Tehnologii informatice de integrare a datelor

Curs 5

Martie 2018

Tehnologii de integrare

1. **Depozite de date**
2. Migrarea datelor
3. Baze de date distribuite

Business Intelligence

- Principala valoare = capacitatea de **a transforma datele în informaţii.**
- Cine nu obtine informaţii suficient de repede şi nu le prelucrează eficient, rămâne în urmă sau dispare, într-un mediu concurenţial din ce în ce mai agresiv.
- **Business Intelligence** se refera la sisteme informatice de identificare, extragere si analizare a datelor disponibile intr-o companie, sisteme al caror scop este **de a oferi un suport real pentru luarea deciziilor de business.**
- O soluţie de Business Intelligence integrează **datele curente** ale afacerii dar şi **date prealabile**, provenind din mai multe programe şi aplicaţii şi le consolideaza într-o singură **bază de date optimizată pentru regăsirea şi analiza informaţiei.**

Analiza datelor

- **Cerinte preliminare**: integrarea datelor
- **Analiza datelor**:
  - Inspectarea, curatarea, transformarea datelor pentru a extrage cunostinte utile
  - Transformarea datelor in informatii si oferirea de raspunsuri descriptive unor intrebari predefinite
- **Data mining**
  - Utilizeaza modelarea datelor pentru descoperirea cunostintelor
- **Business intelligence**
  - Se bazeaza pe analiza datelor si data mining
  - Transforma informatiile si cunostintele in actiuni inteligente
  - Se bazeaza pe diferite instrumente de analiza si pe inteligenta artificiala
- **Instrumente:**
  - Statistica, instrumente de modelare a datelor si de simulare

Niveluri ale BI

- **BI 1.0**
  - Implica mai ales manipularea si prezentarea datelor
  - Ofera o platforma care permite examinarea datelor pentru a oferi informatii necesare pentru luarea deciziilor
- **BI 2.0**
  - Acces in timp real
  - Analize profunde cu instrumente avansate cum sunt score card-uri, metrici KPI, cuburi, panouri de bord
  - Personalizare
- **BI 3.0**
  - Relevanta – adauga la big data informatii de context pentru a imbunatati relevanta analizelor
  - Big data – trateaza volume mari de date nestructurate
  - Flexibilitate ridicata – nu se mai folosesc cuburile preconstruite, ci se decide dinamic care sunt elementele necesare: cube-less BI
  - Sintetice, creative, raspund intrebarilor deschise

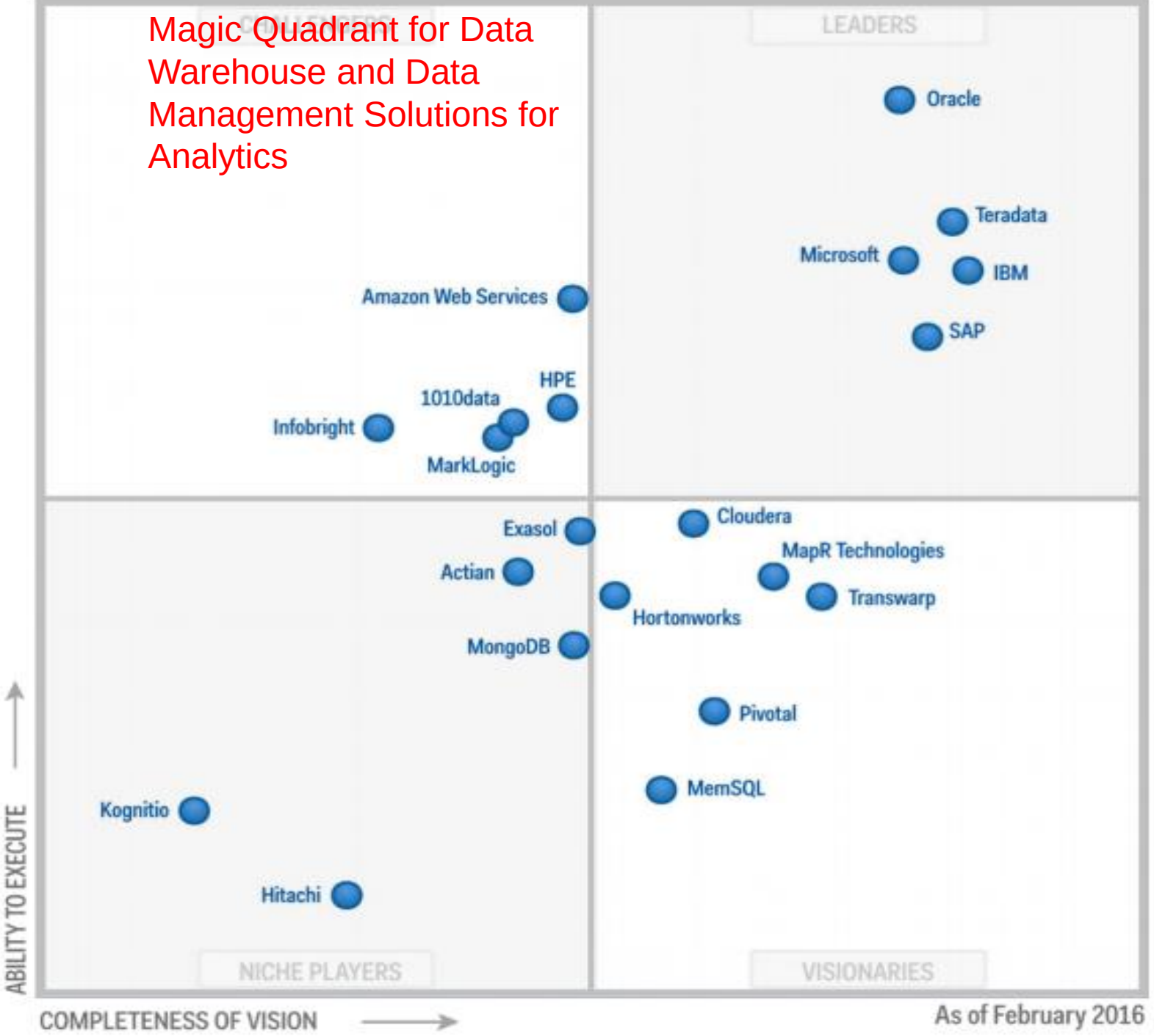
Depozite de date

- **Consiliul OLAP 1995**: o stocare centralizată a datelor detaliate provenite din toate sursele relevante din cadrul unei organizaţii ce permite interogarea dinamică şi analiza detaliată a tuturor informaţiilor.
- **William Inmon**: o colecţie de date orientate pe subiecte, integrate, istorice şi nevolatile destinată sprijinirii procesului de luare a deciziilor manageriale
- **+instrumente de interogare, analiza si prezentare a informatiilor**
  - Instrum. de analiza on-line- OLAP
  - Instrum. de data mining

In cifre…

- **Dimensiunea DW** – TBytes
- **Costul implementarii** – peste 1 mil $
  - Servicii profesionale
  - Software pentru extragere, transformarea, incarcarea si analiza datelor
  - Sisteme hardware si stocarea datelor
- Gartner Group estimează o creştere dublă pe piaţa depozitelor de date în raport cu creşterea globală a pieţei de IT

Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics



Scopul DW – integrarea datelor

- **Integrarea datelor**
  - modalităţi unice de codificare, sistem de unităţi de măsură consistente,
  - sistem stabil de reprezentare fizică a datelor,
  - convenţii clare privind modul de reprezentare a datelor calendaristice,
  - convenţii unice privind denumirile datelor.
- **FLEXIBILITATE** – sa se conecteze la niv. intregii organizatii a.i. servere de la furnizori diferiti sa se poata conecta la depozitul existent
- **ARHITECTURA** – adaptare usoara la modificarile de performante, capacitate si conectivitate
- **Data mart** - >500 GB, <1mil $, <3 luni

Ce este un depozit de date?

- **William Inmon:** este o colecţie de date orientate pe subiecte, integrate, istorice şi nevolatile, fiind destinat fundamentării deciziei manageriale.
- O BD pentru luarea deciziilor, separata de BD operationala a companiei
- Ofera suport pentru procesarea informatiilor, oferind o platforma de date istorice consolidate pentru analiza
- Structurile de date într-un depozit de date sunt optimizate pentru o regasire şi o analiza rapida.

a.DW – orientat pe subiecte (Inmon)

- Organizat pe subiecte importante: client, produs, vanzari.
- Accent pe modelarea si analiza datelor de catre decidenti
- Ofera o perspectiva simpla si concisa asupra anumitor subiecte, excluzand datele care nu sunt utile in procesul de luare a deciziilor

b.DW – integrat (Inmon)

- Integreaza surse de date multiple

- Tehnici de curatare si integrare a datelor.
- Consistenta in
  - conventiile de numire,
  - structura codurilor,
  - unitatile de masura folosite de diferitele surse
    - E.g., Pret hotel: moneda, taxe, mic dejun inclus, etc.

c.DW – istorice (Inmon)

- Datele sunt istorice şi sunt actualizate la intervale regulate.
- Orizontul de timp este mult mai mare decat la sist. operationale (ex: 5-10 ani)
- Fiecare element stuctural cheie al depozitului:
  - Contine o **referire temporala**, implicita sau explicita, ceea ce nu are loc la datele operationale

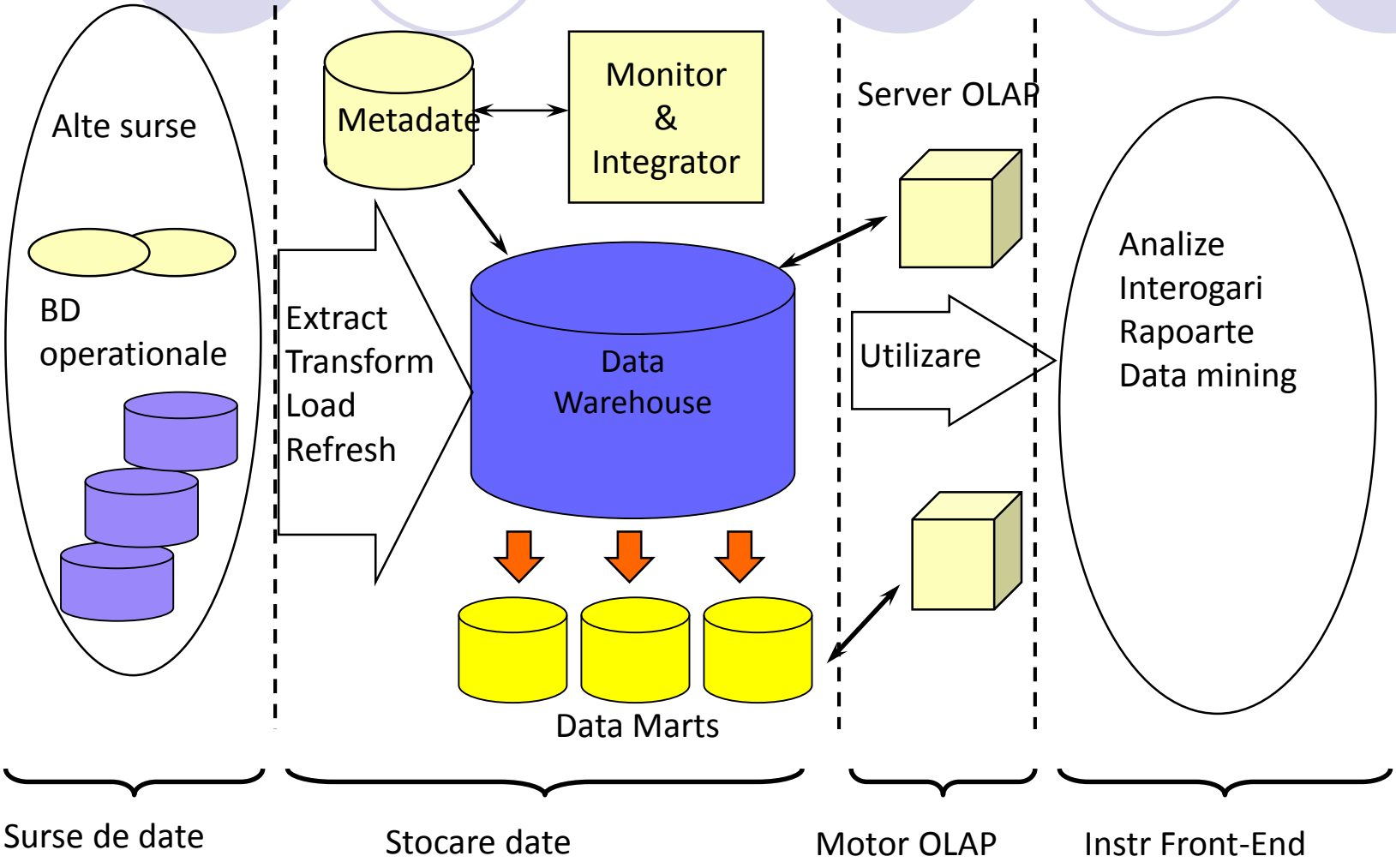
d.DW - nevolatil (Inmon)

- Un depozit separat fizic de date transformate din mediul operational
- In DW **nu au loc actualizari operationale ale datelor**.
  - Nu necesita mecanisme de procesarea tranzactiilor, recuperare si controlul concurentei
  - Sunt necesare doar 2 operatii pentru accesarea datelor :
    - **Incarcarea datelor si accesul la date**.
- **Actualizare** doar adăugarea periodică a unor date extrase din sistemele operationale
- Preocupare pt. **optimizarea accesului la date**: denormalizare, sumarizare, statistici ale accesării şi reorganizare dinamică a indexării

Aplicatii ale depozitelor de date

- **Telecomunicatiile**.
  - folosirea retelei,
  - profilul clientilor care folosesc un anumit serviciu,
  - profitabilitatea produselor si serviciilor oferite.
- **Bancile**
  - gestionarea profitabilitatii generale, prin analizarea profitabilitatii pe produs si pe client.
  - determinarea profilurilor clientilor pentru a directiona cât mai eficient campaniile de marketing.
- **Comertul cu amanuntul**.
  - trendul vânzarilor în functie de anotimp, vacante, campanii de publicitate, activitatea competitorilor.
  - mentalitatile si obiceiurile cumparatorilor = intrari in sistemul de dirijare a actiunilor promotionale si a altor campanii de marketing
  - analiza trendului performantelor
  - vânzari încrucisate
  - profilul consumatorului si piata tinta.

Architectura multinivel



De ce un depozit de date separat?

- **Performante mai bune**
  - **SGBD**— potrivit pt OLTP: metode de acces, indexari, controlul concurentei, recuperare.
  - **Depozit** —potrivit pt OLAP: cereri complexe, perspective multidimensionale, consolidare
- **Functii si date diferite**
  - Date: luarea deciziilor necesita date istorice
  - Consolidarea datelor: luarea deciziilor necesita consolidari de date din surse eterogene
  - Calitatea datelor: datele din surse diferite au reprezentari, codificari si formate diferite care trebuie reconciliate

Tipuri de DW

1 DEPOZITE DE ÎNTREPRINDERE (ENTERPRISE WAREHOUSE)

- întreaga structură organizaţională
- un volum extins de date: atât informaţii **detaliate**, cât şi **agregate**.
- suporturi hardware performante.
- costurile si timpul de proiectare şi implementare sunt considerabile,

2 DATA MART

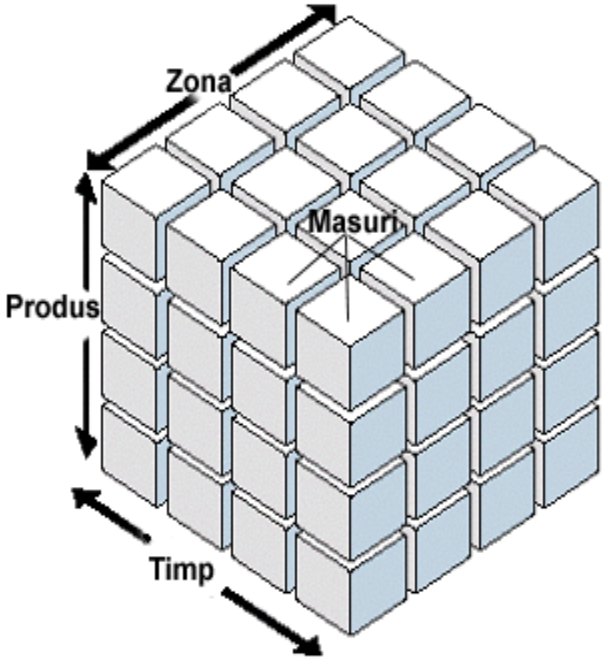
- specific unui anumit subset de cerinte sau unui departament din cadrul organizatiei
- de regulă, datele conţinute într-un data mart sunt **agregate**.
- costurile şi termenele de implementare sunt considerabil reduse

3 DEPOZIT VIRTUAL (VIRTUAL WAREHOUSE)

- o serie de vederi (*views*) realizate direct asupra BD operaţionale.
- procesele de agregare pot afecta capacităţile de prelucrare ale serverelor utilizate în activitatea operaţională,
- aparent uşor de implementat, necesita capacităţi de procesare deosebite.
- necesita curatare si consolidare in timpul rularii

Modelul multidimensional

- permite vizualizarea datelor prin mai multe filtre sau **dimensiuni** in acelasi timp.



- Dimensiuni=coordonate= categorii de informaţie.

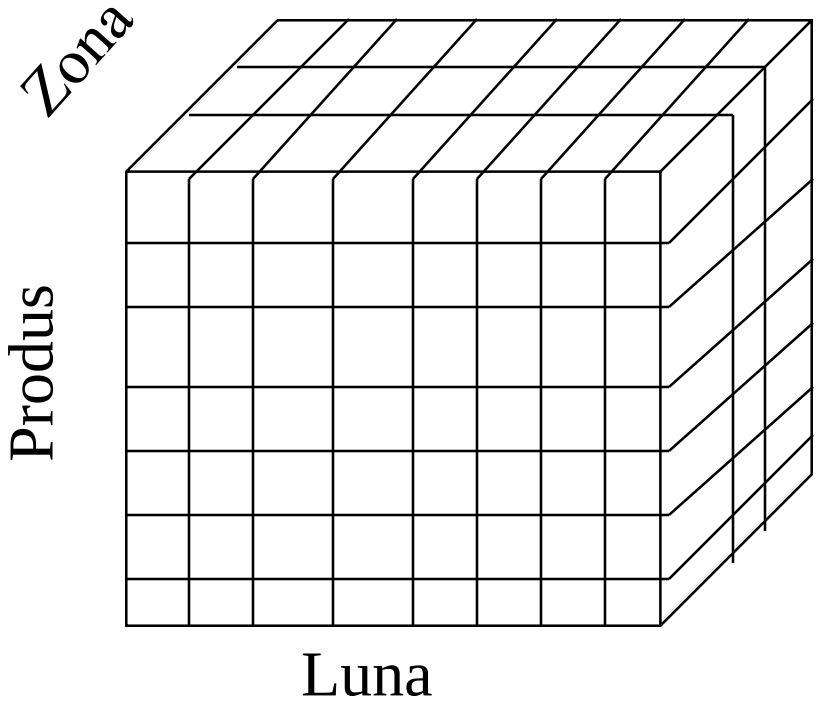
- De ex:
  - Care sunt vanzarile reale in comparatie cu cele previzionate pe zona, pe vanzator, pe produs?
  - Care este profitabilitatea pe produs, pe client?

Obiecte DW

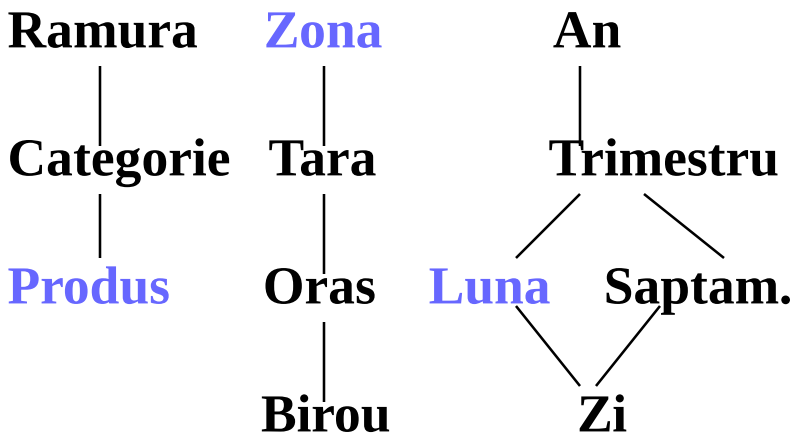
- **Tabelele de fapte** (masuri)
  - conţin faptele şi cheile externe către tabelele de dimensiuni.
  - de obicei date numerice - totalizate şi analizate pe diferite niveluri.
- **Tabele dimensiuni**
  - categorii de informatii care organizeaza datele
  - fiecare tabelă dimensiune are câte o cheie principala
  - câmpurile sunt de obicei textuale - **sursă pentru restricţii** şi pentru **rândurile din rapoarte**.
  - datele sunt de obicei colectate la nivelul cel mai de jos şi mai detaliat şi agregate pe nivelele superioare pentru analiză.
- **Atribut** - un nivel al unei dimensiuni, intr-o **IERARHIE**
- **Ierarhiile**
  - sunt structuri logice utilizate pentru ordonarea nivelelor de reprezentare a datelor.
  - definesc caile de navigare în interiorul datelor, permiţând detalierea graduală a datelor.

Date multidimesionale

- Volumul vanzarilor – functie de produs, luna, si zona



Dimensiuni: Produs, Zona, Timp
Ierarhii:



Exemplu: Vanzari de fructe

Timp	Suma
Trim 1	16000
Trim 2	16000
Total Timp	**32000**

Piata	Suma
Brasov	8000
Sibiu	8000
Arad	8000
Iasi	8000
Total Piata	**32000**

Produs	Suma
Mere	8000
Cirese	8000
Struguri	8000
Pepeni	8000
Total Produs	**32000**

		Brasov	Sibiu	Arad	Iasi	Total
Trim. 1	Mere	-	-	2500	1500	**4000**
	Cirese	-	-	2000	2000	**4000**
	Struguri	1000	3000	-	-	**4000**
	Pepeni	2000	2000	-	-	**4000**
	Total trim 1	**3000**	**5000**	**4500**	**3500**	**16000**
Trim 2	Mere	4000	-	-	-	**4000**
	Cirese	1000	3000	-	-	**4000**
	Struguri	-	-	1500	2500	**4000**
	Pepeni	-	-	2000	2000	**4000**
	Total trim 2	**5000**	**3000**	**3500**	**4500**	**16000**
	Total	**8000**	**8000**	**8000**	**8000**	**32000**

Agregari si granularitate

- **Granularitatea** – reprezinta nivelul de detaliere la care sunt pastrate datele in depozit
- In functie de cerintele de analiza, datele se pot pastra la nivel mai detaliat sau mai agregat (depinde de niv. de detaliere a dimensiunilor)
- **Agregarea** datelor- cresterea performantelor DW
- 10 magazine, 100 produse/marca, vanzari saptamanale

Dacă o interogare necesită...	Atunci trebuie parcurse
1 Produs, 1 Magazin, 1 Săptămână	doar 1 înregistrare din schemă
1 Produs, Toate magazinele, 1 Săptămână	10 înregistrări din schemă
1 Marcă, 1 Magazin, 1 Săptămână	100 înregistrări din schemă
1 Marcă, Toate magazinele, 1 An	52.000 înregistrări din schemă

De la relational la multidimensional

- premise diferite, tehnici diferite şi produc BD cu structuri diferite.
- modul de abordare a datelor (utilizator/date):
  - **model multidimensional** - dimensiuni cât mai apropiate de cele naturale şi de **perspectiva utilizatorului.**
  - **model relational** – perspectiva datelor
- model multidimensional:
  - o BD mult **mai uşor de consultat şi de interogat** la un nivel înalt, sintetic, agregat
  - o BD cu **mai putine tabele şi chei** de administrat decât modelul relational

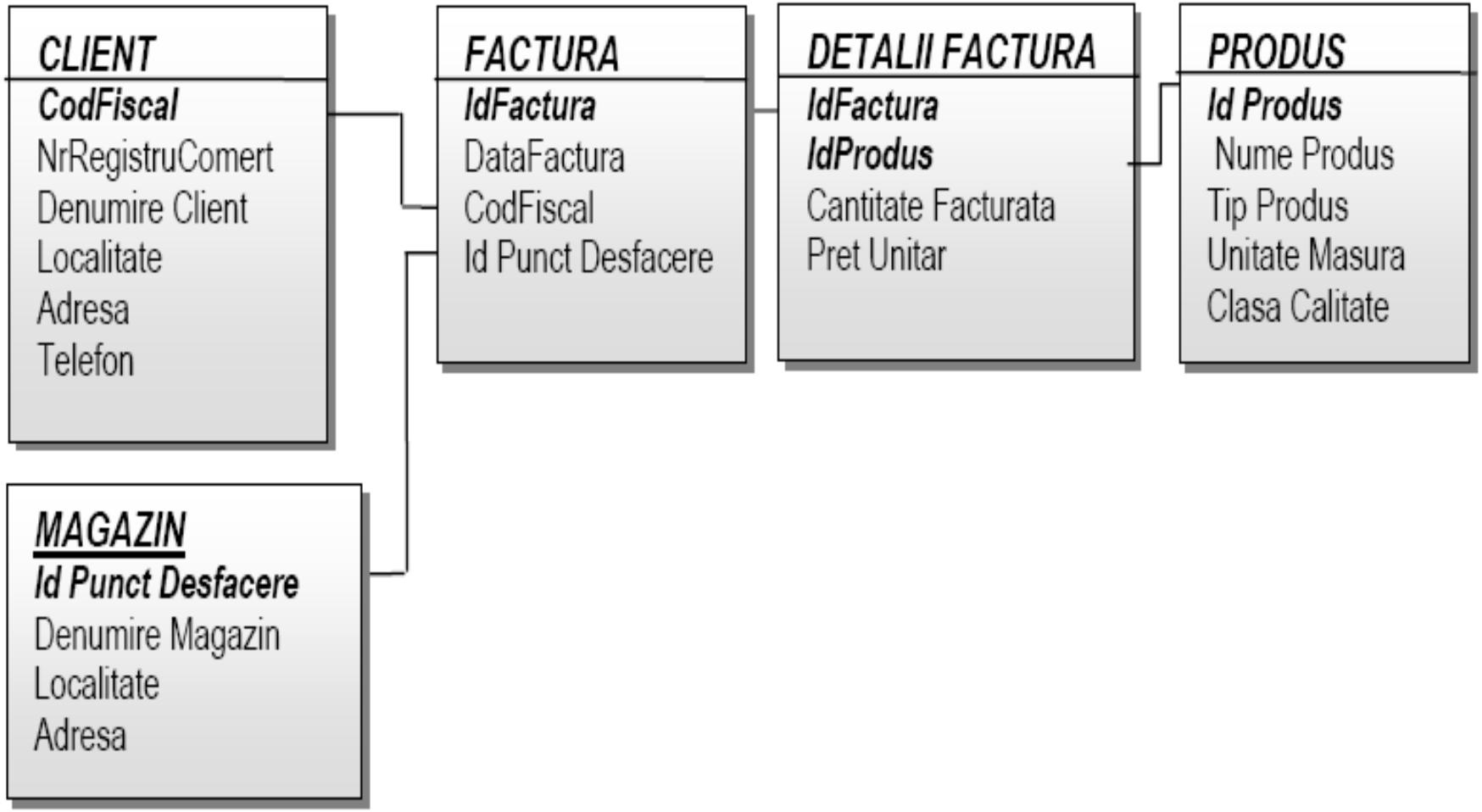
Normalizare in BD, redundante in DW

- procesul de transformare succesivă a unei BDR în vederea aducerii sale într-o formă standard optimizată
- eliminarea anomaliilor, redundanţelor, dependenţelor nedorite între date
- **Anomalii de actualizare**
  - limitarea posibilităţilor de inserare a datelor
  - pierderi de date la ştergere
  - apariţia de inconsistenţe la modificarea datelor
- **Dependente**
  - ***Dependenţă funcţională*** – A depinde funcţional de un B dintr-o tabelă dacă fiecărei valori a lui A îi corespunde numai o valoare a lui B. B ***depinde funcţional complet*** de un grup de atribute dacă B este dependent funcţional de fiecare atribut din grup.
  - ***Dependentă tranzitivă*** –daca B depinde de A şi C depinde de B atunci C se află în dependenţă tranzitivă faţă de A.
  - ***Dependenţă multivaloare*** – dacă valorii unui atribut A îi corespund două sau mai multe valori ale atributului B

Formele normale

- **Forma normală 1 (FN1)** *dacă* atributele sunt la nivel **atomic** şi au fost eliminate **grupurile de atribute** repetitive
- **Forma normală 2 (FN2)** *dacă* este în FN1 şi nu există **dependenţe funcţionale parţiale** pentru atributele non-cheie
- **Forma normală 3 (FN3)** *dacă* este în FN2 şi nu există **dependenţe funcţionale tranzitive** pentru atributele non-cheie
- **Forma normală 4 (FN4)** *dacă* este în FN3 şi există cel mult o dependenţă funcţională multivaloare pentru atributele non-cheie
- **Forma normală 5 (FN5)** *dacă* este în FN4 şi nu există dependenţe joncţiune pentru atributele non-cheie

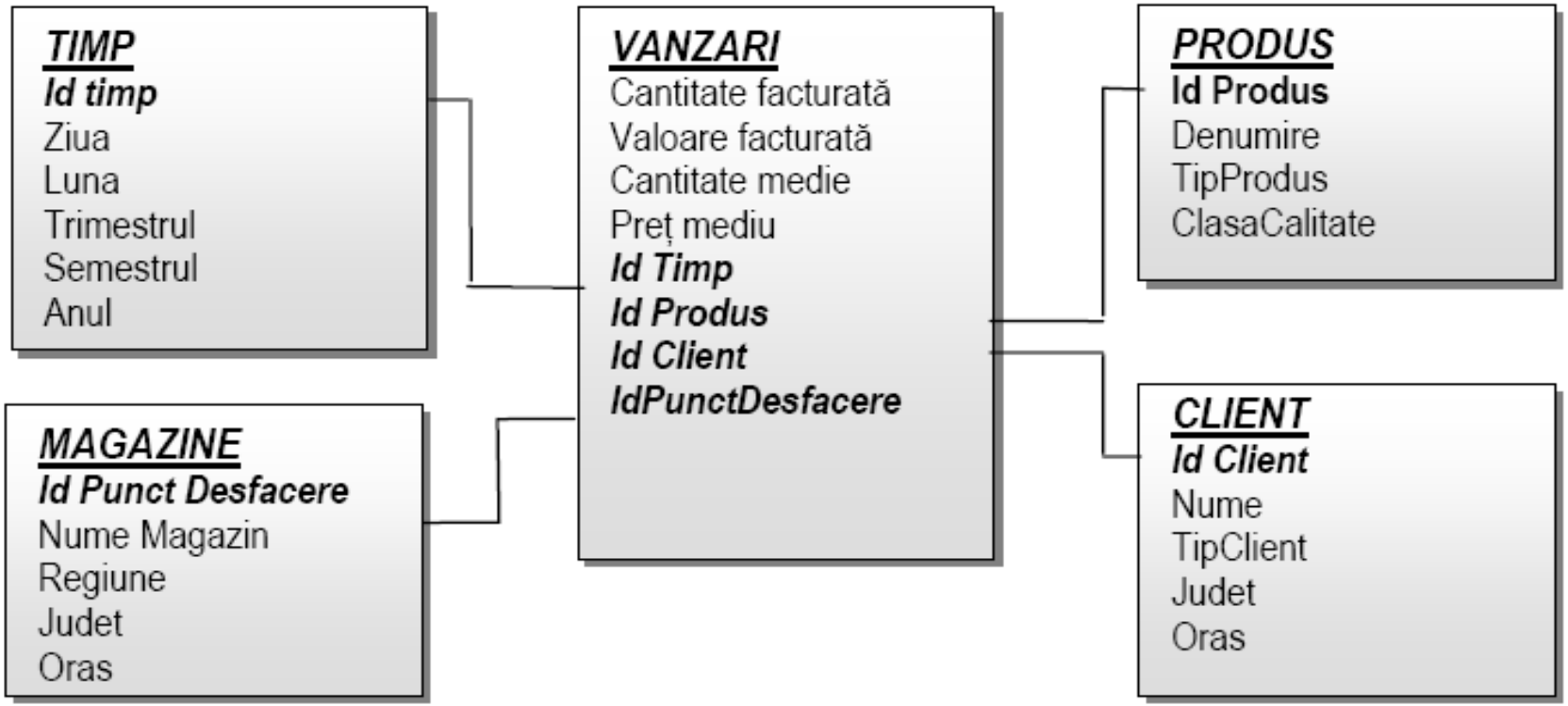
Evidenta facturi – model relational



a.Structura DW – Schema STEA

- cel mai des utilizat model de organizare al depozitelor de date
- **tabela de fapte** cuprinde, fără redundanţe, marea parte a datelor
- tabela de fapte este conectata la tabelele dimensiune pe baza cheilor externe pe care acestea le conţin.
- **star join** = legatura stabilita între un tabel de fapte si tabelele dimensiune
- **star query** = jonctiunea dintre un tabel de fapte si mai multe tabele dimensiune
- **Avantaj**: performante optime pentru interogarile dintr-un depozit de date

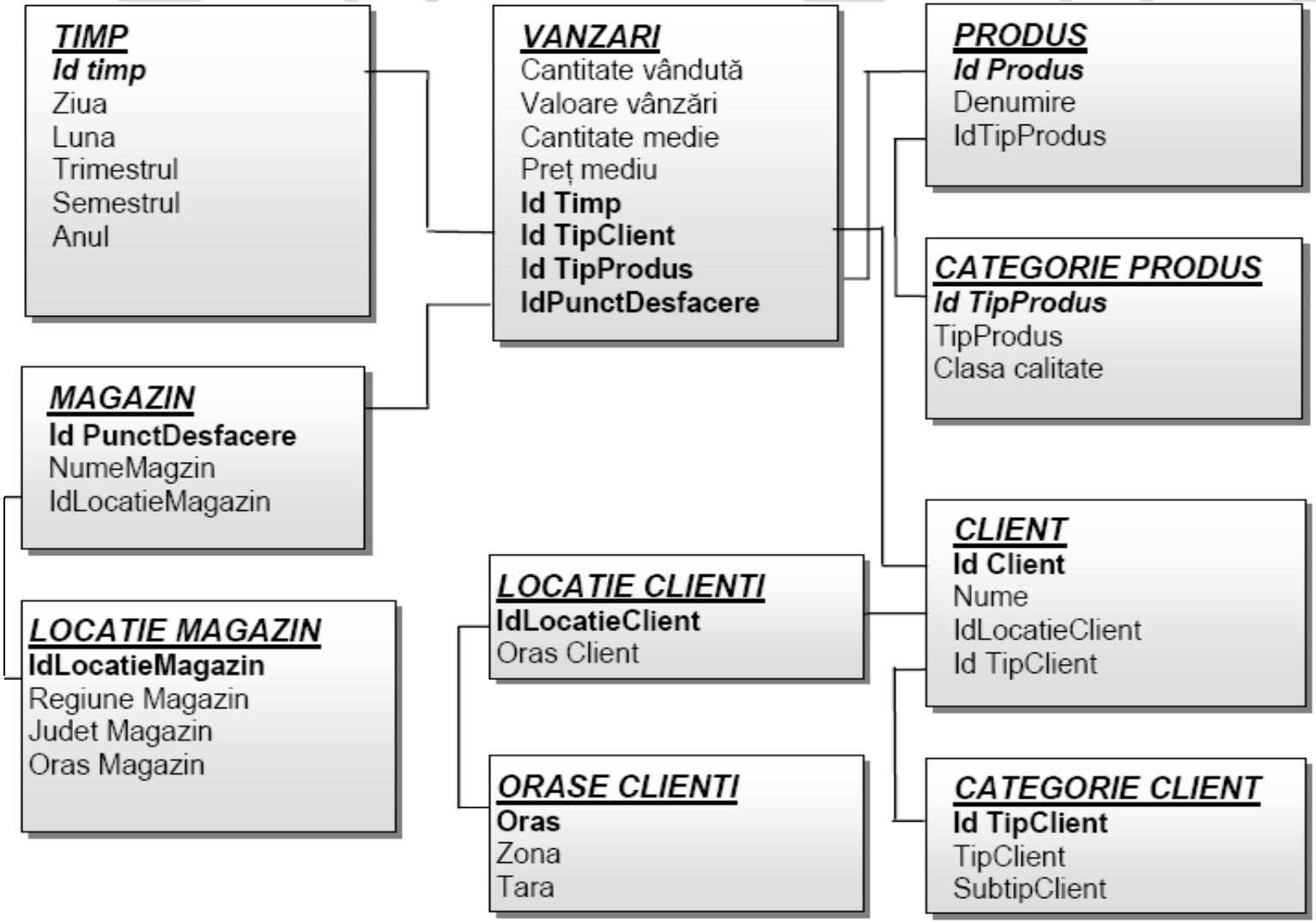
Ex: Schema STEA



b.Structura DW– Schema fulg de zapada

- “**seminormalizat**”, avantajele modelului relaţional.
- tabelele dimensiune respecta regulile de normalizare din modelul relaţional =>economie de spaţiu
- nu va conduce la reducerea spaţiului pt tabela de fapte
- **Avantaje**:
  - Redundanta redusa
  - Usor de întretinut
- **Dezavantaje**: la cereri de interogare complexe(join)=> creste timpul de raspuns

Ex: Schema fulg de zapada

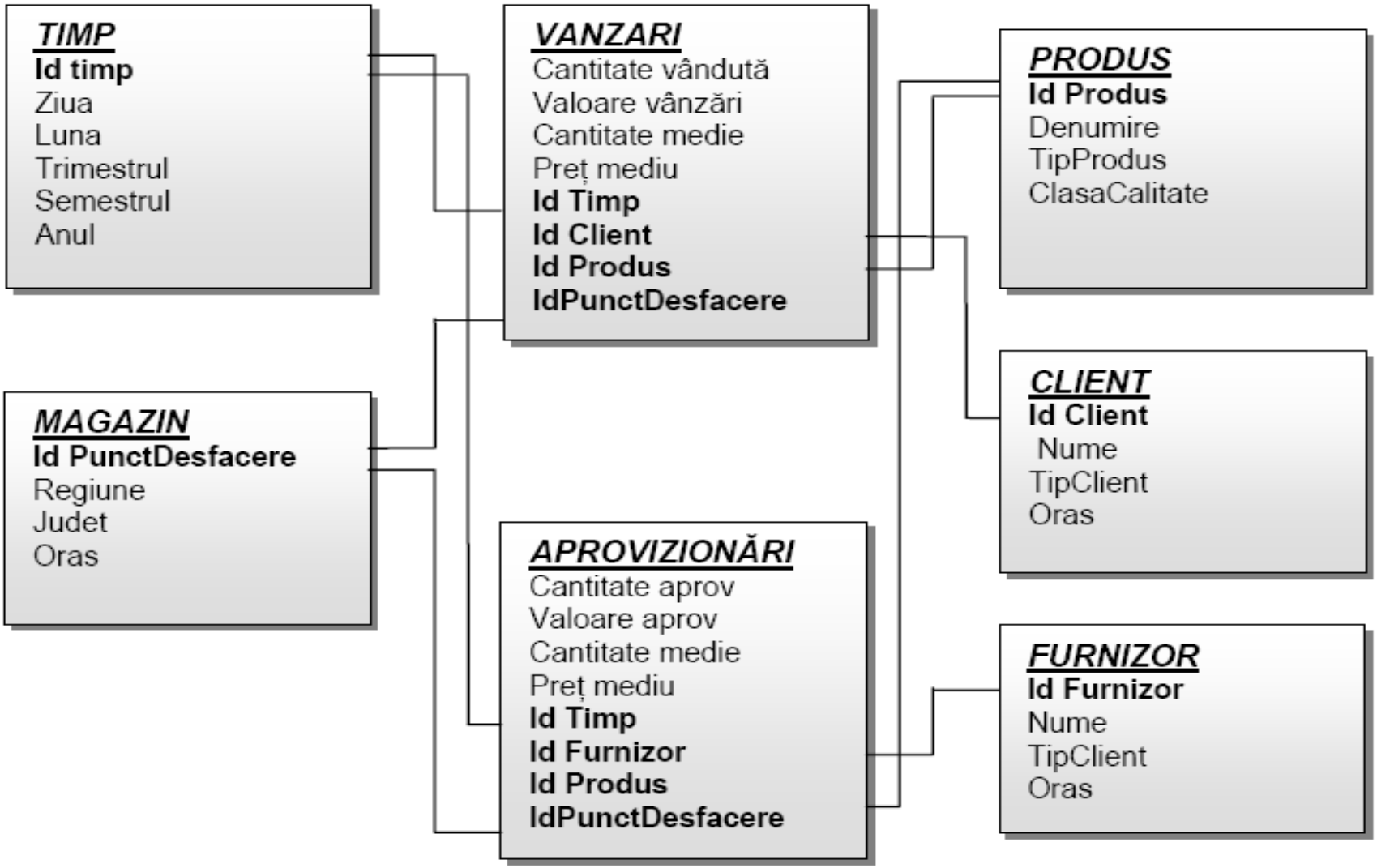


c. Structura DW – Schema constelaţie de fapte

- Schema galaxie
- mai multe tabele de fapte, conectate ce utilizează aceleaşi tabele-dimensiune

- pe lângă tabela de fapte **Vânzări**, o tabelă suplimentară de fapte **Aprovizionări**, legata de dimensiuni

Ex: Schema constelaţie de fapte



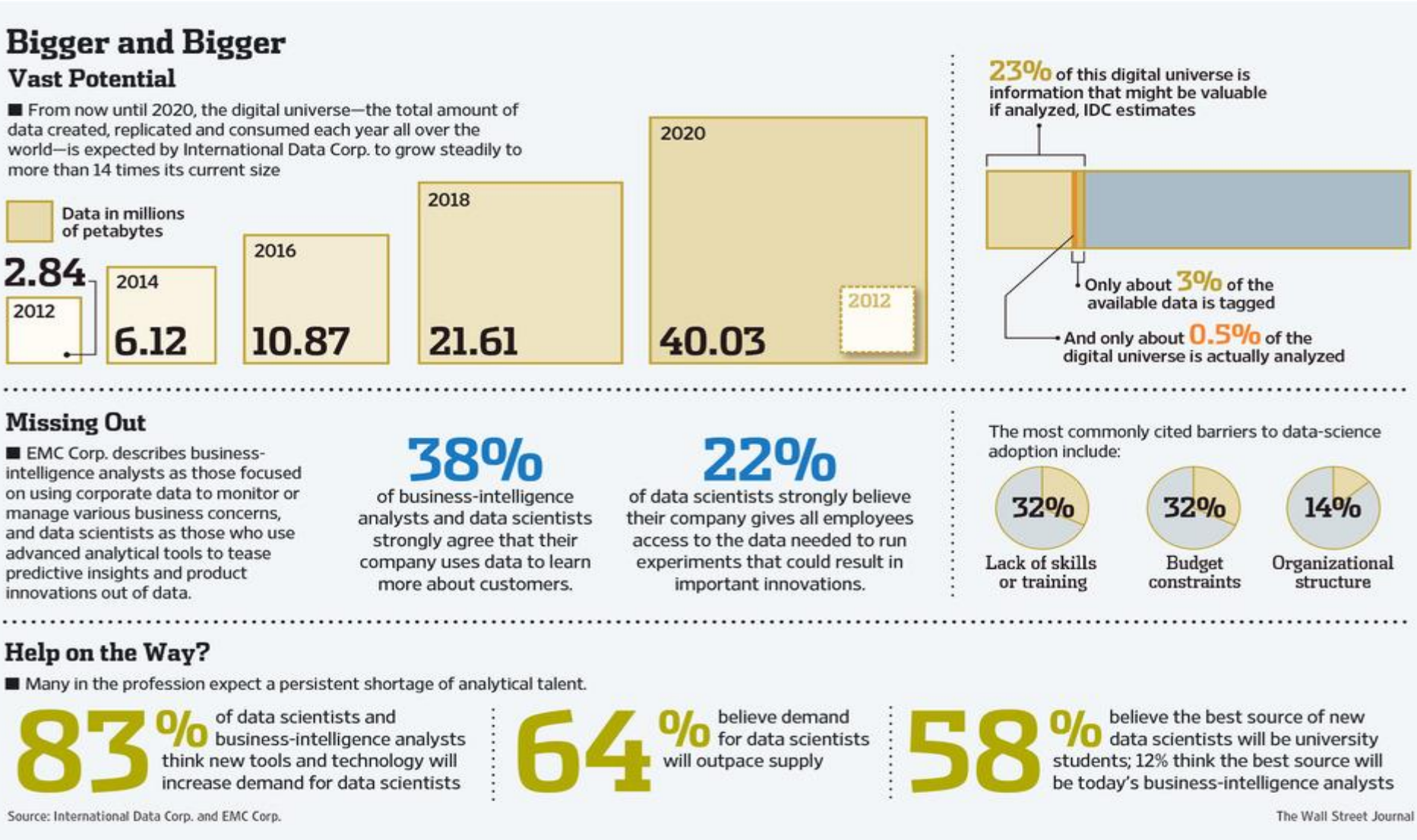
Paralela între prelucrarea relațională şi cea analitică

Caracteristici	Modelul relațional	Modelul multidimensional
Organizarea datelor	Tabela	Dimensiuni, tabele de fapte, cub de date
Nivelul datelor	Detaliu	Agregat
Operația tipică	Actualizare	Raportare şi analiză
Nivelul de analiză cerut	Scazut	Ridicat
Volum de date per tranzacție	Redus	Mare
Vârsta datelor	Curente	Istorice, curente, previzionate

Piata job-urilor

- Se previzioneaza o lipsa mare de personal in urmatorii 5-10 ani pe piata analizei de date
- O piata foarte dinmica
- Acumulari tot mai mari de date
- Noi tehnologii si instrumente
- E nevoie atat de instrumentele traditionale de analytics dar si de expertiza tehnica pentru date nestructurate
  - *IBM are aproape 10000 de consultanti analisti de business si 400 matematicieni*

Nevoia de analiza a datelor



Tehnologii de integrare

1. Baze de date distribuite
2. Depozite de date
3. **Migrarea datelor**

3. Migrarea datelor

- Migrare sau reproiectare la schimbarea BD
- Avantaje reproiectare
  - posibilitatea de a începe de la zero şi a elimina slăbiciunile structurale;
  - adoptarea de noi tehnologii;
  - crearea unei fundaţii proaspete pentru noul sistem
- Dezavantaje reproiectare
  - analiza, proiectarea şi implementarea unui nou sistem solicită mult timp şi resurse
  - este posibil ca noul sistem să fie mai puţin funcţional decât vechiul

Factori ce influenteaza migrarea

- Diferenţele de **sintaxă SQL** între principalele SGBD-uri;
- Integrarea de **restricţii de integritate** şi **algoritmi** atat in BD sursa, cat si in destinatie
- **Asistent de migrare**, care să automatizeze cele mai multe sarcini, iar administratorul BD să faca doar corecţii minore şi de fineţe.
- Interdependenţa dintre obiectele BD
- Volumul mare de date – durata mare transfer

Etapele migrarii datelor

A. Export si conversie
B. Transfer si procesare
C. Import

A. Export si conversie

- Se exporta si se convertesc toate/ o parte din obiectele BD
  - Tabele
  - Viziuni
  - Proceduri/ functii/ pachete stocate
  - Declansatori
- Redenumiri sau schimbari de tipuri

=> **Fisiere ASCII cu comenzi SQL** pentru crearea structurii si cu date pentru popularea BD

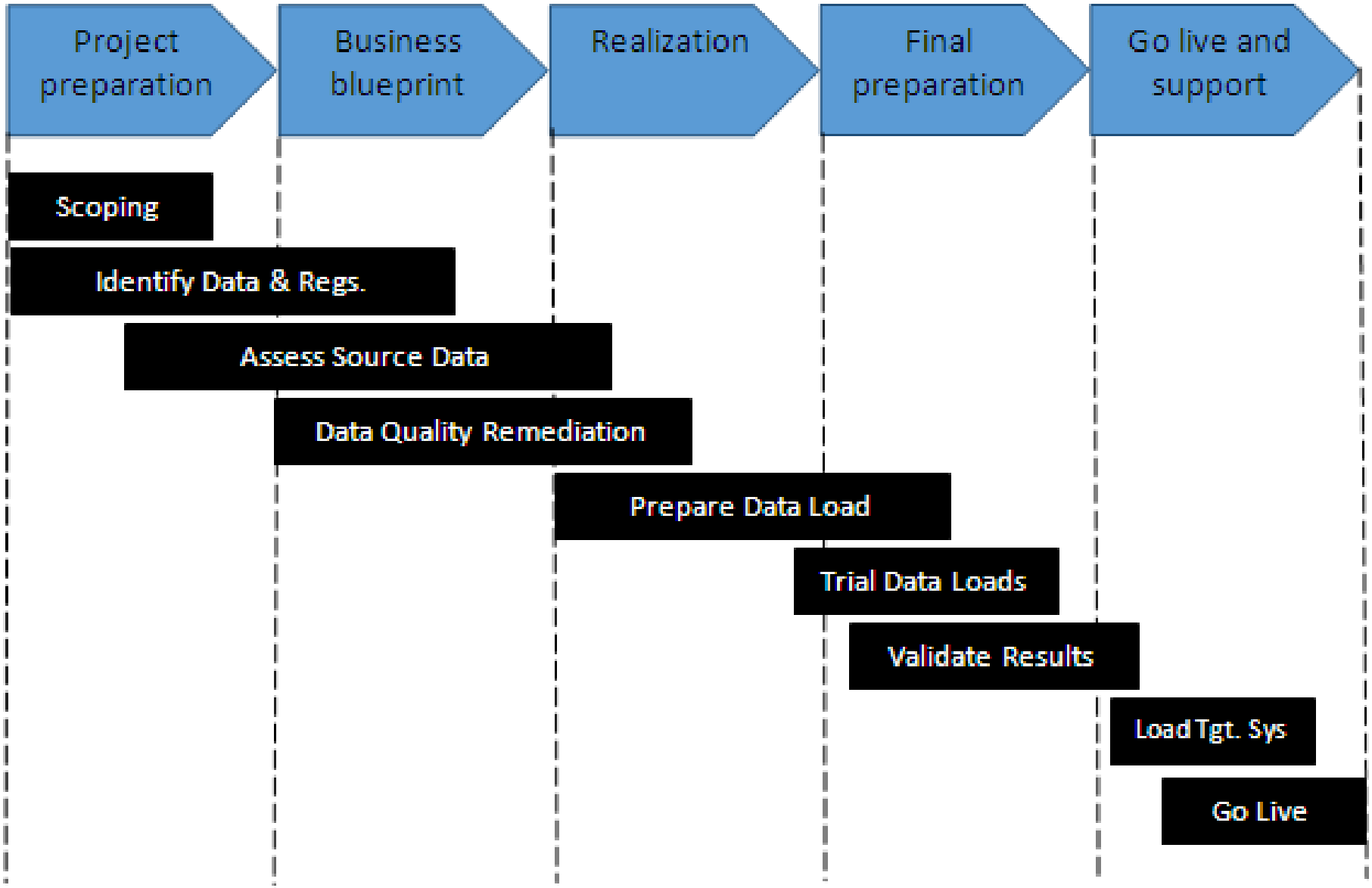
B. Transfer si procesare scripturi

- Optionala, daca e nevoie de transfer
- Procesarea scripturilor transferate – modificari pt nevoi **neacoperite** de agentul de migrare folosit

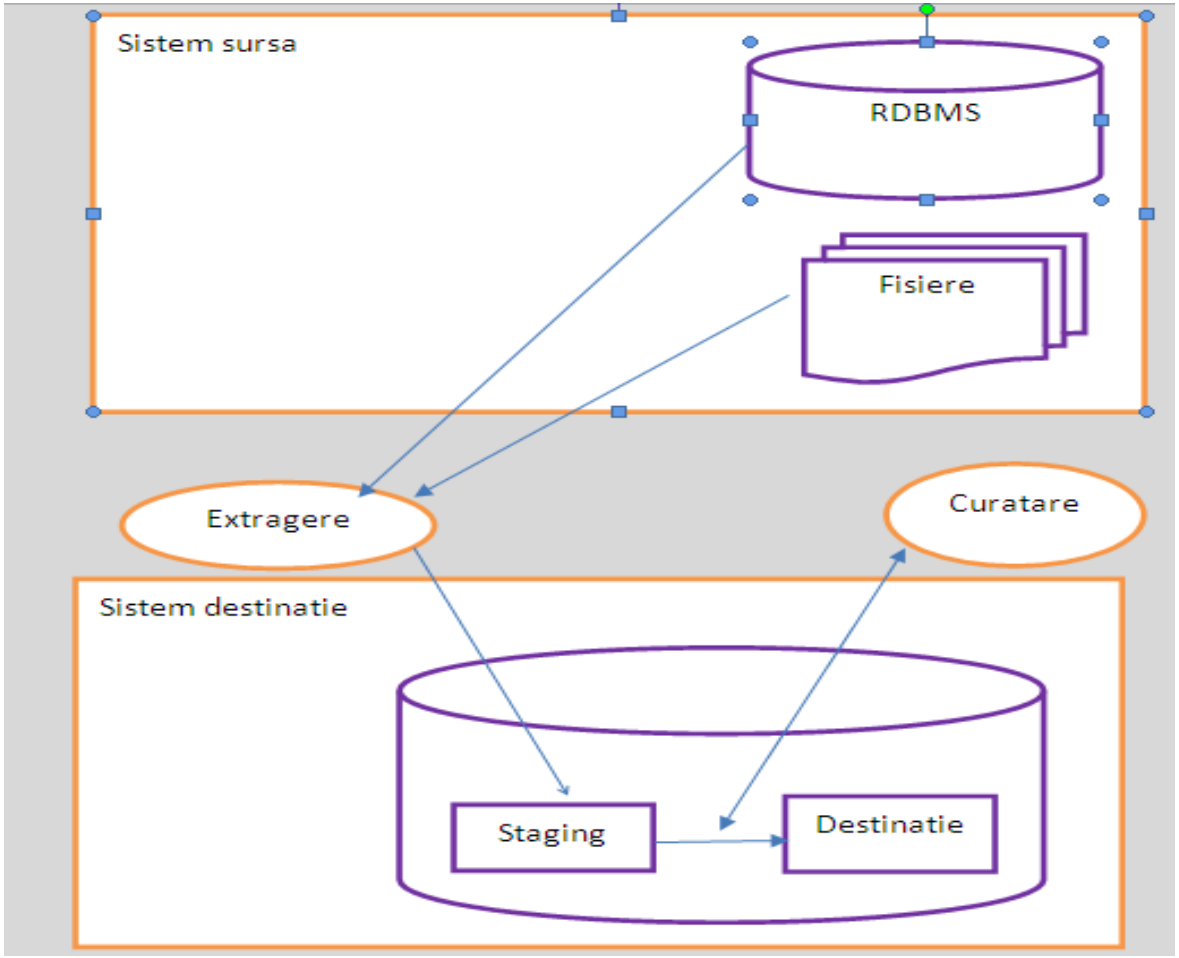
C. Import

- Scriptul creat la A si prelucrat la B e executat pe BD destinatie
- Utilitare pt executia scripturilor:
  - **SQL Plus** pentru Oracle;
  - **CLP** (Command Line Processor) pentru IBM DB2;
  - **ISQL** pentru Ms SQL Server şi SyBase;
  - **linia de comandă** MySQL.
- Utilitare pt. incarcare date din fisiere ASCII:
  - **SQL Loader** pentru Oracle;
  - **LOAD/IMPORT** pentru IBM DB2;
  - **BCP** pentru SQL Server şi Sybase;
  - **LOAD DATA INFILE** pentru MySQL;
  - **BUTIL** pentru Persasive SQL.

Planificarea unui proiect de migrare a datelor

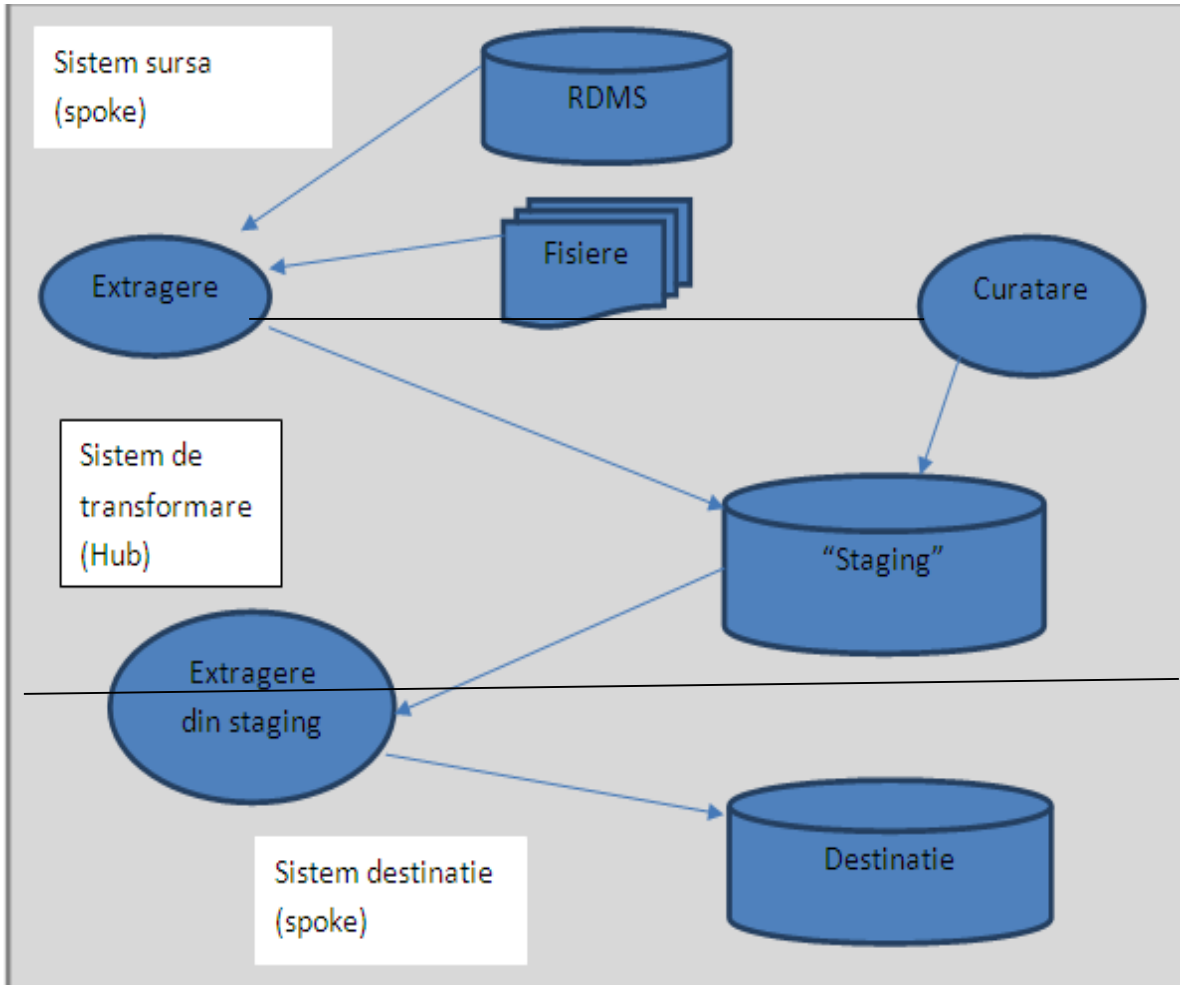


Arhitectura de migrare punct-la-punct



- Transformarile locale inseamna ca procesul de migrare a datelor s-a terminat, datele au ajuns la sistemul destinatie
- Reduce timpul petrecut “in retea”

Arhitectura Hub-Spoke de migrare a datelor



- Poate integra orice numar de sisteme sursa/destinatie
- Regulile datelor sunt pastrate la un nivel separat

Strategii de migrare

- **Strategia “Big Bang”**
  - migrarea tuturor datelor si trecerea la noul sistem in acelasi pas.
  - avantajul -lipsa nevoii interoperabilitatii intre vechiul sistem si noul sistem.
  - dezavantaj- durata mare de “downtime” sau neputinta testarii in productie
- **Strategia “Chicken Little”**
  - sistemul sursa este divizat in unitati cu cat mai putine interdependente
  - vechiul sistem si noul sistem ruleaza in paralel in timp ce modulele sunt transferate
  - migreaza datele incremental, asigurand integritatea informatiei.
  - Avantaj: timpul de stabilizare permis intre migrarile modulelor., testare
- **Strategia “Butterfly”**
  - sursa este migrata iterativ pana cand diferenta dintre cele doua sisteme a atins pragul prestabilit, moment in care, restul informatiei este transferat si noul sistem este pornit
  - nu se foloseste de portalul intre cele doua sisteme.
  - sistemul tinta nu se afla in productie deci nu trebuie sincronizat dupa fiecare migrare.