenzió tengely mentén kevesebb tulajdonság szerepeljen az egyes vevők adatai között.



## MD műveletek

### **Dimenzió kibontása (finomítás, drill down) - kappa**

- attributum szintű

$\kappa_D(F)$

A megadott attributum részletező dimenziójára való áttérés a dimenzió hierarchia mentén, vagy egy újabb attribútumot hozunk be a táblázatba az aktuális dimenzió szintjén.

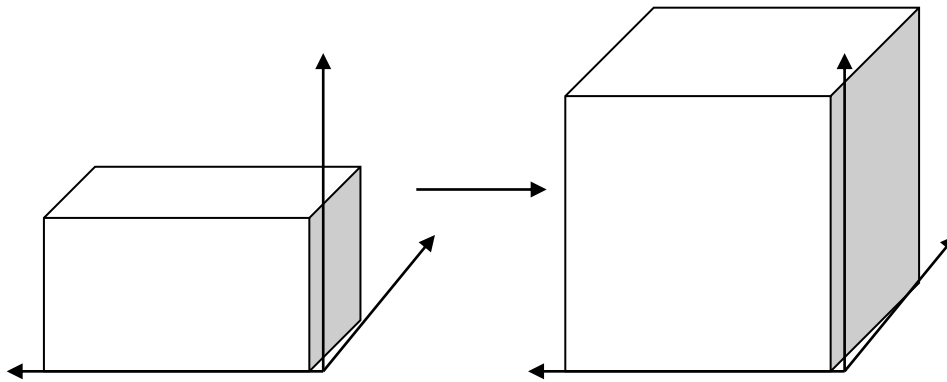
részletezőbb adatokra való áttérés  
megváltozik a szint az aktuális dimenziónál  
finomabb felbontást kapunk

$\kappa_{\text{varos}}$  (ertekekesites)

## Finomítás, drill down

A vásárlások adatkockán az adatok most éppen havi bontásban szerepelnek, a finomítással elérhető többek között az, hogy:

- ne havi bontásban jelenjenek meg az adatok, hanem napi bontásban (egy finomabb dimenzió hierarchia szintre térünk át),
- a vevőknél ne csak a neve szerepeljen hanem az életkora is megjelenjen (a meglévő dimenzió hierarchia szinten maradunk, de egy újabb dimenzió tulajdonságot hozunk be az eredménybe).



## MD műveletek

### Aggregáció (bevonás, fold) - $\phi$

- dimenzió szintű  $\phi_{D, \text{aggr}}(F)$

A megadott dimenziók maradnak meg, az összevont cellák tartalmából az aggr aggregáció alapján képződik az eredő cella.

A dimenziók száma csökken, s a megszüntetett dimenzió értékei bekerülnek a ténycellába mezőként. Erre akkor kerülhet sor, ha valamely dimenzió átmenetileg érdektelenné válik, a tőle való függőségi viszonyt nem fogjuk vizsgálni a közeljövőben.

összesítőbb adatokra való áttérés

szűkül a dimenzió készlet, csökken a dimenziószám

durvább felbontást kapunk

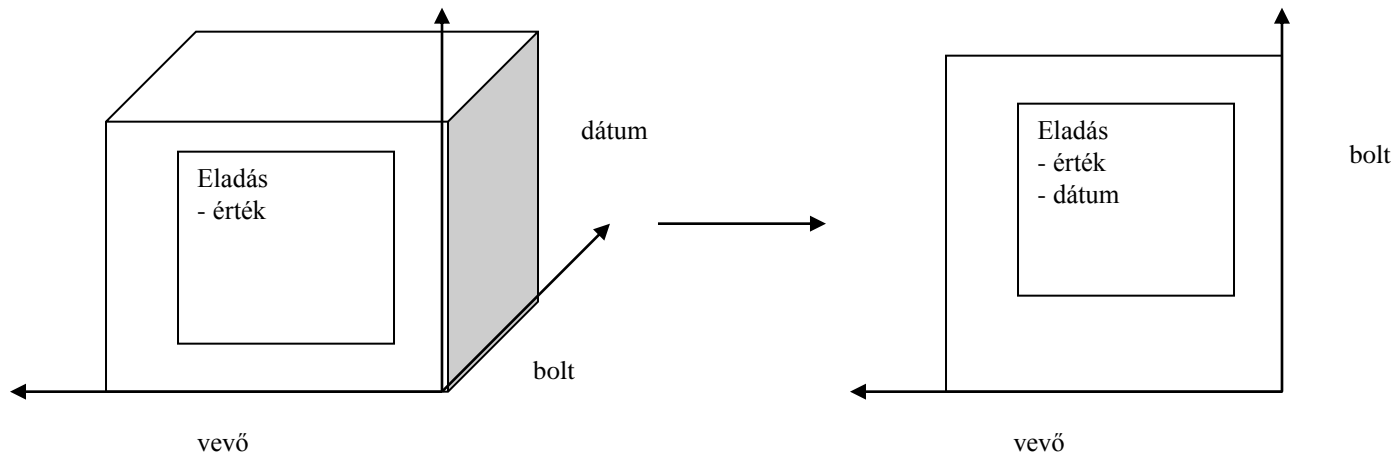
$\phi$  termék, termelte, sum() (ertekeşites)

# Bevonás, fold

Az adatkockában a dátum elemet kivesszük a dimenziók közül, s át tesszük a ténycella mezője közé.

A művelet hatására:

- eggyel csökken a dimenziók száma, és
- eggyel nő a ténycella mezőinek a darabszáma.



# MD műveletek

## **Kiterítés** (kibontás, unfold) - delta

- dimenzió szintű

$\delta_D(F)$

Behozza a kockába a megadott dimenziót (ha lehet).  
A ténystruktúra megadott mezője átmegy dimenzióba,  
így meghatározhatjuk a vizsgált tényadatnak  
az egyik korábbi mezőjétől való függését.

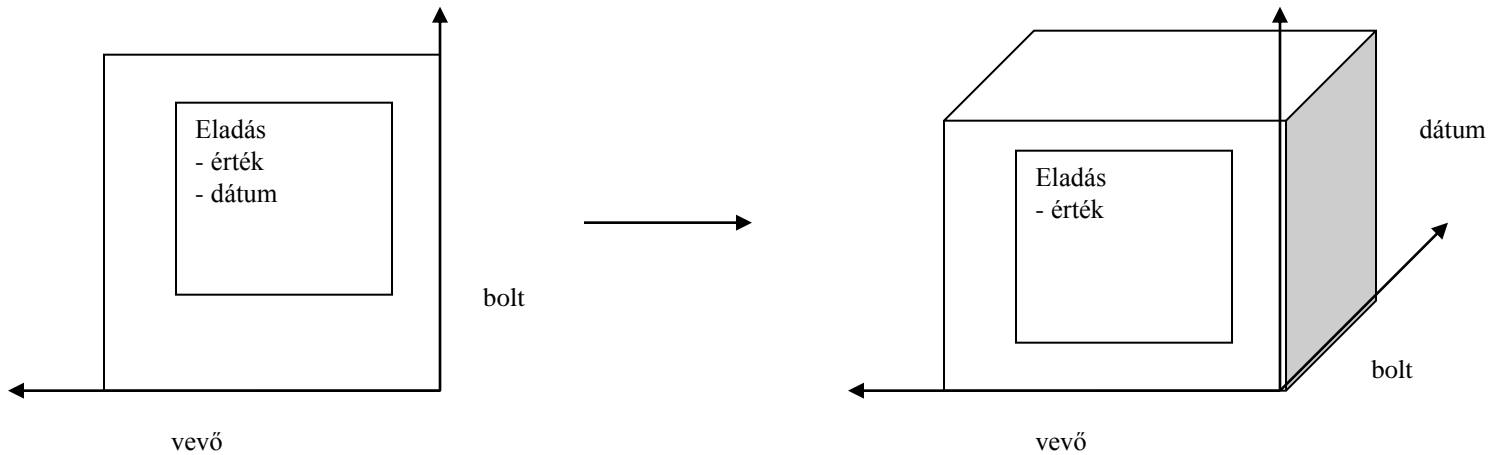
részletezőbb adatokra való áttérés

bővül a dimenzió készlet, növekszik a dimenziószám  
finomabb felbontást kapunk

$\delta_{\text{termek}}$  (ertekecsites)

# Kibontás, unfold

Az adatkockában már lekérdezhető az eladási adatoknak a dátumtól való függése.





## MD műveletek

### Szorzás (összekapcsolás, natural join, drill across)

- adatkocka szintű  $F_1 \times F_2$   
A megadott adatkockákból olyan eredő adatkocka készítése (legyenek közös dimenziók), melyre
  - dimenzióhalmaza az  $F_1$  és  $F_2$  dimenzió halmazának uniója,
  - változólistája a két lista összevonása, párosa,
  - változó értékei a megfelelő koordinátájú értékek párosa.

ertekeites  $\times$  rendeles

# MD műveletek

		$F_1$	
	d24	2	
	d23	1	2
D2	d22	3	
	d21	5	
		d11	d12
		D1	

		$F_2$	
	d26		a
	d24	a	
D2	d23		b
	d21	c	
		d31	d32
		D3	

$$F_1 \times F_2$$

$$(d11, d23, d31) = (1, \text{null})$$

$$(d11, d24, d31) = (2, a)$$

$$(d11, d26, d31) = (\text{null}, \text{null})$$

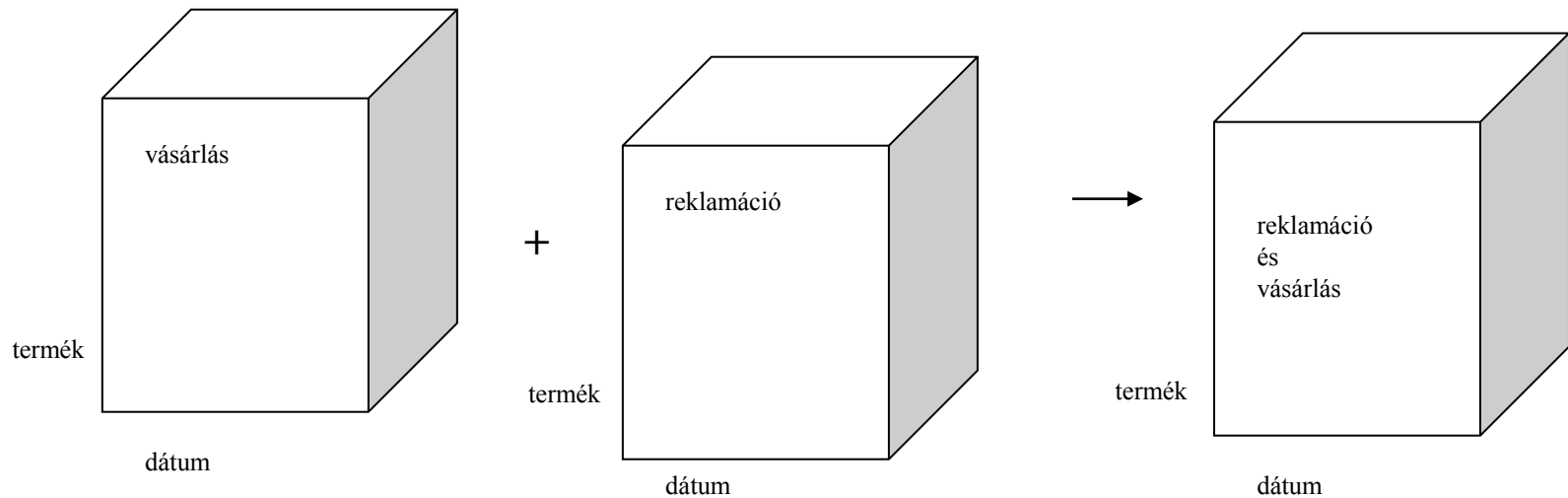
...

# Összekapcsolás, drill across

A vásárlás és reklamáció adatkockák összekapcsolása.

Dátum	Vevő	Termék	Érték	Db		Dátum	Termék	Hiba	Érték	...
10.02	45	786	155	2		10.12	625	77	144	
10.13	67	678	65	1		10.13	678	76	12	
10.16	87	78	83	3		10.16	78	99	1	

Dátum	Vevő	Termék	Eérték	Db	Hiba	HÉrték
10.02	45	786	155	2		
10.13	67	678	65	1	76	12
10.16	87	78	83	3	99	1



## MD műveletek

### **Dimenzió összevonás (cojoin) - kái**

$$\chi_{d_1, d_2}(F)$$

Két dimenzióból egyet állít elő, melyben minden értékpáros előfordul.

A dimenziószám eggyel változik,  
a cellák száma változatlan.

D1 (a,b,c)

D2 (1, 2)

=>

D12(a1,a2,b1,b2,c1,c2)

Ilyen lesz az MDX-ben is, és a PE cojoint dimension is ilyen.

# MD műveletek

## Kiterjesztés (extension) - alfa

- változó szintű  $\alpha_v(F)$   
Új változó hozzáadása, melynek értéke a meglévő változókból származtatható.

## Pivotálás - tau

- adatkocka szintű  $\tau_{D1,D2,aggr}(F)$   
Áttérés aggregációk sorozatával kétdimenziós adatkockára (listára).  
A cellában szereplő értékekből oszlopok/sorok lesznek.

$\alpha_{\text{selejt db} / \text{ossz db}}$  (erteke sites)

$\tau_{\text{termek, datum}}$  (erteke sites)

# Pivotálás

A vásárlások adatkocka celláinak az egyes sorok felelnek meg.

Az oszlopok a tény tagok, illetve a kapcsolódó dimenzió attribútumok.

Dátum	Vevő	Termék	EÉrték	Db	Hiba	HÉrték
10.02	45	786	155	2		
10.13	67	678	65	1	76	12
10.16	45	781	83	3	99	1
10.21	45	786	236	3		
11.10	67	786	287	1		

A tulajdonságként szereplő vevő- és termékkódok oszlopként, sorként fognak szerepelni a művelet után, azaz külön oszlopba vagy sorba kerülnek az egyes kódok. A cellába az adott kódokhoz tartozó összesített eladási érték (EÉrték) fog bekerülni.

	Vevő	45	67	...	...	...
Termék						
678		0	65	...	...	...
781		83	0	...	...	...
786		391	287	...	...	...
....						

# MD műveletek

## Forgatás (rotation) - $\rho$

- dimenzió szintű

$$\rho_{d_1, d_2}(F)$$

A megadott dimenziók helyet cserélnek egymással.

(Egyes esetekben a pivotálás alatt is forgatást értenek.)

részletezés szintje változatlan  
dimenziószám nem változik  
más elrendezést biztosít  
ábrázoláshoz fontos

$\rho_{\text{termek, bolt}}$  (ertekekesites)

# MD műveletek

## Rendezés (ordering) - omikron

- dimenzió szintű

$O_{d,kif}(F)$

A megadott dimenzió előfordulásainak rendezése.

részletezés szintje változatlan  
dimenziószám nem változik  
más elrendezést biztosít  
ábrázoláshoz fontos

$O_{\text{termek, nev}}$  (erteke sites)



# MD műveletek

## Mintapéldák

- Az  $x$ -nél olcsóbb termékekre vonatkozó adatok

$$\sigma_{\text{TERMEK.ear} > x} (\text{ERTEKESITES})$$

- Az  $x$ -nél rosszabb selejtarányú eladásra vonatkozó adatok:

$$\sigma_{\text{selejtdb/osszdb} > x} (\text{ERTEKESITES})$$

- Adatkocka a selejtarányok nélkül

$$\pi_{\text{osszdb}} (\text{ERTEKESITES})$$

- Az  $x$ -nél nagyobb eladások ahol a terméknek csak a neve szerepel

$$\pi_{\text{TERMEK.nev}} (\sigma_{\text{osszdb} > x} (\text{ERTEKESITES}))$$

# MD műveletek

## Mintapéldák

- Értékesítési adatok város szerinti bontásban

$$V_{TELEPHELY.varos} (ERTEKESITES)$$

- Az  $x$ -nél rosszabb selejtarányú eladásra vonatkozó adatok megye bontásban:

$$V_{VAROS.megye} (V_{TELEPHELY.varos} (\sigma_{selejt db/ossz db} > x (ERTEKESITES)))$$

- Áttérés hónap bontásról napi bontásra

$$\kappa_{HO} (C)$$

ahol  $C = V_{DATUM.ho} (ERTEKESITES)$

- Értékesítési adatok termék és idő dimenzióban

$$\phi_{TERMEK, DATUM, Sum} (ERTEKESITES)$$

## MD műveletek

### Mintapéldák

- Az x-nél rosszabb selejtarányú eladásra vonatkozó adatok megye és hónap bontásban:

$$\phi_{MEGYE, HONAP, Sum} (v_{DATUM.ho} (v_{VAROS.megye} (v_{TELEPHELY.varos} (\sigma_{selejtdb/osszdb} > x (ERTEKESITES))))))$$

- Az értékesítési adatok összevonása egy KOLTSEG (ertek, TELEPHELY,HO) adatkockával:

$$\phi_{TELEPHELY, HONAP, Sum} (v_{DATUM.ho} (ERTEKESITES)) \times KOLTSEG$$

- Havi költségadatok megjelenítése dollárban

$$\phi_{TELEPHELY, Sum} (\alpha_{ertek/225\text{ dollar}} (KOLTSEG))$$