Iceland Liechtenstein Norway grants

Data Processing
Data mining

Integrated Applied Genetics Training
Calin Poenaru, MEng, PhD

Analiza datelor – Definitie

- Analiza datelor este procesul de inspectie, “curatate”, transformare si modelare a datelor in scopul descoperirii de informatie utila in stabilirea de concluzii sau pentru suportul deciziilor;
- Data reprezintă o înşiruire de caractere numerice sau alfa numerice, cu o anumită semnificaţie;
- Informaţiile se obţin în general din prelucrarea datelor (nu se confundă însă cu acestea).

Source: Wikipedia

Datele furnizeaza informatii

- Datele “bune” pot fi analizate si apoi sumarizate pentru a furniza informatii utile ☺
- Datele “proaste” pot fi analizate si apoi sumarizate pentru a furniza informatii incorecte/nocive/ non-informative, care conduc la…? ☹

Intrebari la care statistica poate da un raspuns

- Exprimând caracteristicile unui numar de indivizi ai unei populatii prin variabile (unele numerice, altele nu), admitem că prin măsurare sau evaluare vom obţine seturi de date cu care vom construi tabele de date.
- Întrebările esenţiale care se pun de obicei sunt următoarele:
  - cum putem să descriem „sintetic” datele pe care le-am obţinut?
  - cum putem să transmitem altora informaţiile pertinente despre ansamblul indivizilor, fără însă a le transmite toate datele obţinute?

Componente analiza de date – EDA

- Exploratory Data Analysis (EDA) a.k.a. Statistica descriptiva:
  - Descrierea cantitativa a proprietatilor unui set de informatii, ducind la sumarizarea esantionului in cauza
  - Scopul principal este descoperirea de noi proprietati ale esantionului in vederea verificarii prezumtiilor
  - Utilizeaza metode grafice de vizualizare a datelor

Metode EDA

- Ordonarea – diagrame stemplot (stem-and-leaf)
- Grupare – histograme
- Sumarizare – diagrame boxplot

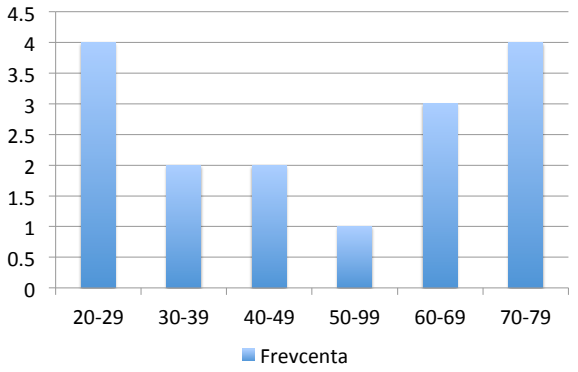
Metode EDA – Ordonare

- Reprezinta baza pentru evaluarea distributiei de frecventa
- Exemplu: 23, 24, 24, 25, 32, 36, 45, 47, 51, 61, 62, 67, 73, 76, 78, 78

Virsta	Frecventa	Cumulativ	Relativ	Observatii
20-29	4	4	0.25	3, 4, 4, 5
30-39	2	6	0.125	2, 6
40-49	2	8	0.125	5, 7
50-99	1	9	0.0625	1
60-69	3	12	0.1875	1, 2, 7
70-79	4	16	0.25	3, 6, 8, 8

Metode EDA – Grupare

- Reprezinta grafic distributia datelor
- Potrivita pentru observatii numeroase



Metode EDA – Sumarizare

- Percentila r: este valoarea care este mai mare sau egala cu r% din cele **n** observatii (sau mai mica sau egala cu 100-r%)
- Quartila: divide un sir ordonat in 4 grupuri

Quartila	Percentila	Mod de calcul
Q1	P25	(n+1)*1/4
Q2 (mediana)	P50	(n+1)*1/2
Q3	P75	(n+1)*3/4

Metode EDA – Sumarizare

- Boxplot (box-and-wisker)
  - reprezinta grafic quartile
- Elemente:
  - Upper hinge = Q3
  - Lower hinge = Q1
  - IQR = Q3 – Q1
    - contine 50% din observatii
  - Upper fence = Upper hinge + 1.5*IQR
  - Lower fence = Lower hinge – 1.5*IQR
- Datele din afara “fence” se numesc “outliers”

Componente analiza de date – CDA

- Confirmatory Data Analysis (CDA) a.k.a. Inferential Statistics
  - Deducerea proprietatilor unei populatii prin analiza datelor oferite de un esantion
    - Populatie: o colectie de observatii
    - Esantion: un subset al unei populatii
  - Propune si testeaza ipoteze asociate cu intreaga populatie, nu doar cu esantionul masurat
  - Utilizeaza modele probabilistice

Componente analiza de date – CDA

Populatie

- N = marimea populatiei
- μ = media (masura a centrului distributiei)
- σ^2 = variatia (masura a dispersiei)
- σ = deviatia standard

Esantion: subgrup al populatiei utilizat pentru a calcula estimari (statistici) care aproximeaza parametrii populatiei

- n = Marimea esantionului
- $\bar{x}$ = media esantionului
- s^2= variatia esantionului
- s= deviatia standard a esantionului

- Populatie: descrisa numeric de parametri
- Esantion: descris numeric de statistici

Ce insemna “Data”?

- O colectie de obiecte si atributele lor
- Un atribut este o proprietate sau caracteristica a unui obiect, iar o colectie de atribute descrie (complet) un obiect
  - Exemple:
    - Culoare ochilor unei persoane
    - Temperatura
    - Cheltuieli anuale
- Atributele sint cunoscute ca si variabile, cimpuri, caracteristici, etc.
- Obiectele sint cunoscute ca inregistrari, puncte, cazuri, esantioane, entitati sau instante

Atribute

Obiecte

Tid	Sex	Starea civila	Limita asigurare	Vaccinat
1	M	Single	125K	Nu
2	F	Married	100K	Nu
3	F	Single	70K	Nu
4	M	Married	120K	Nu
5	F	Divorced	95K	Da
6	F	Married	60K	Nu
7	M	Divorced	220K	Nu
8	F	Single	85K	Da
9	F	Married	75K	Nu
10	F	Single	90K	Da

Valorile atributelor

- Valorile atributelor sint numere sau simboluri asignate unui atribut
- Acelasi atribut poate avea valori diferite (ex. temperatura masurata in °C sau in °F)
- Mai multe atribute diferite pot sa aiba aceeasi valoare (ex. virsta si codul postal sint intregi pozitivi), dar proprietatile asociate valorii atributului pot fi diferite (ex. virsta are valori discrete succesive pina la o valoare maxima, codul postal are valori discrete non-consecutive si non-limitative)

Tipuri de atribute (variabile)

Ce mai intalnita clasificare:

- Calitative - apar atunci când indivizii aparțin/pot fi clasificati in clase separate.
  - Nominale (categoriale)
  - Ordinale
- Cantitative (numerice) - sunt obținute fie prin numărare fie sunt rezultatul unei măsurători.
  - De tip interval
  - De tip raport

Data Mining – Definitie

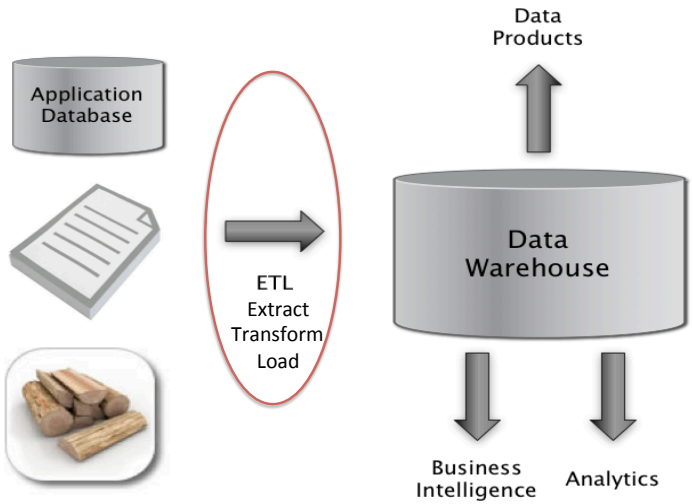
- **Data Mining** = o arie interdisciplinara a Computer Science (CS) care are drept scop descoperirea modelelor din seturi mari de date prin extragerea informatiilor dintr-un set de date si transformarea lor intr-o structura definita utilizabila in viitor
- Implica metode la intersectia statisticii, AI, Machine Learning si bazelor de date
- Este deseori identificata (abuziv) cu orice procesare de cantitati mari de date

Provocari pentru Data Mining

- Scalabilitate
- Dimensionalitate
- Complexitate si eterogenitate
- Calitatea datelor
- Proprietatea asupra datelor si distribuirii lor
- Pastrarea confidentialitatii (privacy)
- Date in flux continuu (streaming data)

DATA PREPARATION & CLEANING

The Big Picture



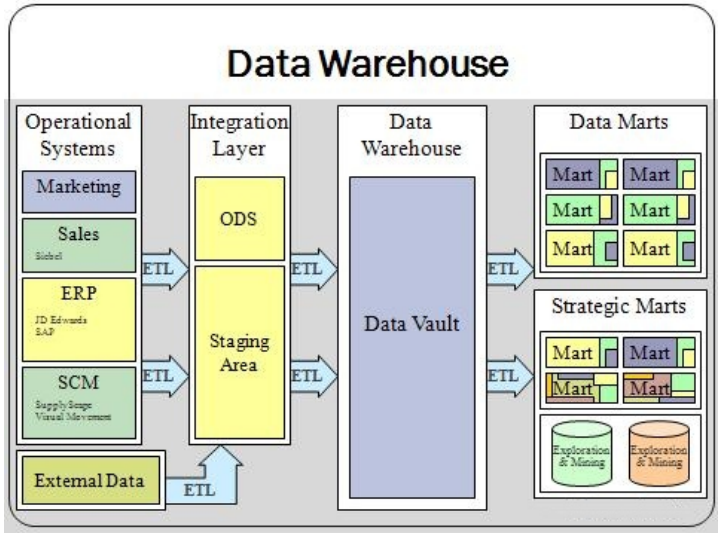
Terminologie suplimentara

Data Warehouse:
- Un sistem utilizat pentru raportare și analiză a datelor.
- DW sunt zone centrale de depozitare de date integrate din una sau mai multe surse diferite
- Datele sunt aduse de la surse operationale (OLTP) într-un singur "depozit" pentru analiză (OLAP)

Data Mart:
- O formă simplă de DW, care este axat pe un singur subiect sau o singura zonă funcțională

Decision Support System (DSS):

Infrastructura pentru analiza datelor si luarea deciziilor
- DW dedicat pentru datele istorice
- DW dedicat pentru predicție



Pregatirea Datelor

- ETL
  - Datele trebuie **extrase (Extract)** de la **sursa(e)**
  - Datele trebuie **transformate (Transform)** la sursa sau destinatie sau chiar in **staging area**
  - Datele trebuie **incarcate (Load)** la **destinatie**
  - Surse: fisiere, database, event log, web site, HDFS…
  - Destinatie: Python, R, SQLite, RDBMS, NoSQL store, fisiere, HDFS…

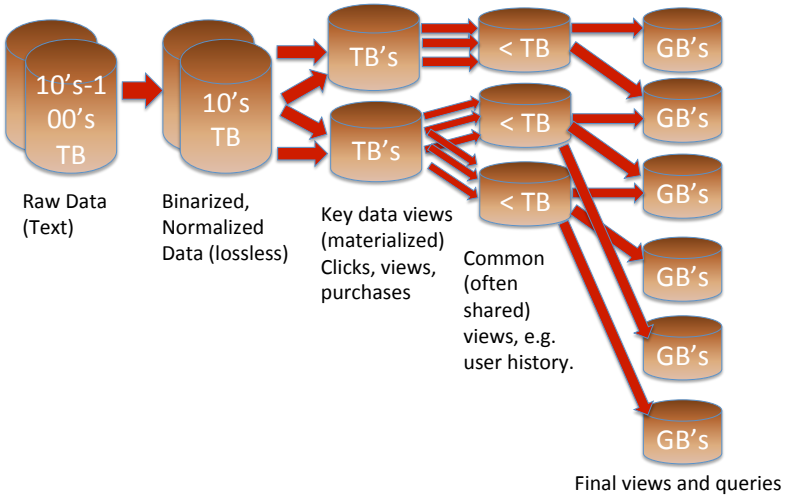
Data Preparation overview

- Abordare din punct de vedere al procesului:
  - Procesul consta in mai multe faze:
    - Caracterizarea datelor
    - Curatarea datelor
    - Integrarea datelor din surse multiple
  - Scopul il constituie eficientizarea (in spatiu si timp) a operatiilor de transfer de date, inclusiv serializare/deserializare (la viteza de 5MB/s a unui disc e nevoie de ~6 ore pentru citirea 1TB)

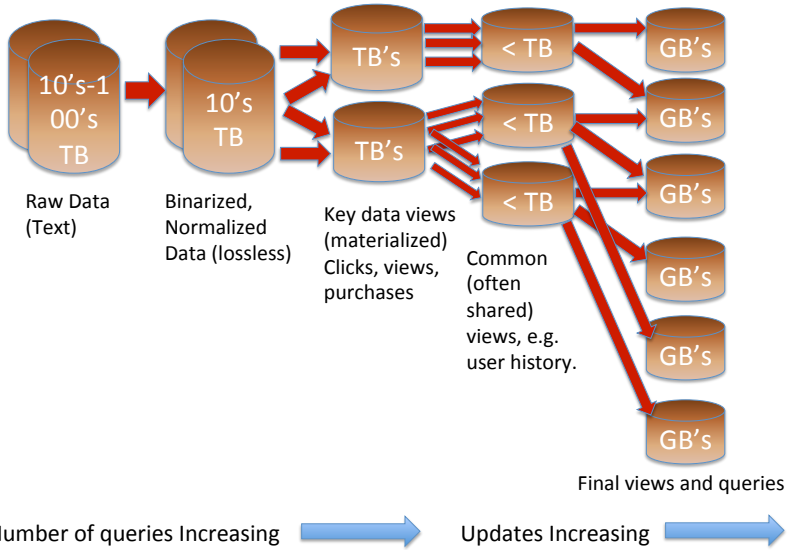
Data Preparation overview

- Implementarea pasilor procesului duce la un “workflow”
- Daca workflow-ul trebuie/poate fi folosit de mai multe ori atunci el poate fi “scheduled” (programat)
- Programarea poate fi facuta:
  - in functie de un moment de timp
  - asociata aparitiei unui eveniment
- Inregistrarea executiei workflow-ului se mai numeste “capturing **lineage** or **provenance”**

Data Pipeline Design



Data Pipeline Design



Data Lakes

Dirty Data

- Punctul de vedere al **Statisticii**:
  - Datele sint produse de un proces
  - Modelarea de esantioane ideale rezultate din proces este imposibila:
    - Distorsiuni – esantioane corupte de proces
    - Selectare distorsionata (bias) – probabilitate ca esantionul sa depinda de valorile sale
    - Dependenta – esantioanele nu sint complet independente
    - Variabilitate – elementele urmarite (ex. pacienti) intra si ies din studiu pe durata acestuia
  - Procesul de prelucrare poate imbunatati rezultatele prin adaugarea de noi modele
  - Dar… nu se pot modela toate tipurile de imperfectiuni
  - Trebuie gasit un echilibru intre acuratete si simplitate (uneori chiar realizabilitate)

Dirty Data

- Punctul de vedere al **Database (CS)**:
  - Acestea sint datele pe care le am
  - Unele din ele nu corespund (sint lipsa, corupte, gresite, duplicate)
  - Rezultatele obtinute in prelucrare sint absolute, determinate de modelul relational
  - Singura modalitate de a avea un raspuns mai bun este imbunatatirea calitatii datelor de intrare

Dirty Data

- Punctul de vedere al Expertului (Domani Knowledge):
  - Este formulat in raport cu un model implicit de date ce este deja asumat (expertiza)
  - Datele nu arata bine → rezultatul obtinut nu poate fi corect
  - Rezultatul nu arata bine → datele initiale nu pot fi corecte

Probleme de calitate a datelor

- La sursa datele pot sa fie implicit eronate (“dirty”).
- Transformarile complexe pot corupe datele (ex. rotunjiri succesive)
- Date eronate pot rezulta din Integrarea datelor din surse curate (“clean”) multiple
- Erori “rare” pot deveni frecvente dupa transformare si integrare
- Datele vechi isi pot pierde in timp precizia (“data/bit rot”)
- Combinatii multiple ale factorilor de mai sus

Surse de “Dirty Data”

- Impartirea textului (parsing) in cimpuri (probleme de separator)
- Conventii de denumire (ex NYC vs New York)
- Lipsa unui cimp necesar (e.g. key field)
- Reprezentare diferita (2 vs Doi)
- Trunchierea unui cimp prea lung
- Inregistrari redundante (exact sau partial)
- Probleme de formatare – in special la reprezentarea datei
- Probleme de acces la date (ex. licentiere/date private)
- …

Sursa: Stanford Data Integration Course

Definirea conventionala a calitatii datelor

- Acuratete
  - Datele au fost inregistrate corect
- Completitudine
  - Toate datele relevante au fost colectate
- Unicitate
  - Fiecare entitate este inregistrata doar o singura data
- Actualitate (Timeliness)
  - Datele sint pastrate in cea mai noua forma (Special problems in federated data: time consistency)
- Consistenta
  - Datele nu au discrepante (data agrees with itself).

Sursa:Ted Johnson's SIGMOD 2003

Probleme

- Calitatea datelor este greu masurabila/ne-masurabila
  - Acuratetea si completitudinea sint foarte greu/imposibil de masurat
- Calitatea datelor (definita prin ceea ce e important) este data de context
  - Precizia e data de scopul prelucrarii (ex. cantitatea de sare din mincarea pasagerilor in perspectiva greutatii avionului)
  - Calculul valorilor agregate tolereaza lipsa punctuala de precizie (ex. consumul de energie electrica)
- Masurarea calitatii este/va fi intotdeauna incompleta
  - Ce parere aveti de metrici ca: interpretabilitatea, disponibilitatea, accesibilitatea, calitatea metadatelor, etc.
- Termenul este vag
  - Nu exista un mod clar in care definitia poate fi actualizata sau imbunatatita in functie de necesitati/utilizare

Sursa:Ted Johnson's SIGMOD 2003

Ce inseamna "Data Quality"

- Exista multe tipuri de date care au utilizari diferite si probleme de calitate diferite:
  - Federated data
  - High dimensional data
  - Descriptive data
  - Longitudinal data
  - Streaming data
  - Web (scraped) data
  - Numeric vs. categorical vs. text data
- Trebuie inteles **cum** si **unde** apar problemele de calitate, deci **regulile** din spatele proceselor

Continuumul calitatii datelor

- Datele si informatiile nu au caracter static, ci urmaresc un proces de la colectare la utilizare:
  - Colectare de date
  - Livrare de date
  - Stocare de date
  - Integrare de date
  - Regasire (retrieval)
  - Data mining/ analiza (analysis)

Metrici pentru Data Quality

- Metricile ajuta la obiectivarea conceptului de “calitate” prin asocierea unor cantitati masurabile:
  - Indica ce e bine si ce e gresit si permite imbunatatirea
  - Realizeaza magnitudinea problemei si impiedica over-engineering-ul
- Tipuri de metrici:
  - Statice vs. dinamice
  - Operationale vs. diagnosticare
- Aplicarea metricilor trebuie sa incurajeze imbunatatirile in utilizarea datelor, nu invers (ex. ignorarea/aruncarea datelor ce nu se potrivesc modelului)
- Deoarece se poate crea un numar foarte mare de metrici, trebuie selectat doar numarul minim ce duce la indeplinirea criteriilor de calitate

Modele de metrici

- Conformanta cu schema
  - Masoara constringerile asociate unui snapshot.
- Conformanta cu “business rules”
  - Masoara constringerile asociate schimbarilor in database.
- Acuratetea
  - Utilizeaza variante cu cost mai mic pentru evaluarea datelor (ex. evaluarea doar a unor esantioane, masurarea numarului de erori de prelucrare, etc.)
- Accesibilitatea
- Interpretabilitatea
- Numarul de probleme ridicate la analiza
- Completarea cu succes a intregului proces (end-to-end)

Abordare pragmatica

- E nevoie de o abordare multidisciplinara in evaluarea calitatii datelor:
  1. Gestiunea proceselor si a procedurilor
  2. Metode statistice in detectarea anomaliilor
  3. Database pentru repararea defectelor si asigurarea consistentei datelor si relatiilor
  4. Expertiza pentru asigurarea interpretabilitatii datelor (datele sint utile scopului propus)

OLAP

- On-Line Analytical Processing (OLAP) este o abordare specifica analizei multi-dimensionale a datelor
- Pentru OLTP (bazele de date relationale) aranjeaza datele in tabele in timp ce pentru OLAP acestea sint reprezentate sub forma unor arii multidimensionale
- Aceast mod de aranjare face ca operatiile de analiza si explorare a datelor sa fie mai usor de facut

Crearea unei arii multidimensionale

- Conversia unui tabel de date intr-o arie multidimensionala implica 2 pasi:
  1. Identificarea acelor atribute care vor fi dimensiunile (valori discrete) si a aceluia care este tinta analizei (valori continue, sau rezultatul unei operatii: sum, count, etc).
     - Atributul tinta poate sa fie numarul obiectelor cu aceleasi valori ale atributelor
  2. Gasirea valorilor pentru fiecare intrare din aria multidimensionala prin insumarea valorilor atributului "target" (tinta), sau numararea tuturor obiectelor care au valorile atributelor corespunzatoare acelei intrari

OLAP Operations: Data Cube

- Operatia cheie in OLAP este formarea "data cube" (aria multidimensionala impreuna cu toate posibilele agregate)
- "Posibilele agregate": agregatele care rezulta din selectarea unui posibil subset de dimensiuni si sumarea tuturor celorlalte dimensiuni ramase
- Exemplu: selectind pentru o dimensiune atributul "specie" si sumind celelalte dimensiuni, rezultatul este o arie unidimensionala cu 3 intrari, fiecare reprezentind numarul florilor din fiecare tip

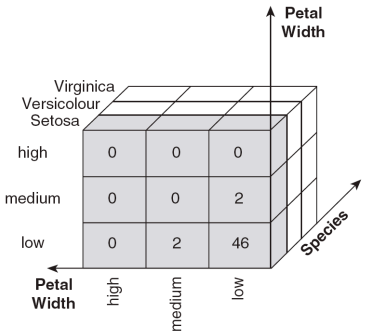
Exemplu: Iris data

1. Discretizam latimea si lungimea petalelor ca sa avem date categoriale: *low, medium, large*
2. Procedind, asa cum am descris anterior, avem tabelul (notati atributul "count"):

Petal Length	Petal Width	Species Type	Count
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44

Exemplu: Iris data (cont)

- Fiecare grupare (tuple) unica de *<l_petala, L_petala, specie>* identifica un element al ariei
- Acestui element ii este asignata valoarea corespunzatoare rezultata din aplicarea "count"
- Pentru toate gruparile inexistente/nespecificate valoarea asignata este 0



Example: Iris data (cont)

"Slices" efectuate dupa dimensiunea "varietate", sint prezentate in tabele de mai jos:

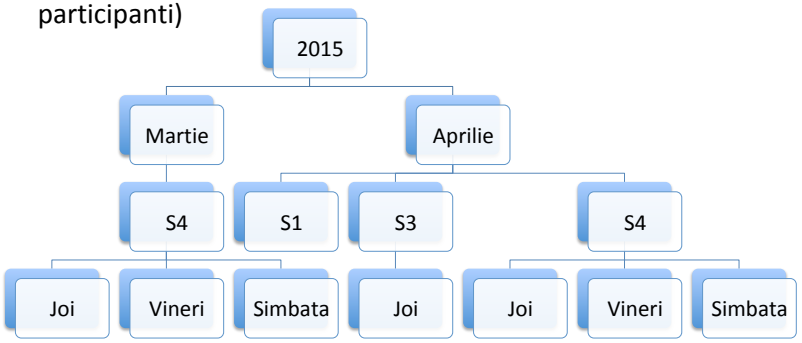
		Width		
		low	medium	high
Length	low	46	2	0
	medium	2	0	0
	high	0	0	0

		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	43	3
	high	0	2	2

		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	0	3
	high	0	3	44

Operatii OLAP: Roll-up and Drill-down

- Atributele au de cele mai multe ori o structura ierarhica:
  - Data: an/luna/zi
  - Locatia geografica: continent/tara/oras/strada/numar/cod
- Aceste categorii pot fi reprezentate ca arbori sau latice (au un element majorant – supremum – si unul minorant – infimum)
- Exemplul de mai jos se refera la acest curs (target = numar de participanti)



Operatii OLAP: Roll-up and Drill-down

- Aceasta structura ierarhica permite "roll-up" (agregare) si "drill-down" (detaliere)
- Pentru proiectul CERO trebuie raportata prezenta pentru fiecare modul de curs:
  - Roll-up: prezenta generala la cursurile modulului de Informatica Medicala (sau prezenta la curs a cursantilor din afara Bucurestiului)
  - Drill-down: prezenta pentru fiecare instanta a cursului de IM (sau prezenta la laboratoarele de IM)
- Operatiile de roll-up/drill-down sint relative la atributul central (fact) existent:
  - in cazul exemplului atributele centrale sint reprezentate de prezenta la fiecare modul: XX la IM, YY la Biostatistica, etc, deoarece acestea dau criteriile de succes . Restul detaliilor sint dimensiuni ale acestora (vezi star-schema)
- Operatiunile de roll-up/drill-down se pot face dupa orice atribut existent, in functie de interes

Data Cube – examplu (cont)

Tabelul de mai jos prezinta una din posibilele agregari bi-dimensionale (dupa locatie) si agregarea zero-dimensionala (total general)

		date				
		Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	total
product ID	1	$1,001	$987	...	$891	$370,000
	⋮	⋮			⋮	⋮
	27	$10,265	$10,225	...	$9,325	$3,800,020
	⋮	⋮			⋮	⋮
	total	$527,362	$532,953	...	$631,221	$227,352,127