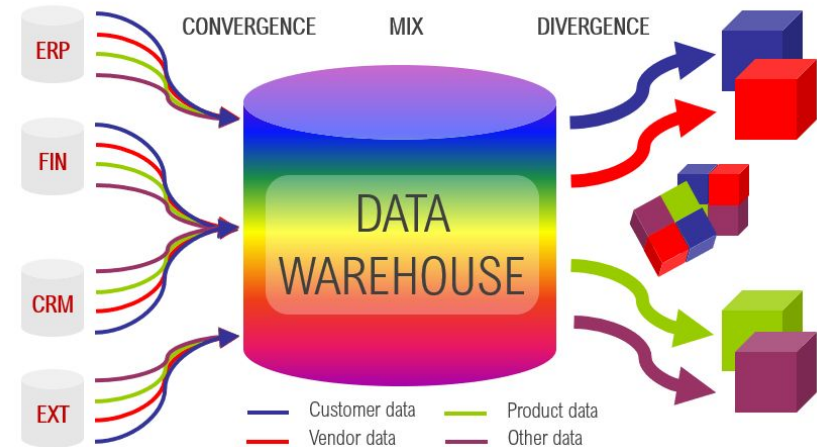


Adattárház rendszerek

Kovács László, ME

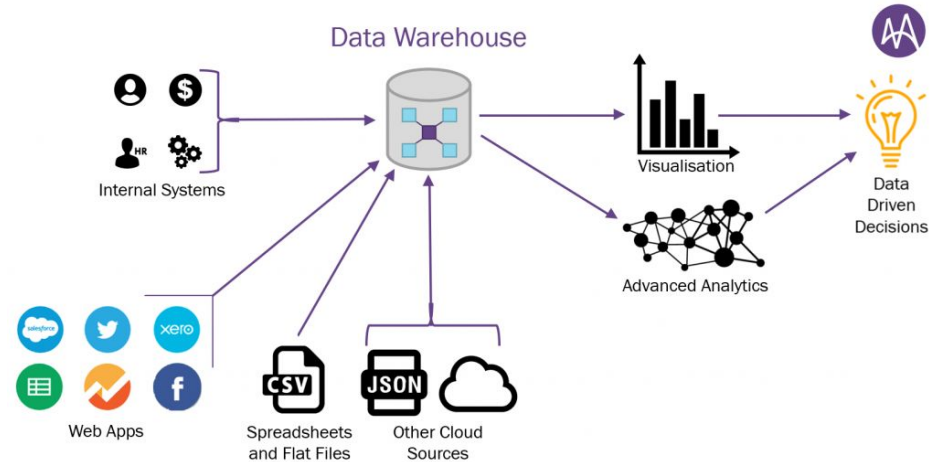
Adattárház

- speciális adatbázis
- egyedi tárolási-kezelési formátum
- adatelemzés orientált
- adatmegjelenítési elemek
- felhasználó orientált műveletek
- egyszerűbb adatstruktúra
- adatelemzés orientált belső szerkezet
- több-dimenziós adatmodell, adatkocka
- MDX

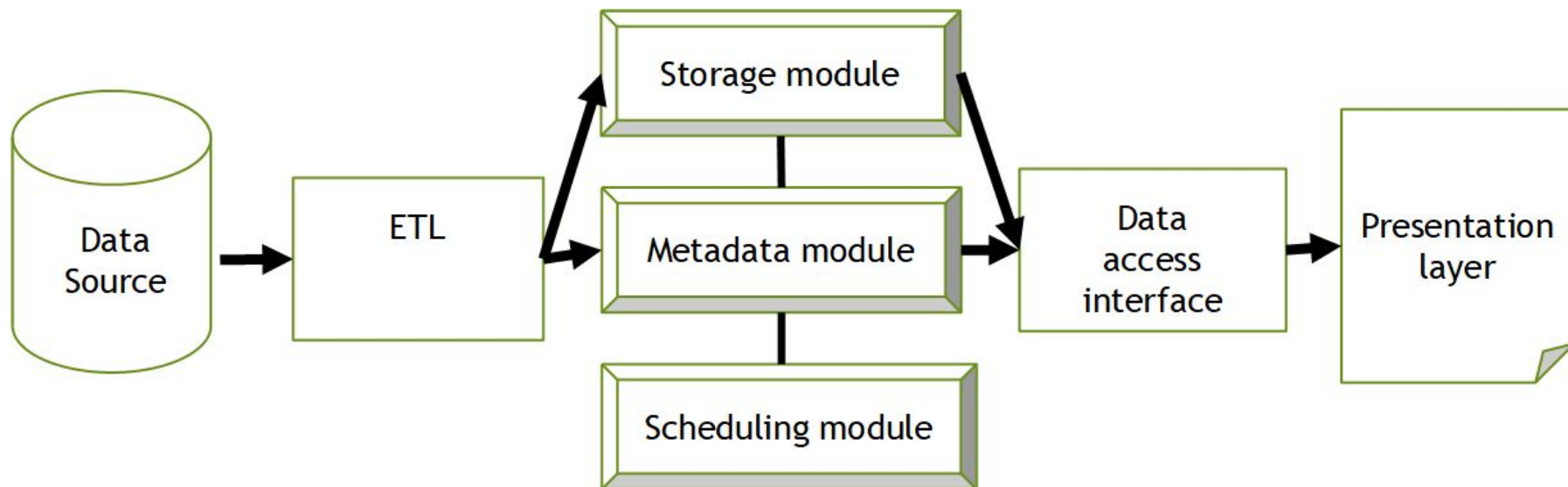


Adattárház

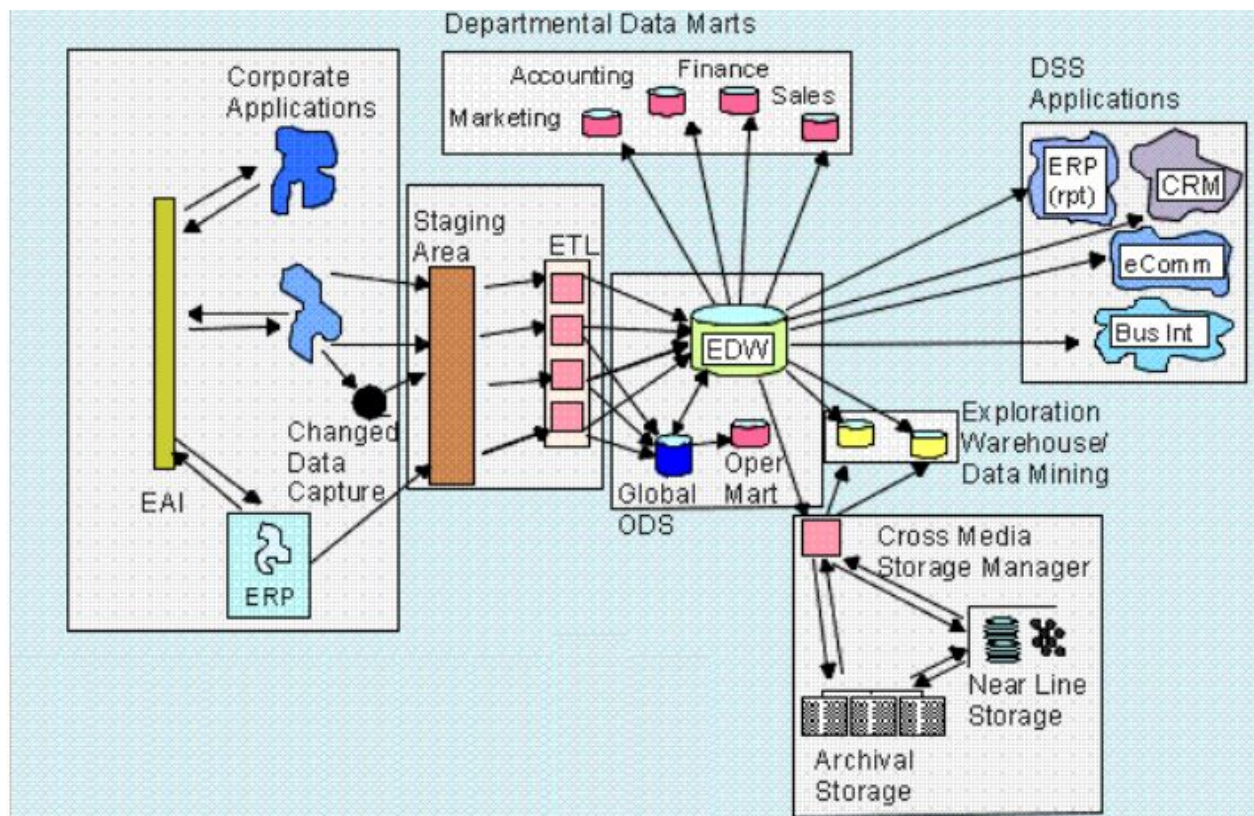
- Inmon (1995)
- OLAP alapú adattárolás
- heterogén adatok integrálása
- múltbeli adatok
- MD modell
- MD algebra
- nagy adatmennyiség



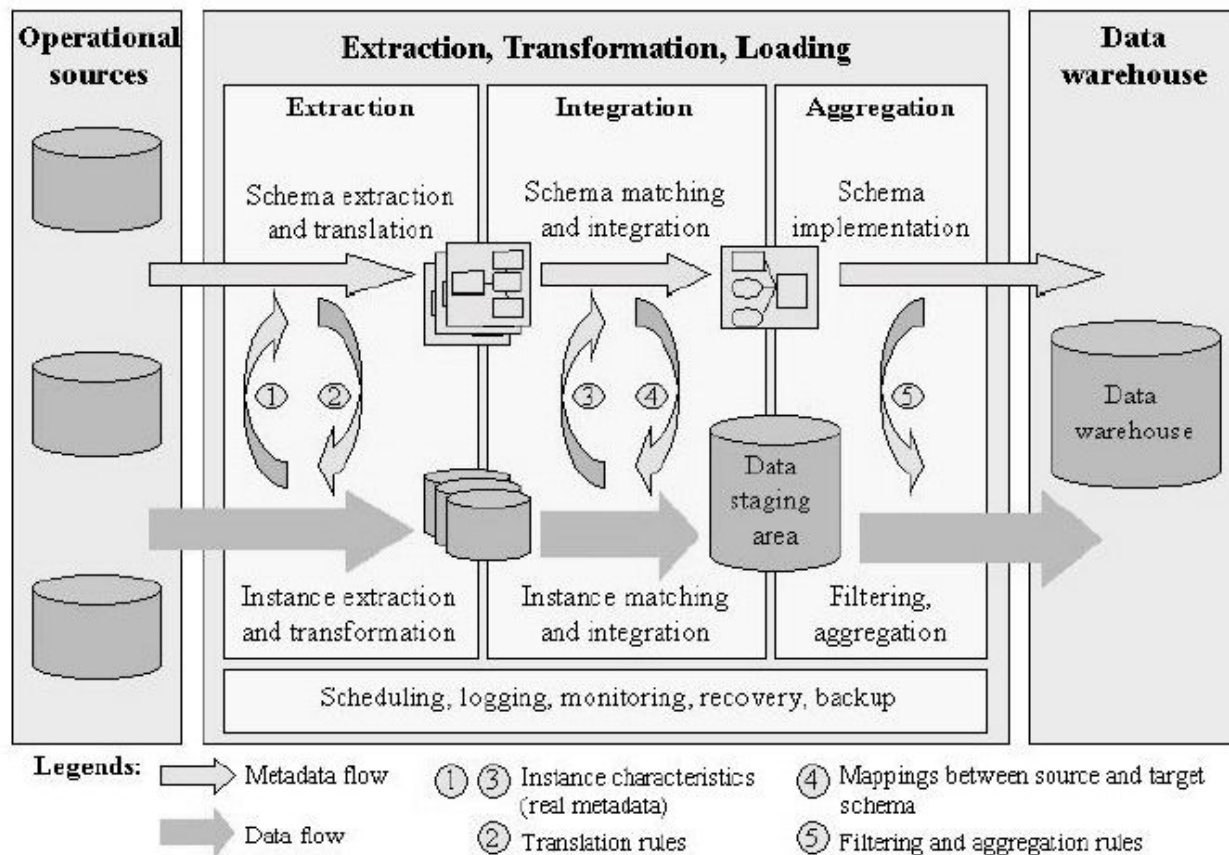
DW architektúra



DW architektúra



ETL architektúra



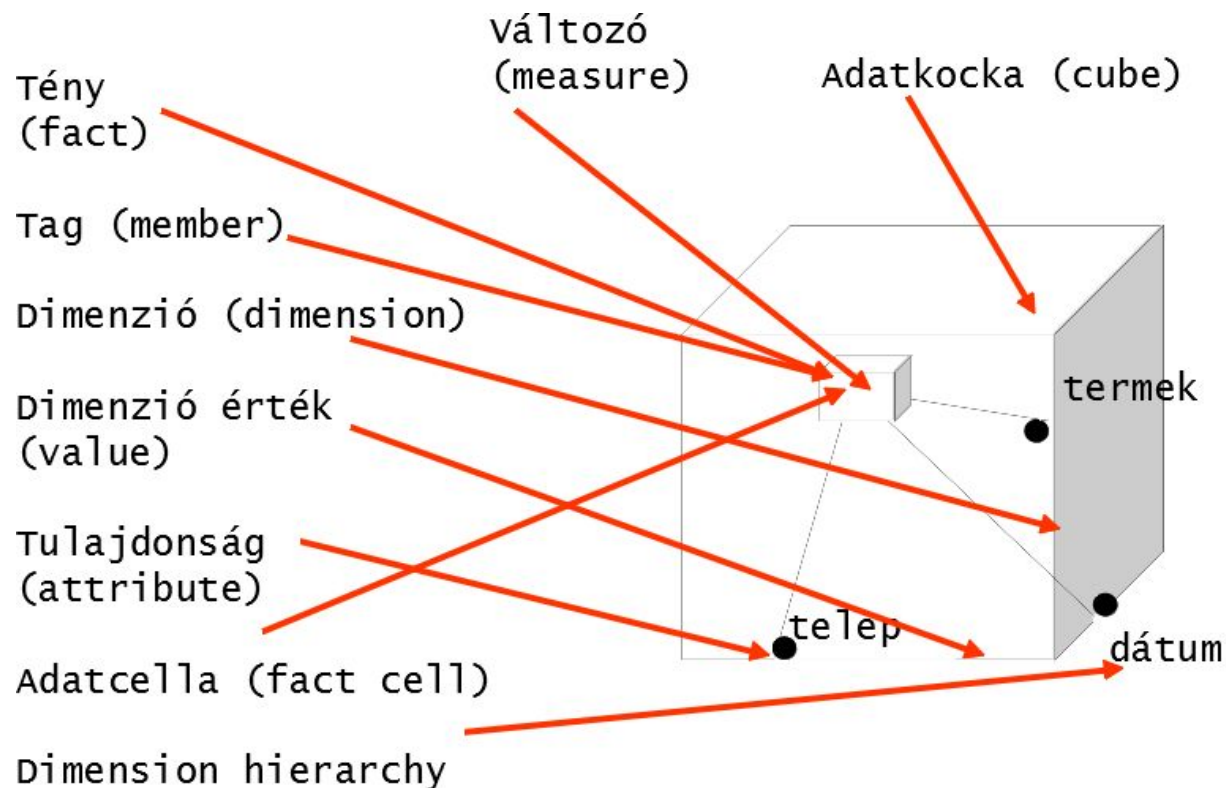
Adattisztítás

Inkonzisztencia okai:

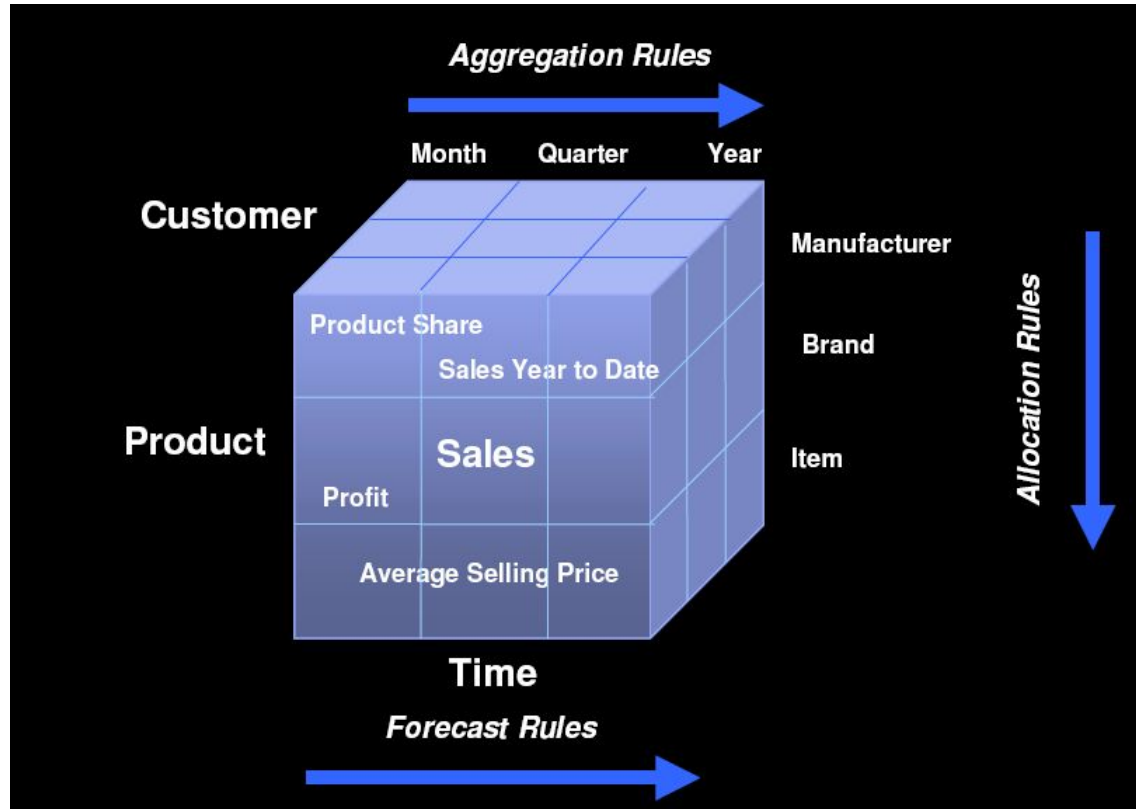
- hiányzó adatelem
- hiányzó adatérték
- hibás adat érték
- hibás számítások
- duplikáció
- eltérő formátum
- eltérő kódolás
- integritási hiba
- név konfliktus
- strukturális konfliktus



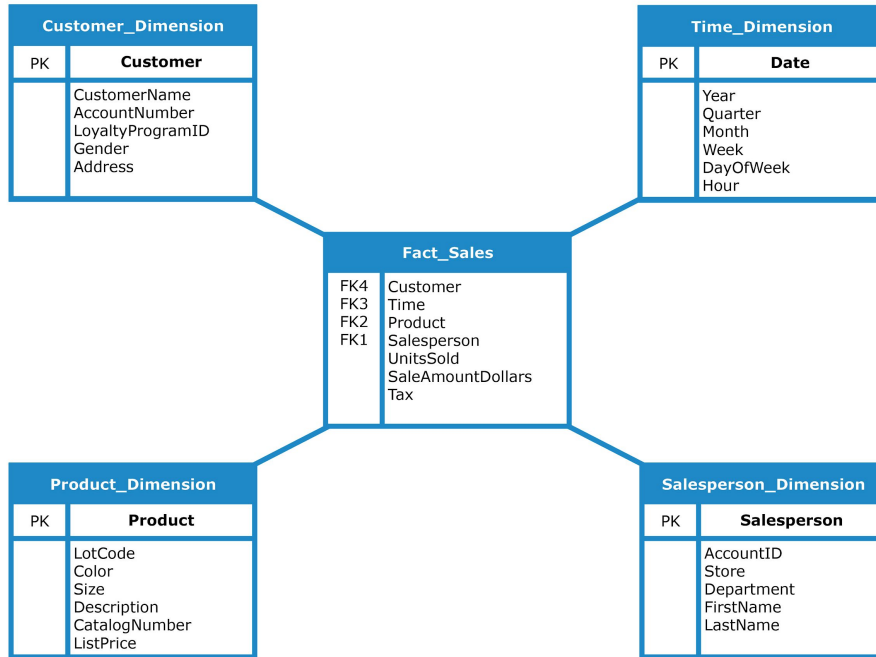
MD adatkocka modell



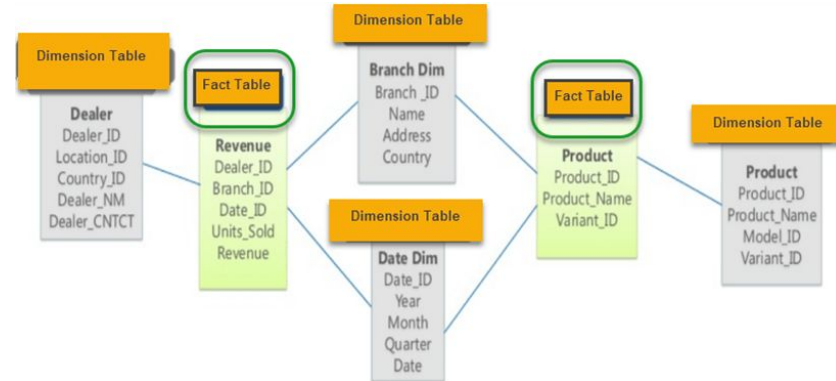
MD adatkocka



MD logikai modell : Star/Csillag séma

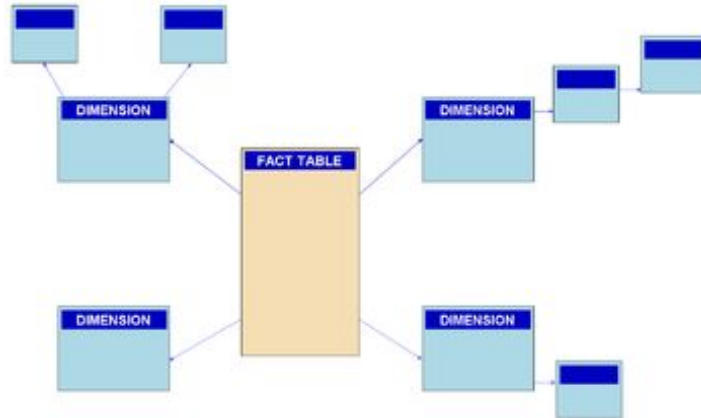


Legend: PK = primary key, FK = foreign key

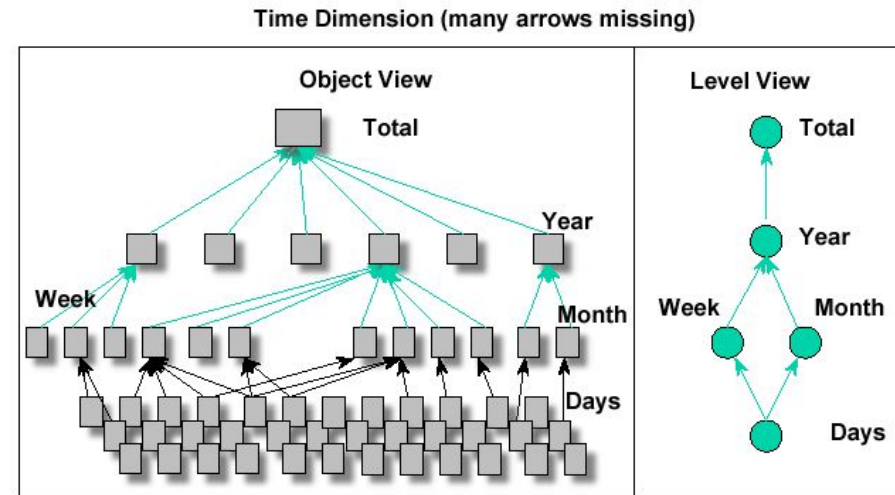


MD logikai modell : Snowflake/Hópehely séma

- a dimenziókhöz szintek (felbontási szintek) rendelhetőek
- szintek hierarchiája
- bázis szint
- PCR kapcsolat

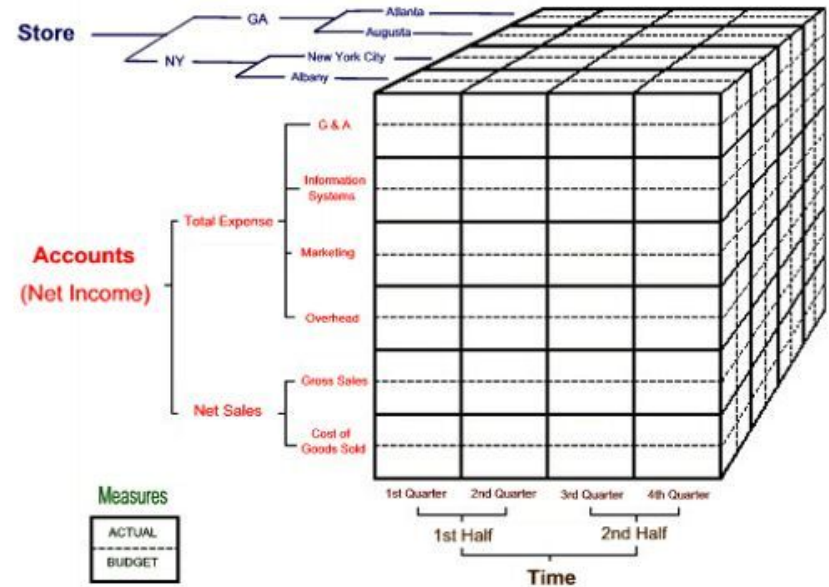


Dimension Hierarchies -- multiple



MD adatmodell

- elemi értékű cellák
- változók, PKI értékek
- dimenzió szintek
- measure dimenzió
- measure értékei változókat jelölnek
- a cellák lehetnek üresek
- member: a dimenzió egy értéke



Ajouter un nouvel ar... Console d'uti... Saiku Developm... Saiku Analytics System Dashboa... Meteorite B1 For... Saiku 3.0 RC2 re...

localhost:8080/pentaho/Home

Fichier Afficher Outils Aide...

Opened

admin

Saiku Analytics

Cubes

SteelWheelsSales

Mesures Add

Mesures

Quantity

Sales

Dimensions

Customers

(All)

Customer

Markets

Order Status

Product

(All)

Line

Vendor

Product

Time

(All)

Years

Quarters

Months

Mesures

Sales

Colonnes

Years

Time

Rangées

Line

Product

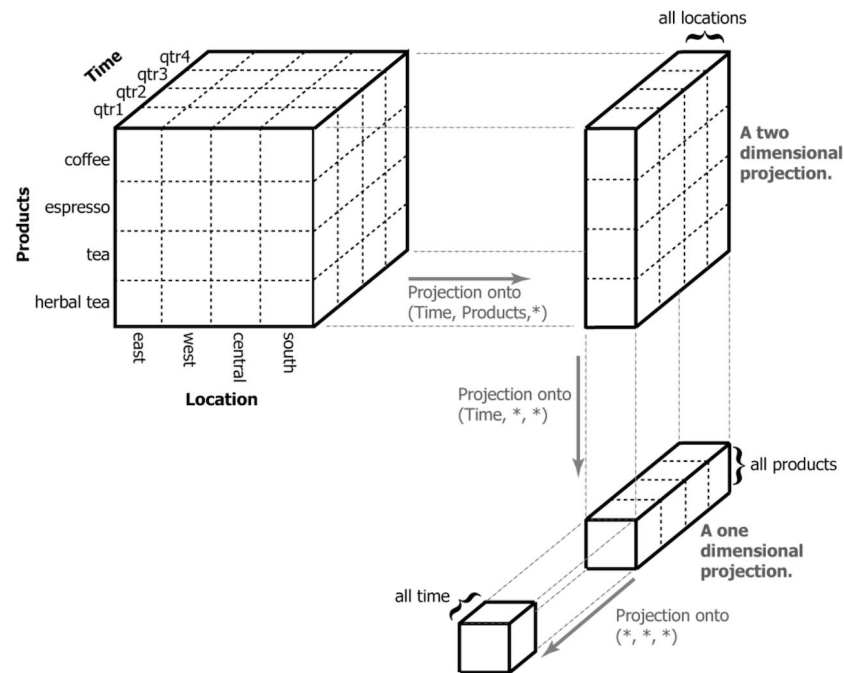
Filtre

Info: 15:02 / 4 x 9 / 0.05s

	2003	2004	2005
Line	Sales	Sales	Sales
Classic Cars	1 514 407	1 838 275	738 738
Motorcycles	397 220	590 580	286 325
Planes	347 755	528 928	200 074
Ships	244 821	375 672	128 178
Trains	72 802	124 750	36 917
Trucks and Buses	420 430	531 976	201 875
Vintage Cars	679 949	997 580	388 718
Grand Total	3 677 384	4 987 740	1 980 825

MD műveletei

- adatlekérdezés orientált
- elemi műveletek
- legkisebb kocka: egy skalár érték
- MD algebra:
 - selection (slice and dice)
 - drill down
 - roll up
 - fold
 - ...



MD művelet: selection

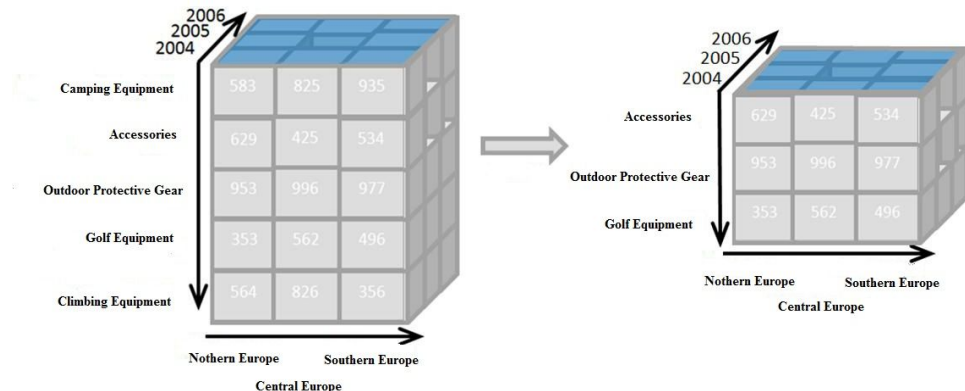
- **Slice:** A slice is a subset of a multi-dimensional array corresponding to a single value for one or more members of the dimensions not in the subset.
- **Dice:** The dice operation is a slice on more than two dimensions of a data cube (or more than two consecutive slices).

slice and dice: részkocka képzése

csak a feltételnek megfelelő cellák
maradnak meg

- variable: $\sigma_{f(v)}(\text{Cube})$
- attribute: $\sigma_{f(d.a)}(\text{Cube})$

$\sigma_{\text{profit} > 100}(\text{Sales})$



MD művelet: drill down and roll up

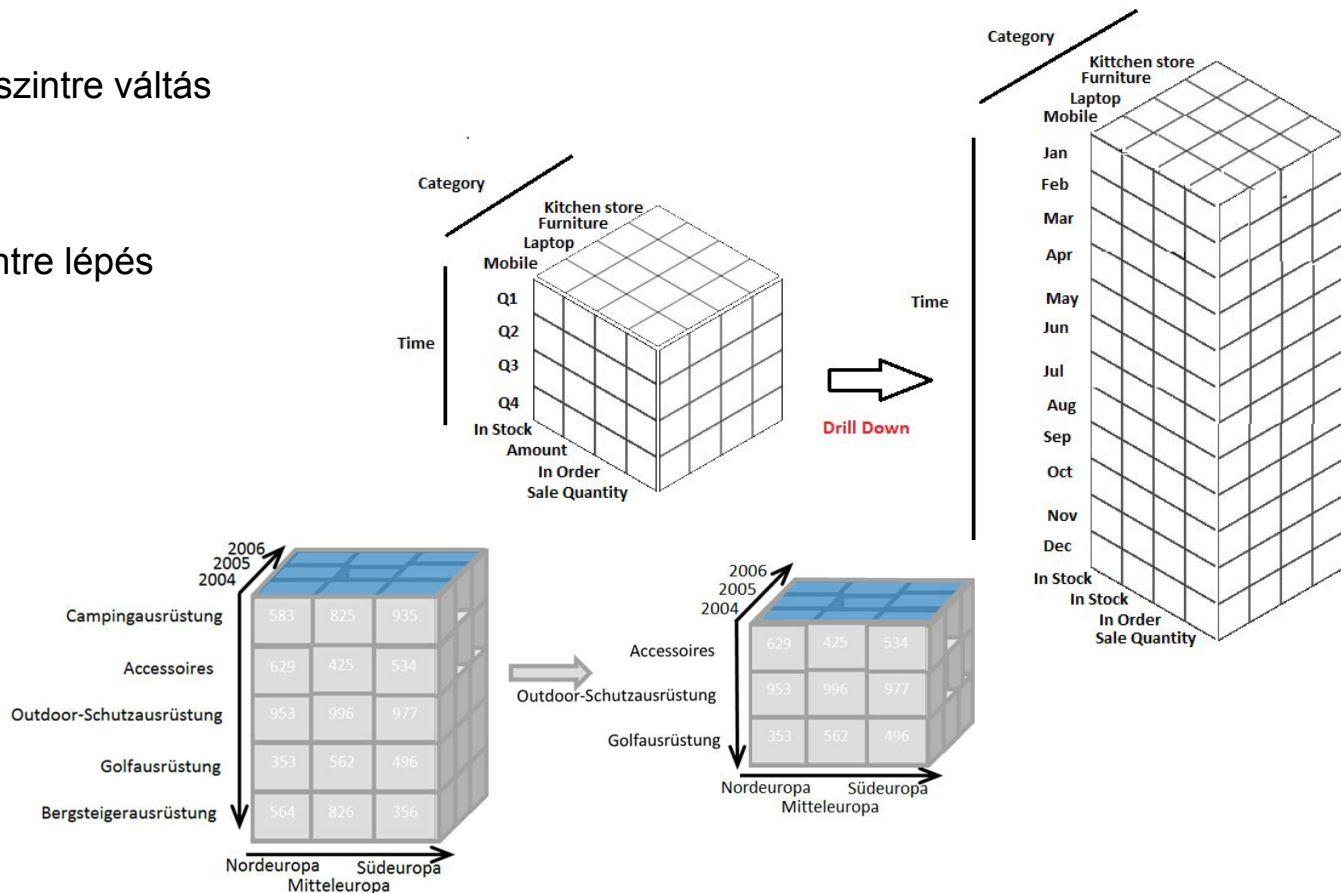
drill down: részletezőbb szintre váltás
(snowflake modell)

Részletek kijelzése

roll up: aggregáltabb szintre lépés
(snowflake modell)

Aggregáció jelzése

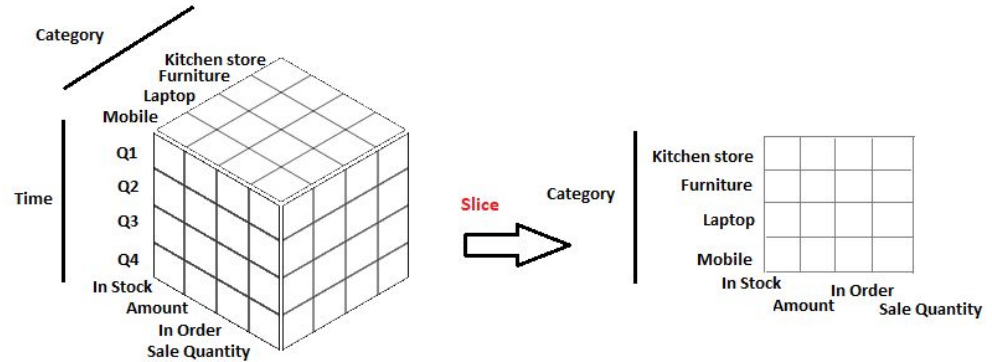
$\kappa_d(\text{Cube})$



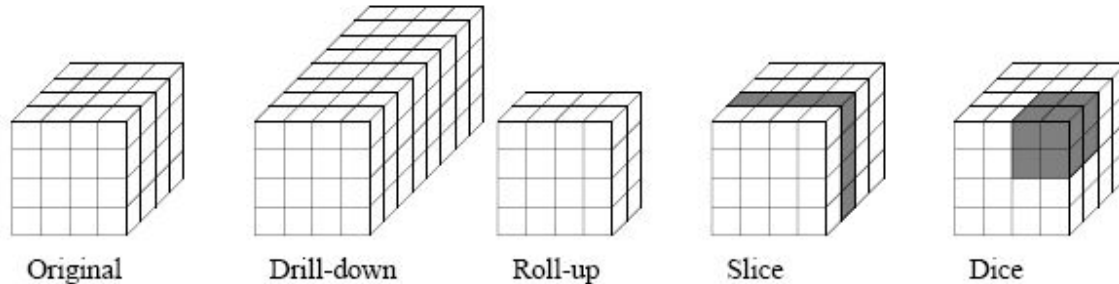
MD művelet: fold

fold: dimenziók megszüntetése

- jelentés: group by
- csökkentett dimenziószám
- a cellák összesített értékeket mutatnak



$$\phi_{d, \text{aggr}}(\text{Cube})$$



MD műveletek

Input kocka : Sales(customers:person,
products:item,time:month)

E1:

eladási adatok 2017-re:

$\sigma_{\text{measure} = \text{sales_amount and time} = 2017} (\kappa_{\text{time:year}}(\text{Sales}))$

E2:

eladott darabszámok magyar vevőkre nézve
2017.ben:

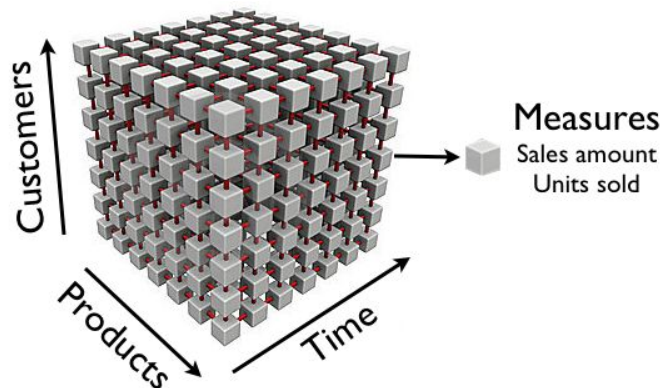
$\sigma_{\text{measure} = \text{sales_amount and customer} = \text{"Hungary"} \text{ and time} = 2017} (\kappa_{\text{time:year, customer: country}}(\text{Sales}))$

E3: össz eladási adatok havi bontásban

$\phi_{\text{time:sum}}(\text{Sales})$

E4: össz eladási adatok évi bontásban

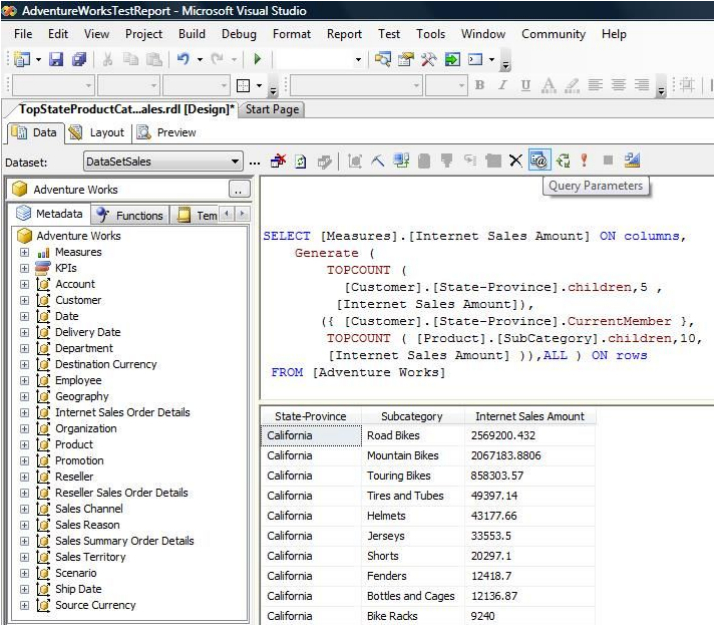
$\phi_{\text{time:sum}}(\kappa_{\text{time:year}}(\text{Sales}))$



MDX nyelv

MDX nyelv modellje:

- adatkocka
- dimenziók
- dimenzió hierarchia
- szintek (level)
- member: dimenzió érték
- Measure
- Tuple: értékek n-es különböző dimenziókból véve
- Set: azonos szerkezetű tuple-ek együttese
- Default member
- Property



The screenshot shows the Microsoft Visual Studio interface with the AdventureWorksTestReport project open. The main window displays an MDX query in the Design view. The query is as follows:

```
SELECT [Measures].[Internet Sales Amount] ON columns,
Generate (
    TOPCOUNT (
        [Customer].[State-Province].children, 5 ,
        [Internet Sales Amount]),
    ({ [Customer].[State-Province].CurrentMember },
    TOPCOUNT ( [Product].[SubCategory].children, 10,
    [Internet Sales Amount] )), ALL ) ON rows
FROM [Adventure Works]
```

Below the query, the results are displayed in a table with three columns: State-Province, Subcategory, and Internet Sales Amount. The results are as follows:

State-Province	Subcategory	Internet Sales Amount
California	Road Bikes	2569200.432
California	Mountain Bikes	2067183.8806
California	Touring Bikes	858303.57
California	Tires and Tubes	49397.14
California	Helmets	43177.66
California	Jerseys	33553.5
California	Shorts	20297.1
California	Fenders	12418.7
California	Bottles and Cages	12136.87
California	Bike Racks	9240

MDX nyelv

lekérdezés operátora:

```
SELECT [<axis_specification>  
      [, <axis_specification>...]]  
FROM [<cube_specification>]  
[WHERE [< slicer_specification>]]
```

```
<axis_specification> ::= <set> ON <axis_name>  
<axis_name> ::= COLUMNS | ROWS |  
                PAGES | SECTIONS | CHAPTERS | AXIS(<index>)
```

SELECT FROM Sales

SELECT {[MEASURES].[unit sold]} ON COLUMNS FROM Sales;

SELECT {[MEASURES].[unit sold]} ON AXIS(0), {Products.Fruit.MEMBERS} ON AXIS(1) FROM Sales;

MDX minta lekérdezés

```
SELECT Measures.MEMBERS ON COLUMNS,  
  
[Store].MEMBERS ON ROWS  
  
FROM [Sales]
```

```
SELECT Measures.MEMBERS ON COLUMNS,  
  
{[Store].[Store State].[CA], [Store].[Store State].[WA]} ON ROWS  
  
FROM [Sales]
```

```
SELECT [Product].[Product Family].MEMBERS ON COLUMNS,  
  
[Customers].[City].MEMBERS ON ROWS,  
  
[Time].[Quarter].MEMBERS ON PAGES  
  
FROM [Sales]  
  
WHERE (Measures.[Unit Sales])
```

Cubes



Electricity



Measures

Add

▼ Measures

Quantity

Avg

Dimensions

▶ Interval

▶ NMI

▶ Reading Type

▼ Time

(All)

Year

Month

Day

▶ TimeDayOfYear

▶ TimeMonthOfYear

```
1 WITH
2 SET [~ROWS] AS {[Time].[Day].[2014-01-01]}
3 SELECT
4 NON EMPTY {[Measures].[Quantity]} ON COLUMNS,
5 NON EMPTY [~ROWS] ON ROWS
6 FROM [Electricity]
```

1, 0

Day	Quantity
2014-01-01	6.284