Business Intelligence
Cursul 3

Prof. Ramona Bologa,
ASE Bucuresti

Agenda

1. **OLAP**
   - Cerinte functionale OLAP
   - Arhitecturi OLAP: ROLAP, MOLAP, desktop OLAP si Hybrid OLAP
2. **Data mining**
3. **Integrarea datelor**
   - BI si ERP;
   - Descrierea unui sistem ERP (SAP) integrat cu software BI

3. OLAP

- Cerinte functionale OLAP
  - Regulile lui Codd
  - Regulile FASMI
- Arhitecturi OLAP: ROLAP, MOLAP, desktop OLAP si Hybrid OLAP

Ce este OLAP

- Dc instrumentele de interfata si structura BD suporta analiza multidimesionala, acces instantaneu si manipulare usoara => **online analytical processing**
- **Codd**, parintele acestui termen a evidentiat diferentele OLTP-OLAP- 1993 - **criterii generale** pentru BD OLAP.
- **ANALIZA MULTIDIMENSIONALA**
  - Aplicarea de **formule şi modele** asupra dimensiunilor şi ierarhiilor;
  - Vizualizarea datelor prin **mai multe filtre sau dimensiuni** in acelasi timp
  - **Analiza de trend** pe perioade diferite de timp;
  - Analiza în adancime (**drill-down**);
  - **Extragerea unui subset** de date pentru vizualizare(slice);
  - **Rotaţii** în cadrul dimensiunilor;

OLAP si DW

- Sistemele OLAP şi DW - sisteme suport de decizie orientate pe date şi sunt similare.
- **DW** pune accentul pe procesele ce asigură consistenţa, corectitudinea şi valabilitatea datelor la utilizatori,
- **sistemele OLAP** pun accentul pe *cerinţele analitice* şi *procesele de modelare şi calcul* necesare.

Cerinte functionale OLAP - Codd

Caracteristici de bază

1: O viziune conceptuală multidimensională
2: Manipularea intuitivă a datelor
3: Accesibilitate
4: Surse de date variate
5: Modele de analiză OLAP
6: Arhitectura client/server
7: Transparenţă
8: Suport multiutilizator

Caracteristici speciale

9: Denormalizarea datelor
10: Stocarea rezultatelor generate de instrumentul OLAP
11: Manipularea valorilor lipsă
12: Modul de tratare a valorilor lipsă

Modul de prezentare a datelor

13: Flexibilitatea rapoartelor
14: Performanţa raportării
15: Ajustarea automată a nivelului fizic

Controlul dimensiunilor

16: Dimensionalitate generică
17: Dimensiuni şi niveluri de agregare nelimitate
18: Operaţii între dimensiuni nerestrictive

Regulile FASMI (1995 Nigel Pendse)

- **Fast Analysis of Shared Multidimensional Information**
- ***FAST*** - cat mai multe raspunsuri utilizatorilor intr-un termen mai scurt de 5 sec
- ***ANALYSIS*** - orice logica de afaceri si analiza statistica relevanta pentru aplicatie sau utilizator, suficient de simplu pentru utilizatorul final
- **SHARED** - toate cerintele de securitate pentru confiden-tialitate, dar si blocarea actualizarilor concomitente, daca este necesar accesul multiplu la scriere
- **MULTIDIMENSIONAL** - viziune **conceptuala multidimensionala** asupra datelor, incusiv suport complet pentru ierarhii si ierarhii multiple
- **INFORMATION** reprezinta toate datele si informatiile derivate necesare oriunde se afla si in orice masura este relevanta pentru aplicatie.

a. Arhitecturi OLAP - ROLAP

- **Relational OnLine Analytic Processing**
- tehnologia relaţionala, adaptată şi extinsă
- agregările sunt stocate în cadrul BD relaţionale sursă
- cea mai lentă soluţie, ex: DSS Server/Microstrategy
- Avantaje:
  - se integrează cu tehnologia şi standardele existente;
  - actualizarea sistemelor MOLAP este dificilă;
  - ROLAP sunt adecvate pentru a stoca volume mari de date, prin utilizarea **procesării paralele** şi a **tehnologiilor de partiţionare**;
  - ROLAP sunt recomandate pentru aplicaţiile cu volatilitate ridicată a datelor (antecalcul agregari)

ROLAP atunci cand:

a) Volumul de date este ***prea mare*** pentru a fi duplicat.
b) Datele sursă ***se modifică frecvent*** şi este mai bine de a citi în timp real decât din copii;
c) Se doreşte ***integrarea cu alte sisteme*** informatice relaţionale existente;
d) Firma are o ***politică de neduplicare*** a datelor, pentru securitate sau alte motive, chiar dacă aceasta conduce la aplicaţii mai puţin eficiente

b. Arhitecturi OLAP - MOLAP

- **Multidimensional OnLine Analytic Processing**
- stocarea datelor în formă multidimensională, folosind **structuri de date vector (tehnica matricilor rare)**
- atât datele sursă, cât şi agregările sunt stocate în format multidimensional
- indexare rapida a datelor preagregate
- opţiunea cea mai *rapidă pentru consultare*
- necesită cel mai *mult spatiu* de disc
- stocarea fizică a datelor multidimensionale, precum şi fenomenul de împrăştiere sunt preocupări majore
- Ex: Oracle Essbase

Avantaje MOLAP

- tabelele nu sunt potrivite pentru date multidimensionale;
- **matricile multidimensionale** permit stocarea eficientă a datelor multidimensionale;
- **limbajul SQL** nu este corespunzător pentru operaţii OLAP

c. Arhitecturi OLAP - HOLAP

- **Hybrid OnLine Analytic Processing**
- combinaţie a primelor două modele
- Arhitecturi HOLAP
  - agregările - stocate in structură multidimensională, nivelul celulelor de bază în formă relaţională
  - cele mai recente felii de date stocate in MOLAP si restul in ROLAP
- oferă performanţele MOLAP atunci când este nevoie de preluarea datelor din tabele
- Ex: Microsoft SQL Server OLAP Services

Caracteristici HOLAP

- transparenţa locaţiei şi a accesului
- transparenţa fragmentării
- transparenţa performanţei
- un model de date comun
- alocarea optimă în sistemele de stocare

Solutii OLAP

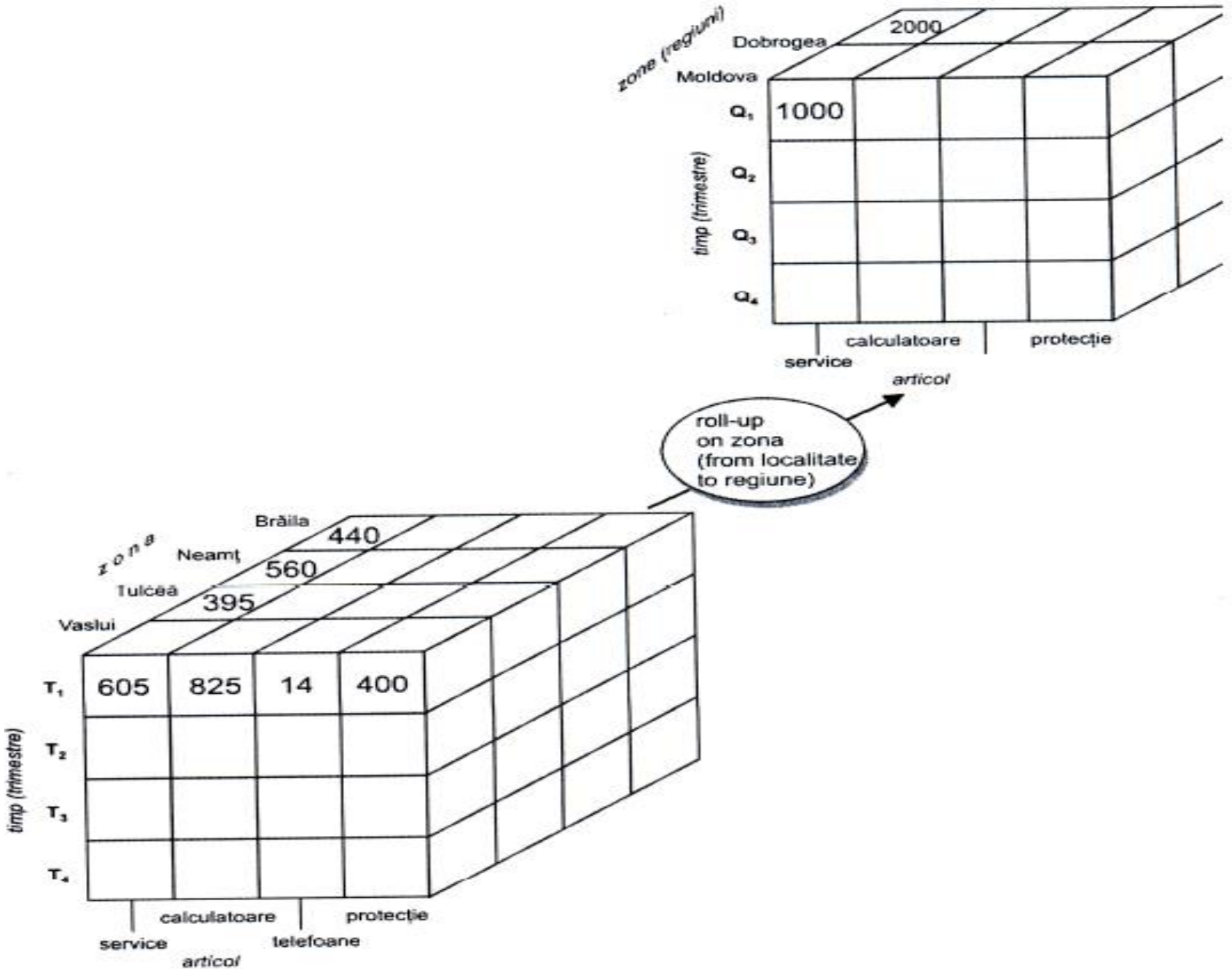
	Essbase	SSAS	Mondrian	Oracle OLAP	SAS OLAP	SAP Netweaver BW	Microstrategy Intelligence Server	Cognos TM1
Firma	Oracle	Microsoft	Pentaho	Oracle	SAS Institute	SAP	Microstrategy	IBM
Arhitectura	MOLAP ROLAP HOLAP	MOLAP ROLAP HOLAP Power Pivot pt excel	ROLAP	MOLAP ROLAP HOLAP	MOLAP ROLAP HOLAP	MOLAP ROLAP	MOLAP ROLAP HOLAP	TM1-MOLAP
Lb MDX	DA	DA	DA	NU	DA	DA	DA	DA
Lb SQL	NU	DA	NU	DA	NU	NU	DA	NU
Standard XML for analysis	DA	DA	DA	NU	DA	DA	DA	DA
Masuri semiaditive	DA	DA	NU	DA	DA	DA	DA	DA
SO	Win Linux Unix	Win	Win Linux Unix	Win Linux Unix	Win Linux Unix	Win Linux Unix	Win Linux Unix	Win Linux Unix

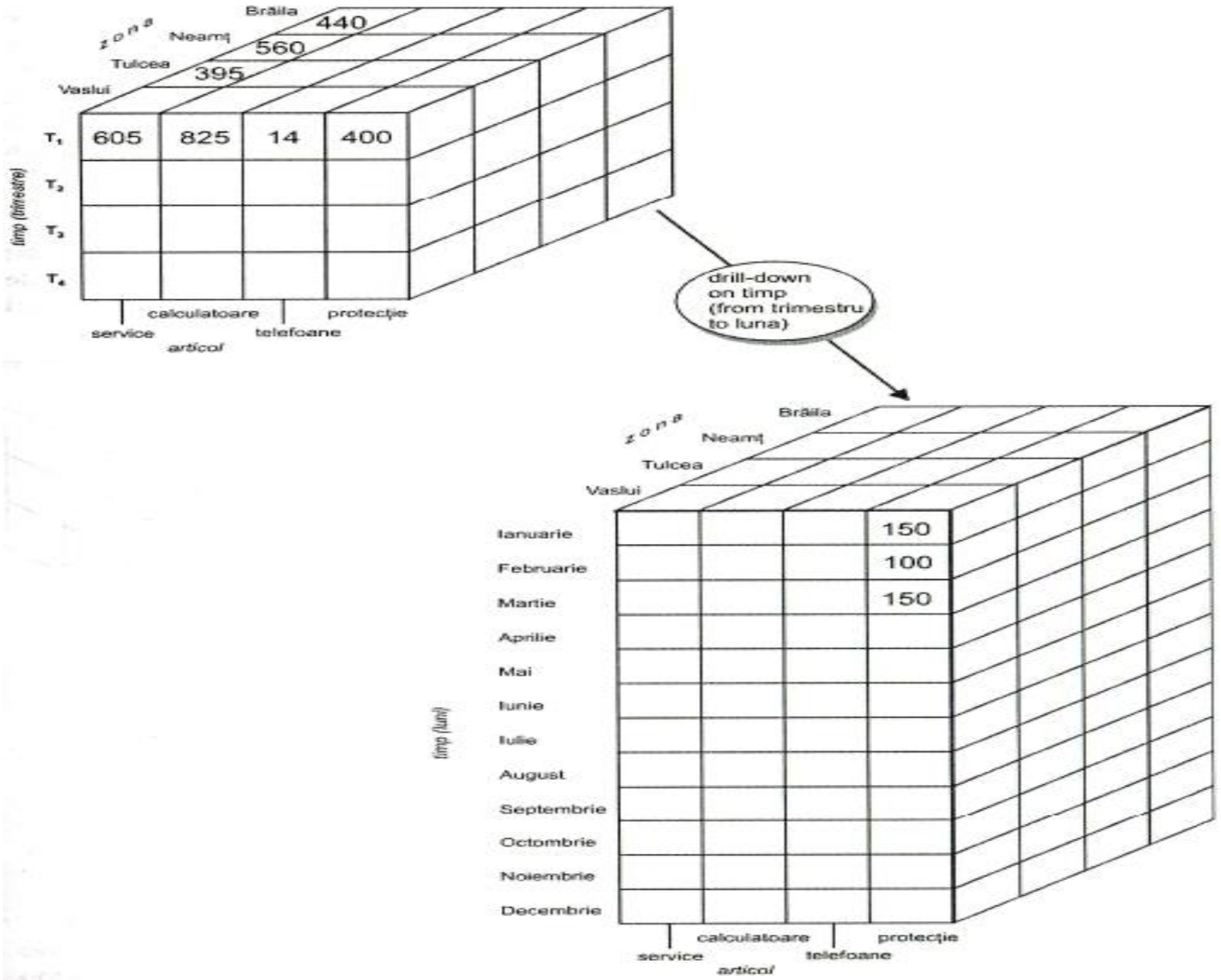
Operatori OLAP: Slice, dice, pivot

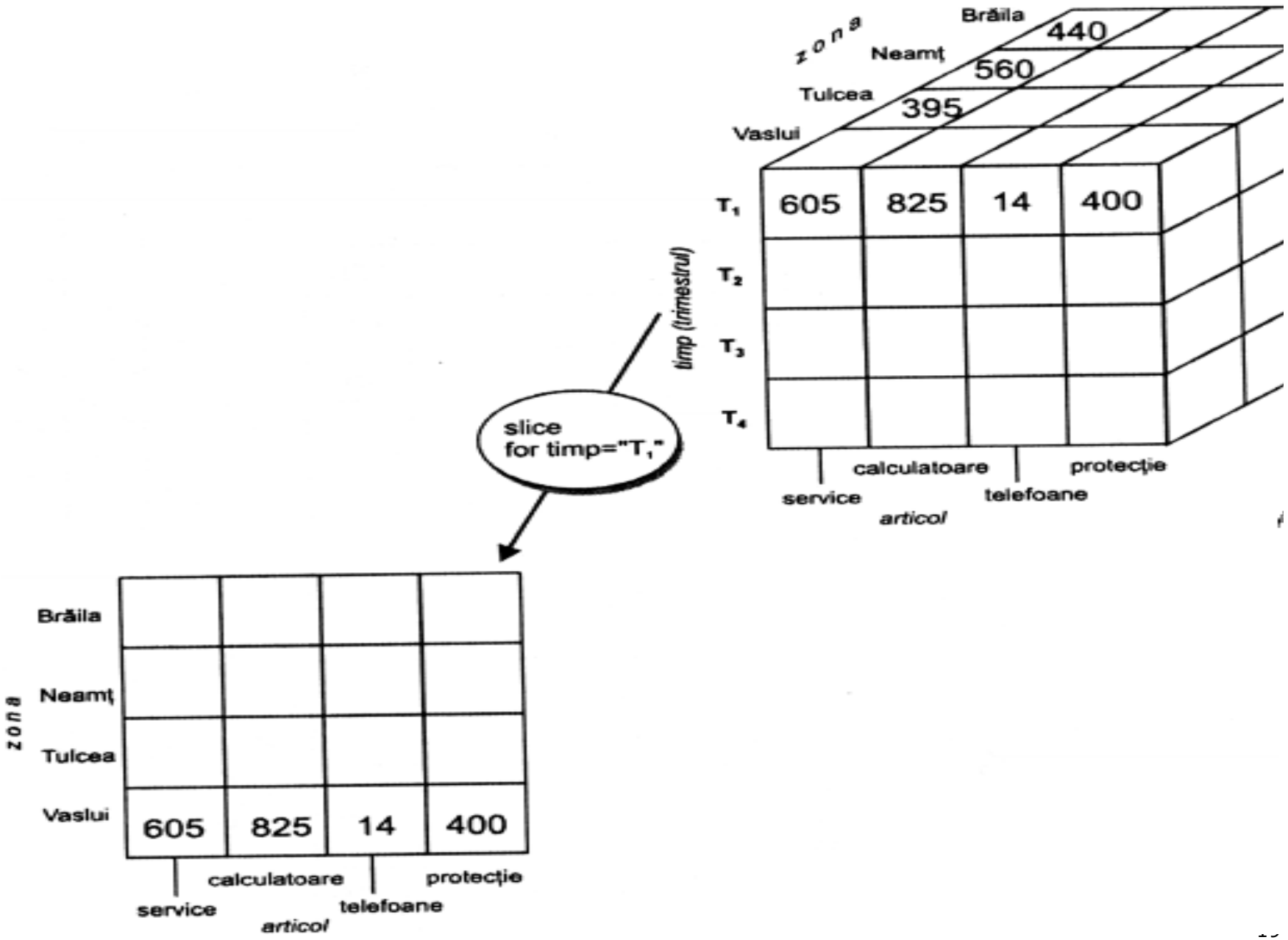
- **Secţionarea** (**slice**) –selectarea unui membru a unei dimensiuni, crearea unei serii de intersecţii cu alte dimensiuni al acelui membru pentru secţiunea respectivă.
- **Decupare (dice)** – crearea unui subcub al cubului de date prin selectarea unor dimensiuni si a intervalelor de valori pentru acestea. De exemplu, vanzari după luna, după regiune, după client. Acest "după" ne indică cum putem realiza **rotirea (dicing)** datelor.
- **Pivotarea** şi **imbricarea** dimensiunilor. Pivotarea presupune inlocuirea intre ele a dimensiunilor in cadrul unei vizualizari, trecerea de pe linii pe coloane si invers.

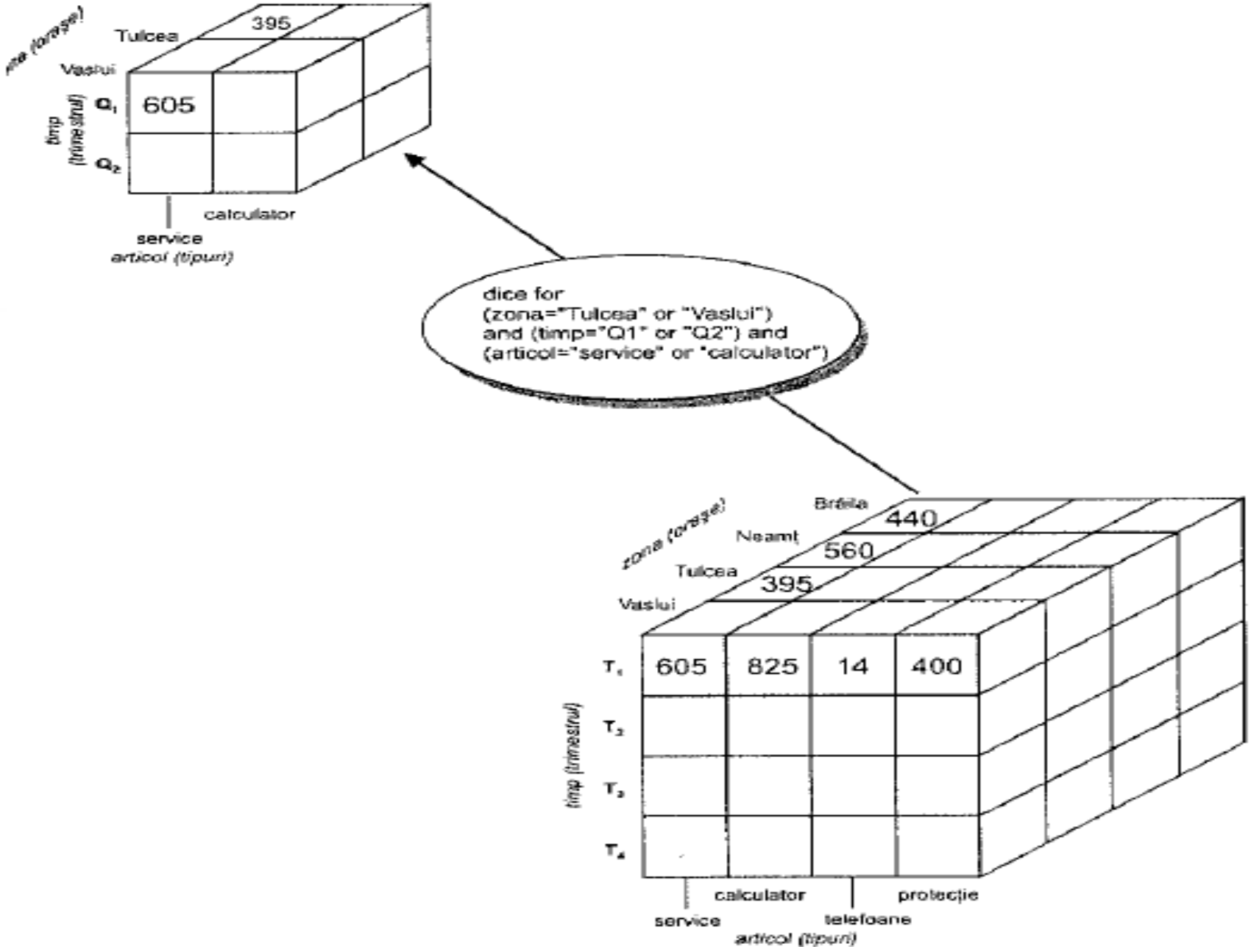
Ierarhii si navigare

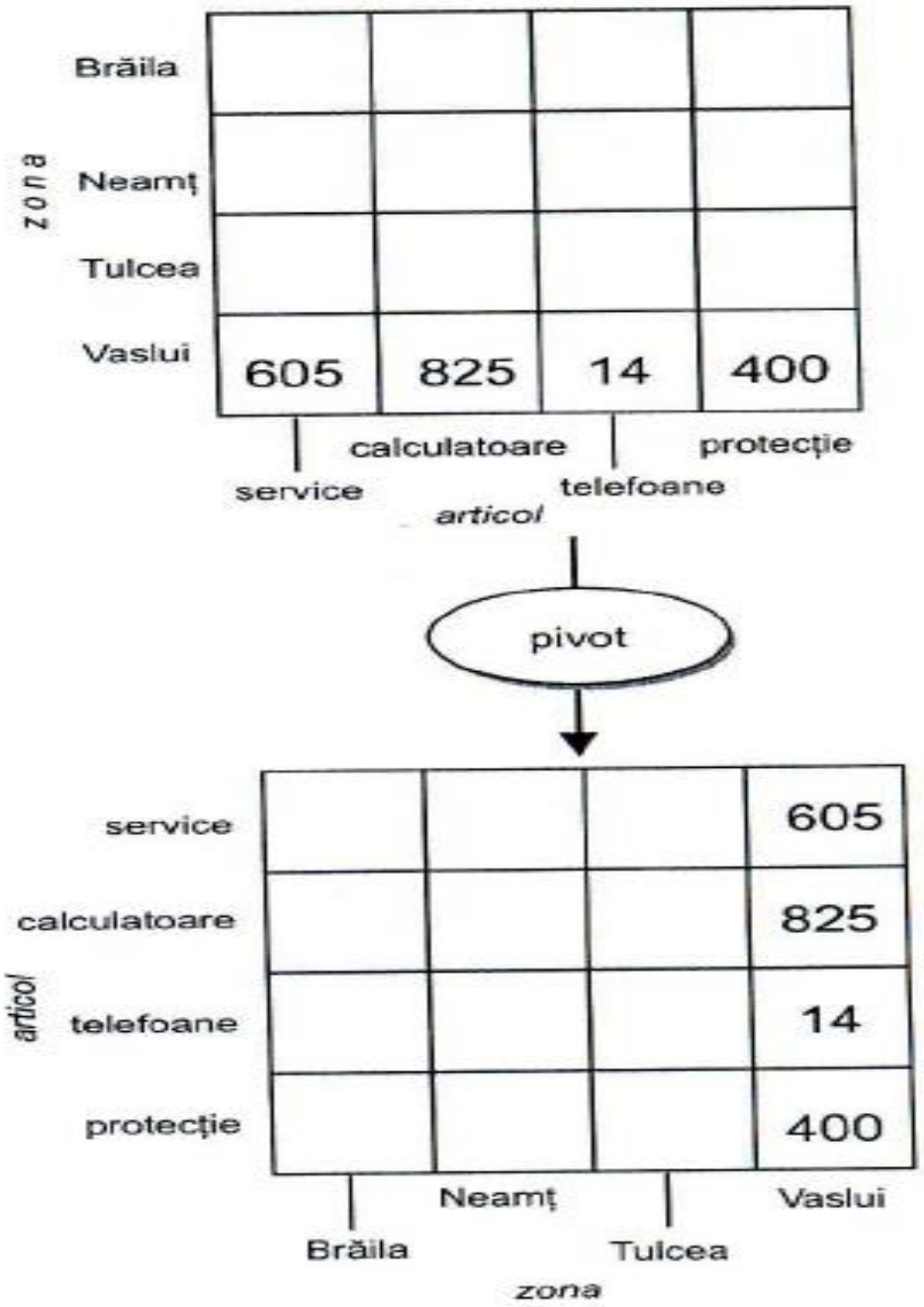
- Sistemele OLAP permit agregarea sub formă de ierarhii care realizează agregarea detaliilor la nivele din ce în ce mai înalte.
- De exemplu, datele lunare pot fi agregate (**roll-up**) la nivel de trimestru sau la nivel de an,
- Plecarea din vârful ierarhiei dimensiunilor şi detalierea (**drill-down**)
- Ia nastere astfel un nou proces, denumit «**analiză ad-hoc**», care permite :
  - Răspunsul la diferitele întrebări ale managerilor în doar câteva minute de navigare în date
  - Formatarea rapoartelor prin pivotarea şi imbricarea dimensiunilor
  - Învăţarea rapidă a utilizării unui astfel de sistem de către orice persoană, mai ales managerii


zone (regiuni)
Dobrogea
2000
Moldova
Q1
1000
Q2
Q3
Q4
timp (trimestre)
calculatoare
protecție
service
articol
roll-up
on zona
(from localitate
to regiune)
Brăila
440
zone
Neamț
560
Tulcea
395
Vaslui
T1
605
825
14
400
T2
T3
T4
timp (trimestre)
calculatoare
protecție
service
telefoane
articol


zonă
Brăila
Neamț
Tulcea
Vaslui
440
560
395
T1
T2
T3
T4
605
825
14
400
timp (trimestre)
service
calculatoare
telefoane
protecție
articol
drill-down on timp (from trimestru to luna)
zonă
Brăila
Neamț
Tulcea
Vaslui
Ianuarie
Februarie
Martie
Aprilie
Mai
Iunie
Iulie
August
Septembrie
Octombrie
Noiembrie
Decembrie
150
100
150
timp (luni)
service
calculatoare
telefoane
protecție
articol


zona
Brăila
Neamț
Tulcea
Vaslui
440
560
395
T1
T2
T3
T4
605
825
14
400
timp (trimestrul)
calculatoare
protecție
service
telefoane
articol
slice
for timp="T1"
Brăila
Neamț
Tulcea
Vaslui
zona
605
825
14
400
calculatoare
protecție
service
telefoane
articol


zona (oraşe)
Tulcea
395
Vaslui
timp (trimestrul)
Q1
605
Q2
calculator
service
articol (tipuri)
dice for
(zona="Tulcea" or "Vaslui")
and (timp="Q1" or "Q2") and
(articol="service" or "calculator")
zona (oraşe)
Brăila
440
Neamt
560
Tulcea
395
Vaslui
T1
605
825
14
400
T2
T3
T4
timp (trimestrul)
calculator
protecţie
service
telefoane
articol (tipuri)


Brăila
Neamț
Tulcea
Vaslui
zona
605
825
14
400
calculatoare
protecție
service
telefoane
articol
pivot
service
calculatoare
telefoane
protecție
articol
605
825
14
400
Neamț
Vaslui
Brăila
Tulcea
zona

2. Data Warehouse si Data mining

- Trei tipuri de aplicatii de DW
  - **Procesarea informatiilor**
    - Interogari, analize statistice de baza, raportari folosind tabele, grafice, figuri
  - **Procesare analitica**
    - Analiza multidimensionala a datelor DW
    - Operatii OLAP de baza, navigare prin date, pivotari, rotatii, sectionari
  - **Data mining**
    - Descoperire de cunostinte din modele ascunse
    - Asocieri, construire de modele analitice, realizare de clasificari si predictii, si prezentarea rezultatelor cu instrumente de vizualizare
    - OLAM –Online Analitycal Data Mining

Despre data mining

- Converteste **datele** in **cunostinte valoroase** care pot fi folosite ca suport pentru decizii
- Este o *colectie de metodologii, tehnici si algoritmi de analiza a datelor* pentru descoperirea de **modele noi** in date
- Este folosit pentru **seturi mari de date**
- Procesul este **automatizat**, nu e necesara interventia umana
- ***Data mining*** si ***Knowledge Discovery in Databases (KDD)*** sunt considerate de unii autori ca reprezentand acelasi lucru. Altii considera data mining-ul ca fiind *pasul de analiza* in procesul KDD, dupa curatarea si transformarea datelor si inainte de vizualizare/ evaluarea rezultatelor

Despre data mining

- Modelele trebuie sa fie *valide*, *utile* si *inteligibile*
- Implica **metode** care sunt la intersectia intre inteligenta artificiala, invatare automata (machine learning), statistica si sisteme de baze de date.
- Cele mai valoroase **rezultate** obtinute prin DM sunt: clusterizarea, clasificarea, estimarea, predictia si gasirea lucrurilor care apar impreuna.
- Principalele instrumente de DM includ:

  - Arbori de decizie,
  - Retele neuronale
  - Instrumente de vizualizare
  - Algoritmi genetici
  - Logica fuzzy
  - Metode statistice clasice.

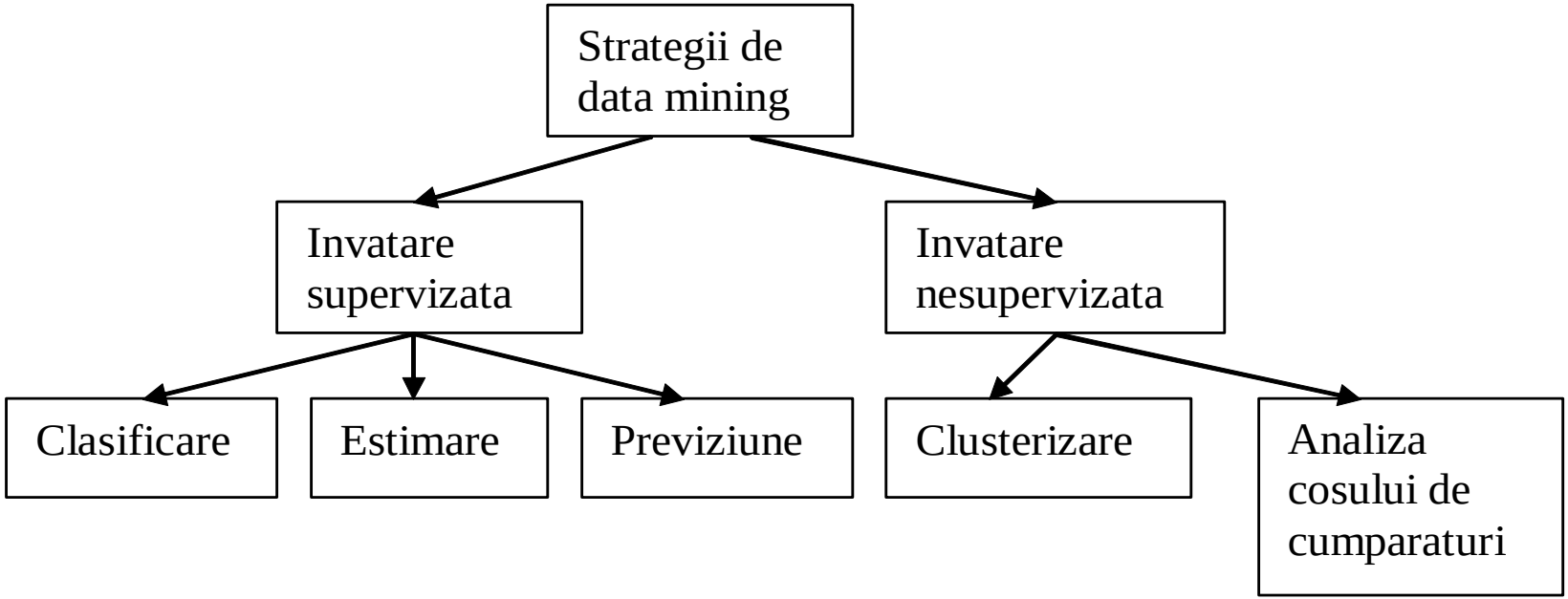
De unde provine?

Povesti de succes

- **Arbori de decizie** construiti pe baza istoricului **imprumuturilor** acordate de banci pentru a decide daca sa acorde un imprumut
- **Modele ale comportamentului de calatori** utilizate pentru a gestiona vanzarea cu discount a biletelor de avion, camerelor de hotel, etc.
- **"Scutece si bere"** Observatia ca acei clienti care cumpara scutece au probabilitate mai mare decat media de a cumpara si bere a permis supermarket-urilor sa amplaseze Berea langa scutece. Si chips-urile langa. Au crescut vanzarile la toate trei.
- Skycat si Sloan Sky Survey: **clusterizarea corpurilor ceresti** in functie de nivelul lor de radiatii in diferite benzi, a permis astronomilor sa diferentieze intre galaxy, stele in formare si alte tipuri de obiecte ceresti.
- **Compararea genotipului** unor persoane care au/ nu au anumite afectiuni au permis descoperirea unor seturi de gene responsabile pentru cele mai multe din cazurile de diabet. Acest gen de DM va deveni tot mai util, odata ce genomul uman este construit

Data mining

- **Pasi**:
  - i: Culegerea datelor
  - ii: Preprocesarea (pregatirea) datelor de analizat.
  - iii: Analiza datelor sau aplicarea unui algoritm/metode de DM
    - Invatare supervizata
    - Invatare nesupervizata
  - iv: Vizualizarea si interpretarea rezultatelor algoritmului
  - v: Aplicarea rezultatelor obtinute la noi probleme.



Pasii procesului de data mining

1. **Culegerea datelor**: colectarea datelor din baze de date sau prin cautari pe Web
2. **Preprocesarea datelor**
   a) ***Curatarea datelor***: inlocuirea (sau stergerea) valorilor lipsa, eliminarea sau doar identificarea valorilor extreme, eliminarea zgomotelor din date, inlaturarea inconsistentelor.
   b) ***Integrarea datelor***: datele sunt preluate din surse multiple, cu tipuri de date si structure diferite, sunt integrate si se elimina duplicatele si inconsistentele
   c) ***Transformarea datelor***: normalizarea (sau standardizarea), sumarizari, generalizari, construirea de noi atribute, etc.
   d) ***Reducerea datelor*** (sau extragerea caracteristicilor): doar atributele relevante sunt selectate pentru procesare ulterioara
   e) **Discretizarea**: deoarece unii algoritmi lucreaza doar cu valori discrete, valorile atributelor continue trebuie inlocuite cu valori discrete dintr-o lista predefinita

Pasii procesului de data mining

3. **Analiza datelor**: se aplica algoritmii de DM si se realizeaza extragerea si descoperirea de modele.
4. **Vizualizarea**: deoarece DM extrage proprietati si informatii ascunse din date, pentru a intelege si evalua rezultatele e necesara vizualizarea lor
5. **Evaluarea rezultatelor**: nu toate rezultatele obtinute prin DM sunt informatii valoroase. Pot rezulta adevaruri statistice sau informatii care nu sunt utile in activitatea analizata. Expertii sunt cei care vor evalua rezultatele.

Metode de data mining

- **Metode predictive** – utilizeaza niste variabile pentru a *prezice valoarea* altor variabile.
  - **Clasificarea** – se bazeaza pe date cunoscute, etichetate si algoritmii de clasificare construiesc modele pentru a clasifica date noi
  - **Regresia**
  - **Detectarea deviatiilor** (ex: fraude, intruziuni)
- **Metode descriptive**: algoritmii gasesc modele care descriu *structura interna* a setului de date.
  - **Clusterizarea** – identifica grupuri de obiecte similare (CLUSTERE) din setul de date, dar si posibile obiecte izolate, valorile extreme.
  - Descoperirea **regulilor de asociere**
  - Descoperirea **pattern-urilor secventiale**

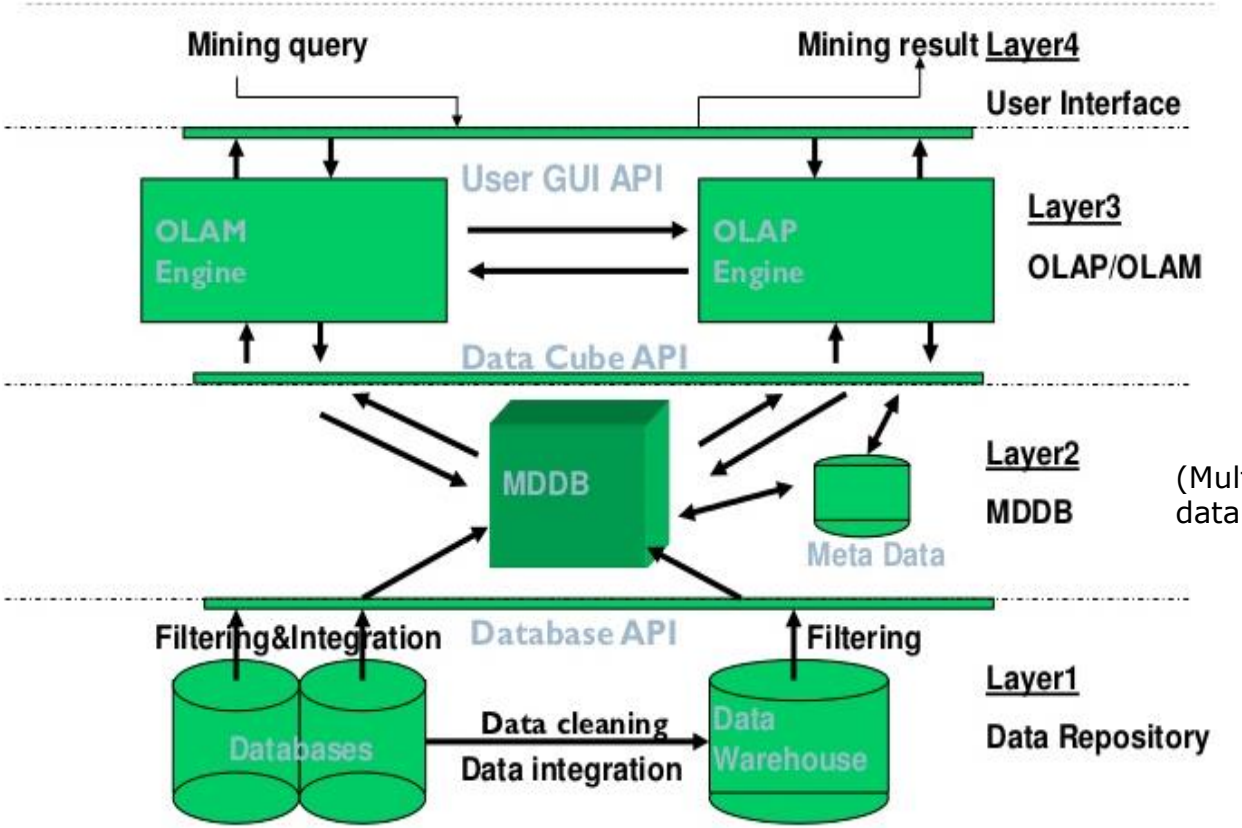
Intrebare pentru voi!

- Care sunt diferentele intre Data mining si Machine learning?
- Exemple

DM in DW

- **Volume f. mari de date** – milioane de inregistrari, mii de atribute
- Se realizeaza **procesul ETL** si se incarca si gestioneaza datele in sistem multidimensional
- Se ofera acces utilizatorilor de business care isi vor realiza **analizele dorite** prin aplicatii software specifice
- **Rezultatele** sunt prezentate sub forma de tabele sau grafice

Arhitectura sistem OLAM



Aplicatii DM

- **AT&T** utilizeaza o aplicatie de data mining pentru identificarea *apelurilor internationale frauduloase;*
- sistemul american **FAIS** (Financial Crimes Enforcement Network AI System) utilizeaza data mining pentru identificarea activitatilor de spalare a banilor in cadrul tranzactiilor foarte mari de bani;
- **Banca Americii** utilizeaza data mining pentru identificarea clientilor care utilizeaza anumite produse ale bancii si care sunt produsele preferate ale clientilor, in scopul crearii de mixuri de produse care sa satisfaca exigentele clientilor.
- **US West Communications**, furnizor de servicii de comunicatii cu peste 25 milioane de clienti, utilizeaza data mining pentru a determina tendintele si nevoile clientilor pe baza unor parametri de tipul: dimensiunea familiei, varsta medie a membrilor familiei si adresa de rezidenta.
- **Twentieth Century Fox** analizeaza incasarile de box-office pentru a identifica care actori, filme si scenarii vor fi apreciate in diverse arii de marketing.

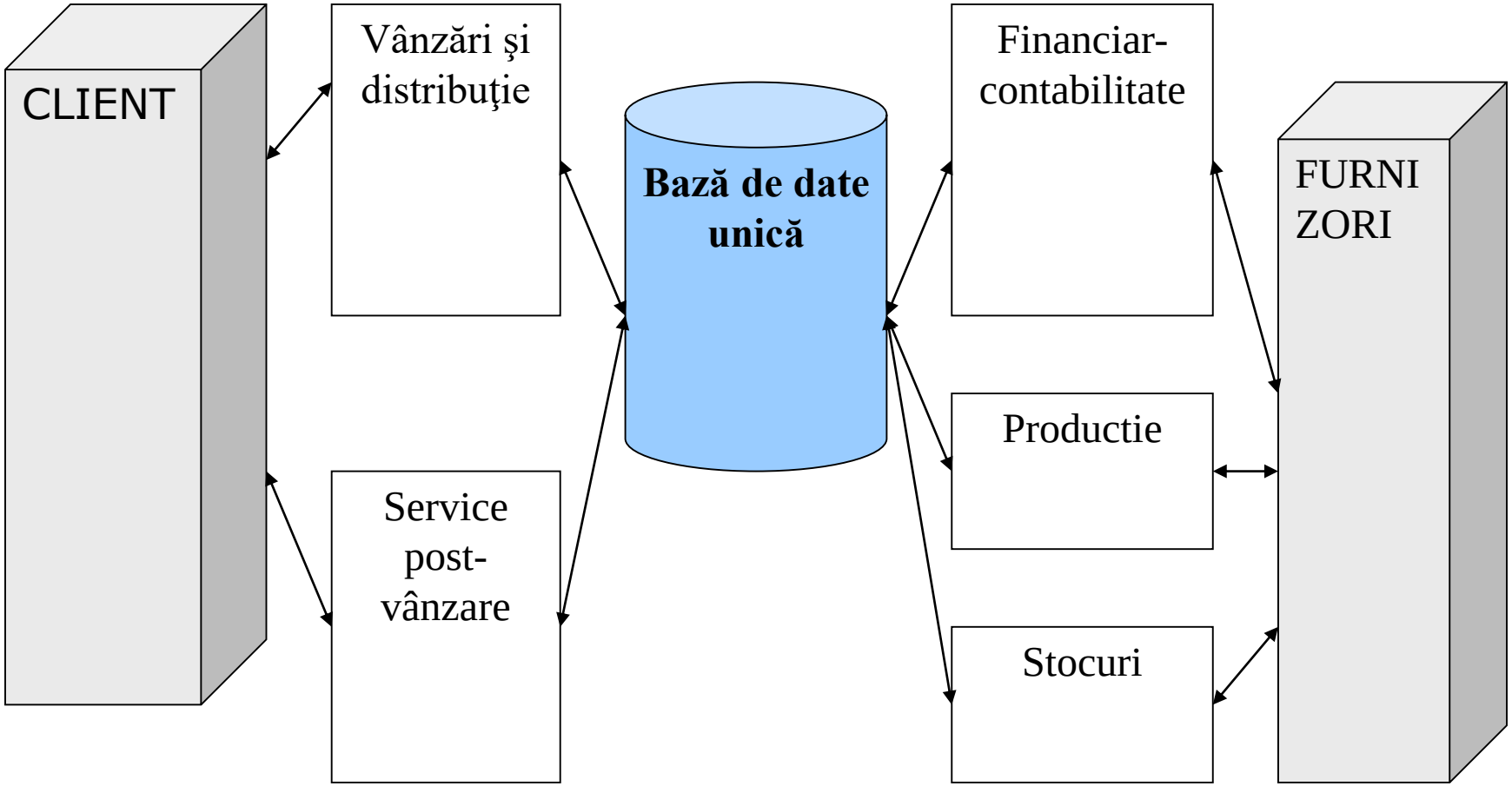
3. Integrarea datelor

- BI si ERP;
- Descrierea unui sistem ERP (SAP) integrat cu software BI

Ce este un sistem ERP

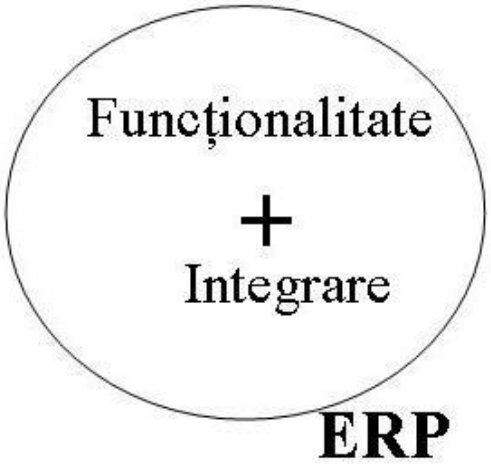
- "un pachet care promite integrarea completă a tuturor informaţiilor din cadrul unei organizaţii" [Davenport]
- "infrastructură software, multimodulara ce oferă suport de gestiune şi coordonare a diferitelor structuri şi procese din companie, în vederea realizării obiectivelor de afaceri" [Fotache]
- Ofera accesabilitate, vizibilitate şi consistenţa informaţională în întreaga organizaţie
- Dezvoltare cu instrumente CASE

Arhitectura client-server



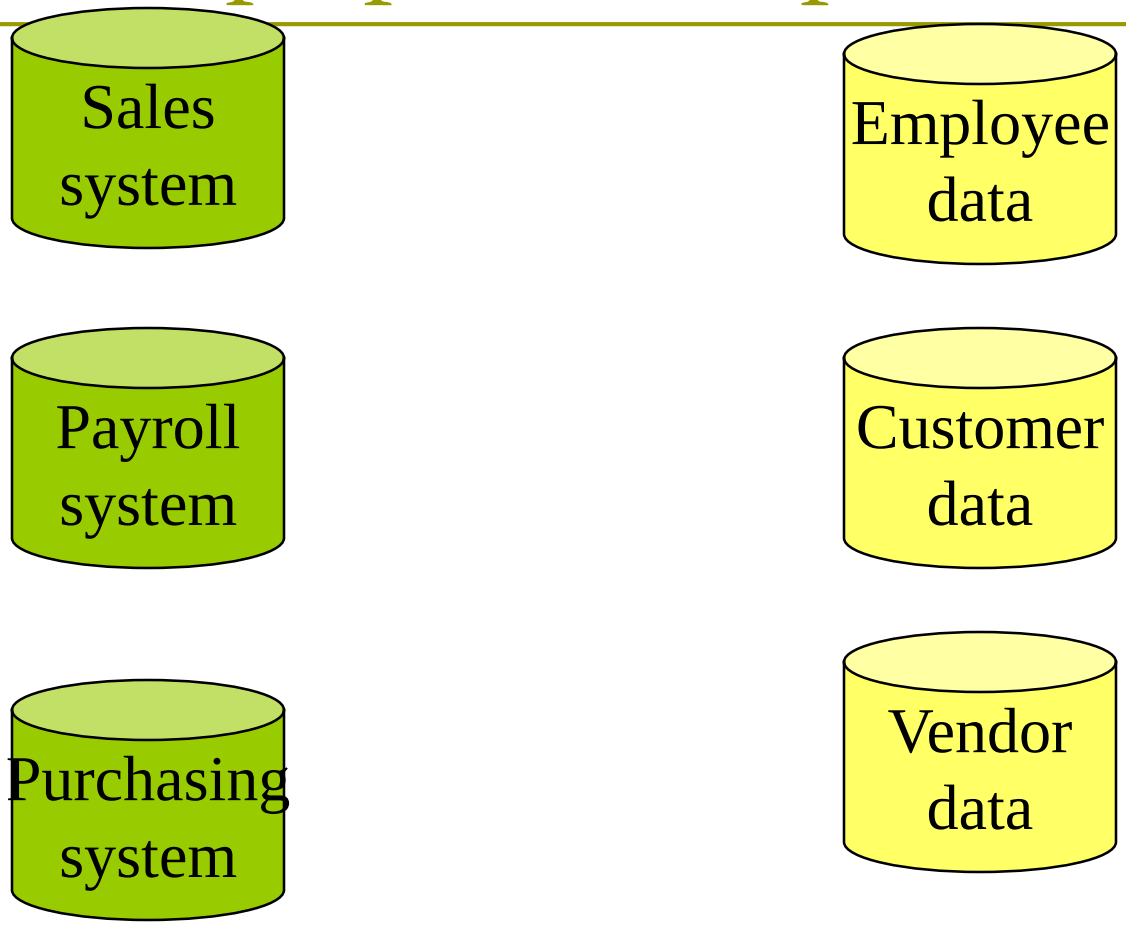
Proprietati fundamentale

- **Integrarea** asigură conectivitatea între fluxurile de procese economice funcţionale



- **Functionalitatea** a unui sistem ERP asigură fluxurile de procese economice din cadrul fiecărei funcţiuni

Orientare pe procese/ pe subiecte



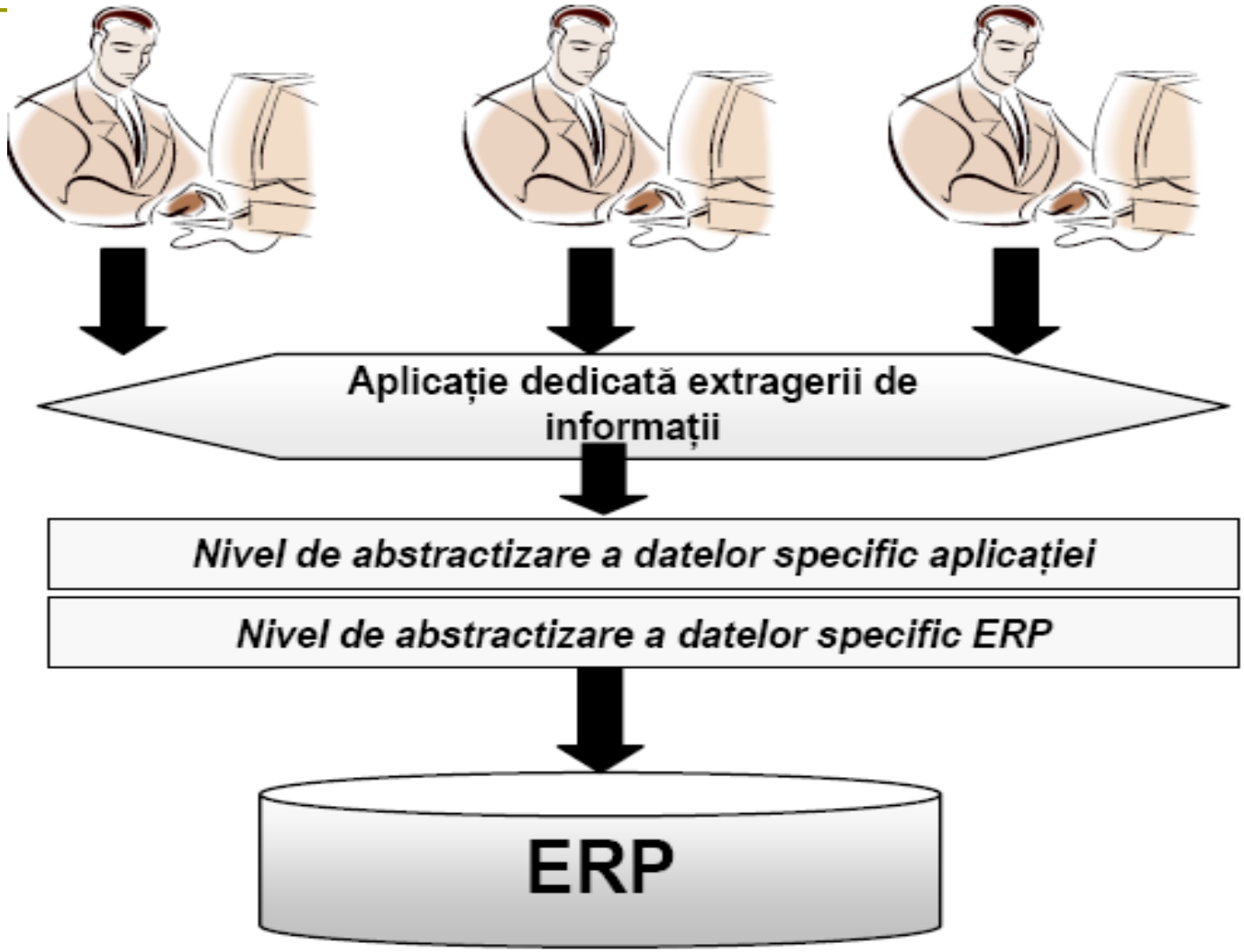
BI si ERP

- ERP orientarea pe **procese economice**
- DW orientarea pe **subiecte**
- ERP -BD unica, imensa, cu mii de tabele, care nu se preteaza pentru interogari ad-hoc si analize complexe
- ERP – avantaj pentru proiectarea şi implementarea DW
  - **omogenitatea** sistemelor sursă şi, implicit,
  - modalităţi mult mai facile de **achiziţie** a datelor şi de **asigurare a calităţii**
  - posibilitatea **consolidării datelor** la nivel de companie în cazul firmelor cu mai multe filiale

a.Arhitecturi: Sistem BI cu acces direct la datele din sistemul ERP

- integrat prin intermediul unor aplicaţii specifice de interogare a datelor.
- suprapun peste primul nivel de abstractizare al modelului ERP un nivel de abstractizare propriu, specific fiecărui utilizator
- sunt realizate interfeţe dedicate fiecărui modul din sistemul integrat.
- **Dezavantaje**
  - limitele impuse de suporturile tehnice.
  - viziunea istorică se suprapune rareori cu necesităţile sistemelor tranzacţionale
  - este o soluţie de compromis ce poată fi exploatată temporar

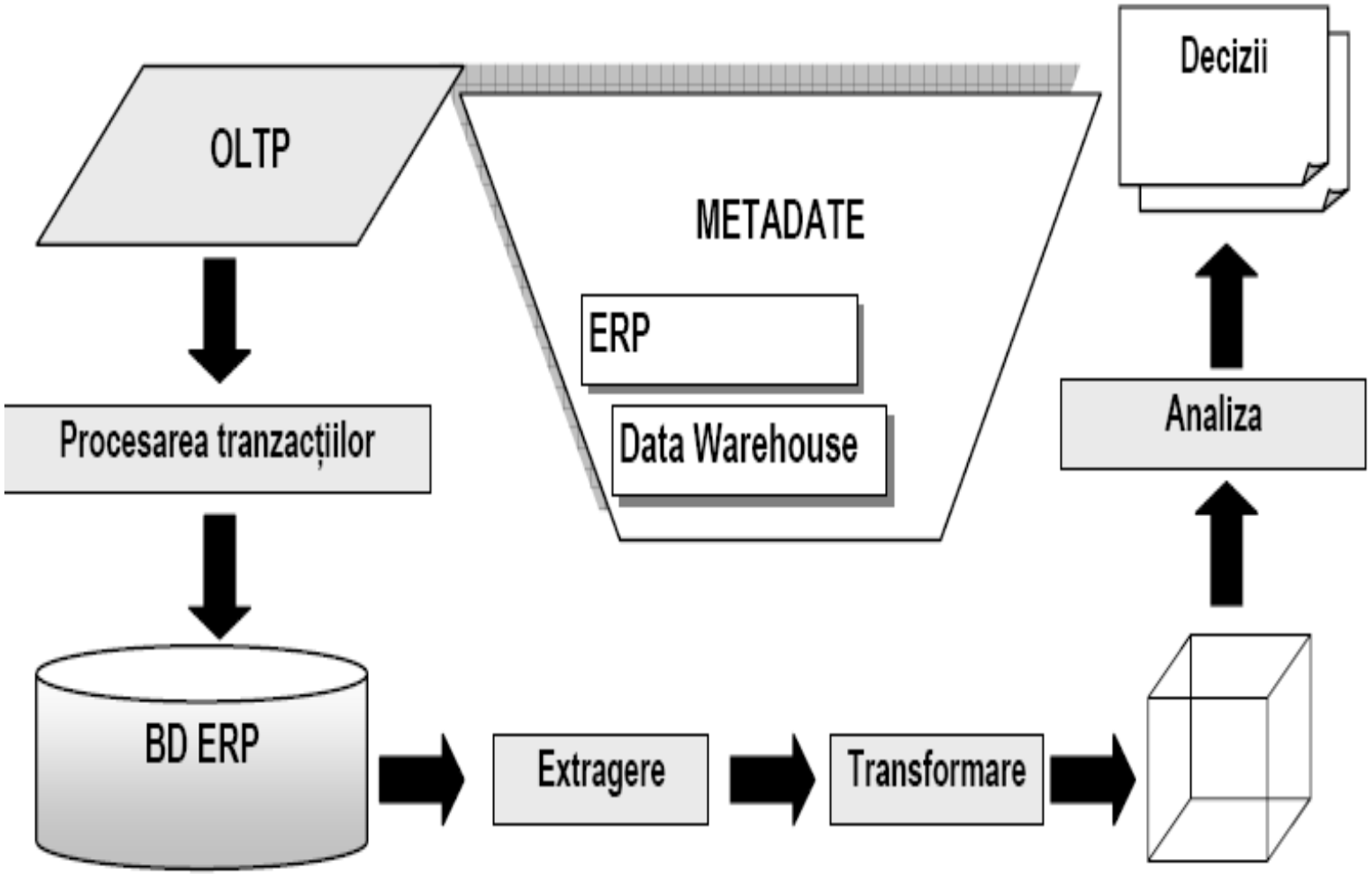
Acces direct la datele ERP



b.Arhitecturi: Depozit de date ataşat ERP

- sistem de asistare a deciziei **specializat**, construit pe baza unui **depozit de date** sau a unei colecţii de **data marts**.
- **dicţionar de date propriu**
- ca aplicaţie independentă sau ca un modul al ERP (SAP BI)
- eforturi considerabil mai mari atât în etapele de proiectare si implementare => avantaje prin prisma performanţelor în exploatare

Depozit de date ataşat ERP



SAP BI (Business Information Warehouse)

- **Business Content** = container ce cuprinde
  - Infocuburi (peste 420),
  - Query-uri (peste 1700),
  - Rapoarte si
  - Roluri utilizator

  cu specific industrial si functional= solutii preconfigurate pentru diferite industrii
- **Extractori** („plug-in“) - extragerea datelor din SAP ERP si incarcarea in SAP BW
  - complet (full extraction) sau
  - partial (delta extraction).