



*Universitatea Tehnică a Moldovei
Departamentul Informatică
și Sisteme Inginerești*



Depozite de Date

-DW-

Note de curs sintetice

Chișinău, 2024

© *Dr. Ion GANEA,
Lector univ.*



Repere

1. DEPOZITE DE DATE
2. DATA MINING

**STRICT PENTRU
UZ PERSONAL !!!**



Depozite de Date

- *Colecție de date orientate pe subiecte, integrate, istorice și nevolatile, 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 analiză;*
- *Structurile de date într-un depozit de date sunt optimizate pentru o regasire și o analiza rapida.*

Caracteristicile principale ale DD

- *Dimensiunea foarte mare, volumul de date fiind de la sute de GB la TB;*
- *Este organizat la nivelul unei organizații;*
- *Cererile de regăsire sunt ad-hoc și implică foarte multe înregistrări;*
- *Sursele de date sunt variate: istorice statistice, demografice etc.;*
- *Se pot folosi modele de date multidimensionale și tehnologia **OLAP**;*
- *Necesitatea DD se referă la îmbunătățirea calității informațiilor din organizații.*

OLAP (OnLine Analytical Processing)

- *Tehnologie care permite analiza multidimensională a datelor, esențială pentru aplicațiile de Business Intelligence (BI).*
- *OLAP permite utilizatorilor să extragă și să vizualizeze date din mai multe perspective, facilitând astfel raportarea complexă și analiza trendurilor.*
- *Datele sunt organizate în structuri de tip cub, permițând interogări rapide și agregări eficiente pe dimensiuni precum timp, locație sau produse.*
- *Această tehnologie este ideală pentru analiza datelor istorice, identificarea patternurilor și suportul decizional strategic.*

OLAP vs OLTP

- *OLAP și OLTP sunt două tehnologii fundamentale, dar destinate unor scopuri diferite în gestionarea și analiza datelor.*
- *OLAP este orientat spre analiza datelor și suportul decizional, lucrând cu seturi mari de date agregate,*
- *OLTP se concentrează pe gestionarea tranzacțiilor curente și eficiente într-un mediu operațional.*
- *Ambele sunt esențiale în ecosistemul informațional al unei organizații, dar servesc scopuri foarte diferite.*

Caracteristici OLAP

- **Scop**: Analiză complexă a datelor, suport pentru decizii strategice.
- **Tip de Date**: De obicei, date agregate și istorice, stocate în depozite de date (Data Warehouses).
- **Interogări**: Se axează pe interogări complexe, cum ar fi raportarea, analiza trendurilor și identificarea patternurilor.
- **Performanță**: Optimizat pentru citirea rapidă și analiza datelor pe mai multe dimensiuni.
- **Structură**: Folosește structuri de tip cub multidimensional, facilitând explorarea datelor din diferite perspective.
- **Utilizatori**: Manageri, analiști de afaceri, și alte persoane implicate în luarea deciziilor.

Caracteristici OLTP (OnLine Transaction Processing)

- **Scop**: Gestionarea și procesarea eficientă a tranzacțiilor de zi cu zi, operațiuni de bază.
- **Tip de Date**: Date curente și detaliate, stocate în baze de date operaționale.
- **Interogări**: Interogări simple și frecvente, cum ar fi inserarea, actualizarea și ștergerea datelor.
- **Performanță**: Optimizat pentru tranzacții rapide și eficiente, cu timp de răspuns minim.
- **Structură**: Baze de date normalizate pentru a evita redundanța și pentru a asigura integritatea datelor.
- **Utilizatori**: Personal operațional, aplicații de front-end care necesită acces rapid la date.

Scheme de depozitare a datelor

- *La fel ca un sistem **OLTP** (**bază de date**), un sistem **OLAP** (**depozit de date**) este, de asemenea, necesar pentru a menține o schemă.*
- *O bază de date utilizează un*

MODEL ENTITATE - RELAȚIE

- *Un depozit de date utilizează un*

MODEL DIMENSIONAL

Colecția de date de tip DD

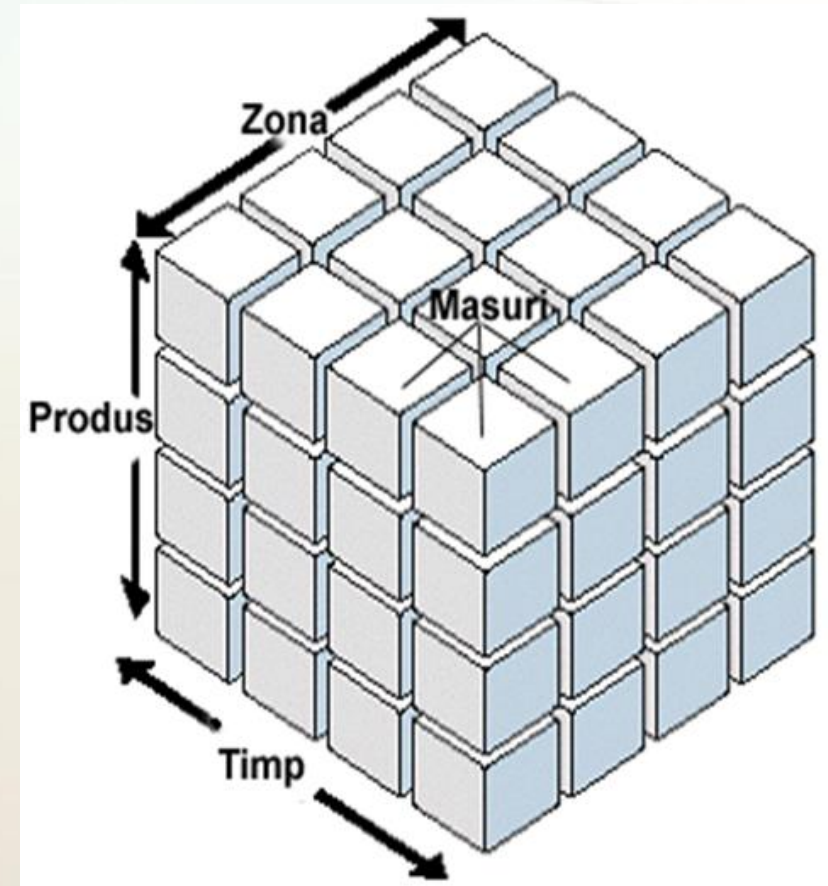
- ***Este orientată pe subiectele importante din organizație: clienți, produse etc.***
- ***Integrată deoarece datele provin din surse diferite printr-un proces anterior de curățare / filtrare și prelucrare;***
- ***Non-volatilă pentru că operațiile efectuate sunt în special de regăsire și periodic de încărcare, nu de actualizare;***
- ***Dependentă de timp pentru că datele sunt stocate pentru o perioadă de 5-10 ani;***

- Permite vizualizarea datelor prin mai multe filtre sau dimensiuni in acelasi timp.
- **Dimensiuni=coordonate=categorii de informație.**

De ex:

Care sunt **vanzarile reale** in comparatie cu cele previzionate pe **zona**, pe **vanzator**, pe **produs**?

Care este **profitabilitatea** pe **produs**, pe **client**?



Model Relațional vs Model Multidimensional

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

Tipuri DW

1 DEPOZITE DE ÎNTREPRINDERE (ENTERPRISE WAREHOUSE)

- întreaga structură organizațională un volum extins de date: atât informații detaliate, cât și agregate.
- suporturi hardware performante.
- costurile și timpul de proiectare și implementare sunt considerabile,

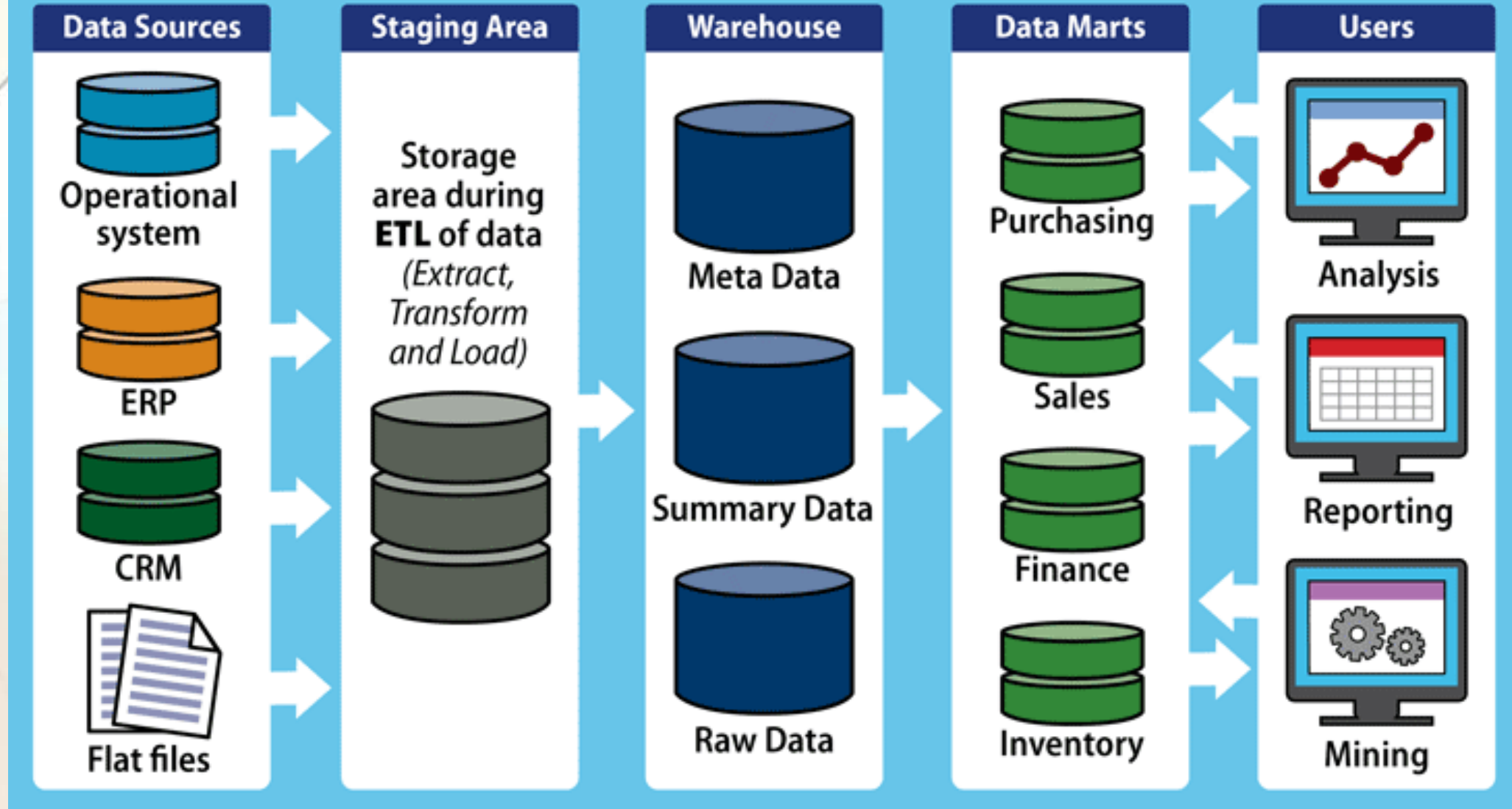
2 DATA MART

- specific unui anumit subset de cerințe sau unui departament din cadrul organizației de regulă, datele conținute într-un data mart sunt agregate.
- costurile și termenele de implementare sunt considerabil reduse

3 DEPOZIT VIRTUAL (VIRTUAL WAREHOUSE)

- o serie de vederi (views) realizate direct asupra BD operaționale.
- procesele de agregare pot afecta capacitățile de prelucrare ale serverelor utilizate în activitatea operațională,
- aparent ușor de implementat, necesită capacități de procesare deosebite.
- necesită curățare și consolidare în timpul rularii.

Data Warehouse Architecture



SURSA: <https://todobi.com/comparativa-de-las-15-mejores-soluciones-de-data-warehouse/>

Structuri de date într-un Depozit de Date

- Într-un DD, structurile de date sunt esențiale pentru organizarea și gestionarea eficientă a datelor.
- Structurile sunt proiectate pentru a optimiza stocarea, interogarea și analiza datelor:
 - **Tabele FAPTE** (Fact Tables):
 - **Tabele DIMENSIUNE** (Dimension Tables)
 - **Cuburi OLAP** (OLAP Cubes)
 - **Scheme** (Schemas)
 - **Indici** (Indexes)
 - **Vederi Materializate** (Materialized Views)
 - **Partiționarea Datelor** (Data Partitioning)
 - **Diagrama Relațională** (Entity-Relationship Diagram-ERD)

Tipuri de tabele în DD

1. TABELE DE FAPTE (MĂSURI)

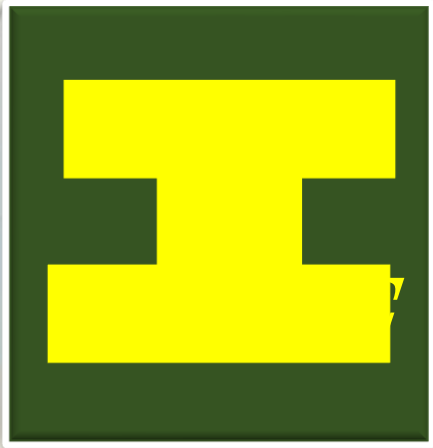
- **Scop:** Conțin faptele și cheile externe către tabelele de dimensiuni.
- Sunt măsurători numerice (numere întregi, valori reale) care reprezintă evenimente sau activități.
- Sunt totalizate și analizate pe diferite niveluri (agregare).
- **Caracteristici:**
 - Conțin o cheie primară compusă din cheile străine ale dimensiunilor.
 - Sunt foarte mari și conțin un volum mare de date.
 - Sunt optimizate pentru operații de agregare și sumarizare.
- **Exemple:** Vânzări, costuri, profit, număr de clienți, cantități.

2. TABELE DIMENSIUNE

- **Scop**: Oferă contextul pentru măsurători, descriind dimensiunile diferitelor aspecte ale datelor:
 - a) Categoriile de informații care organizează datele;
 - b) Fiecare tabelă dimensiune are câte o cheie principală;
 - c) Câmpurile sunt de obicei textuale - sursă pentru restricții și pentru rândurile din rapoarte;
 - d) Datele sunt de obicei colectate la nivelul cel mai de jos și mai detaliat și agregate pe nivelele superioare pentru analiză.
- **Caracteristici**:
 - Conțin atribute descriptive (text, date) despre dimensiuni.
 - Sunt mai mici decât tabelele de fapte.
 - Sunt utilizate pentru a filtra și segmenta datele.
- **Exemple**: Timp (ani, luni, zile), produse, clienți, locații geografice.

Ierarhiile

- **Definiție:** Structuri logice de attribute utilizate pentru ordonarea nivelelor de reprezentare a datelor.
- Permite utilizatorilor să navigheze prin date la niveluri diferite de granularitate.
- Definesc căile de navigare în interiorul datelor, permițând detalierea graduală a datelor (relațiile de tip „părinte-copil” între attribute).
- **Exemplu:** Într-o tabelă dimensiune Time (Timp), o ierarhie tipică ar putea fi Year > Quarter > Month > Day. Aceasta ierarhie permite utilizatorilor să analizeze datele la nivel de an, trimestru, lună sau zi, oferind perspective diferite asupra datelor.
- **RoI:** Ierarhiile facilitează agregarea datelor și permit utilizatorilor să navigheze între diferite niveluri de detaliu în analizele OLAP.
- De exemplu, într-o analiză de vânzări, utilizatorii pot începe cu o vedere de ansamblu la nivel de an și apoi pot să aprofundeze la nivel de lună sau zi pentru a vedea tendințele detaliate.
- Ierarhiile organizează attributele (un nivel al unei dimensiuni) într-o formă care reflectă relațiile naturale și ierarhice dintre acestea, făcând analizele și rapoartele mai intuitive și mai ușor de interpretat.
- **Utilizare:** Attributele sunt esențiale pentru definirea detaliilor specifice ale datelor, în timp ce ierarhiile permit organizarea și agregarea acestor detalii în moduri care reflectă structurile logice și operaționale ale unei afaceri.



Importanță tipuri structuri

- **Performanță**: Structura specifică a depozitelor de date permite interogări și analize foarte rapide, chiar și pe volume mari de date.
- **Flexibilitate**: Permit efectuarea de analize multidimensionale, oferind o imagine completă a afacerii.
- **Scalabilitate**: Pot fi ușor extinse pentru a include noi date și dimensiuni.
- Înțelegerea diferenței dintre tabelele de fapte și dimensiune este esențială pentru a proiecta și utiliza eficient un depozit de date.
- Aceste structuri de date permit organizațiilor să obțină informații valoroase din datele lor istorice și să ia decizii mai bune.

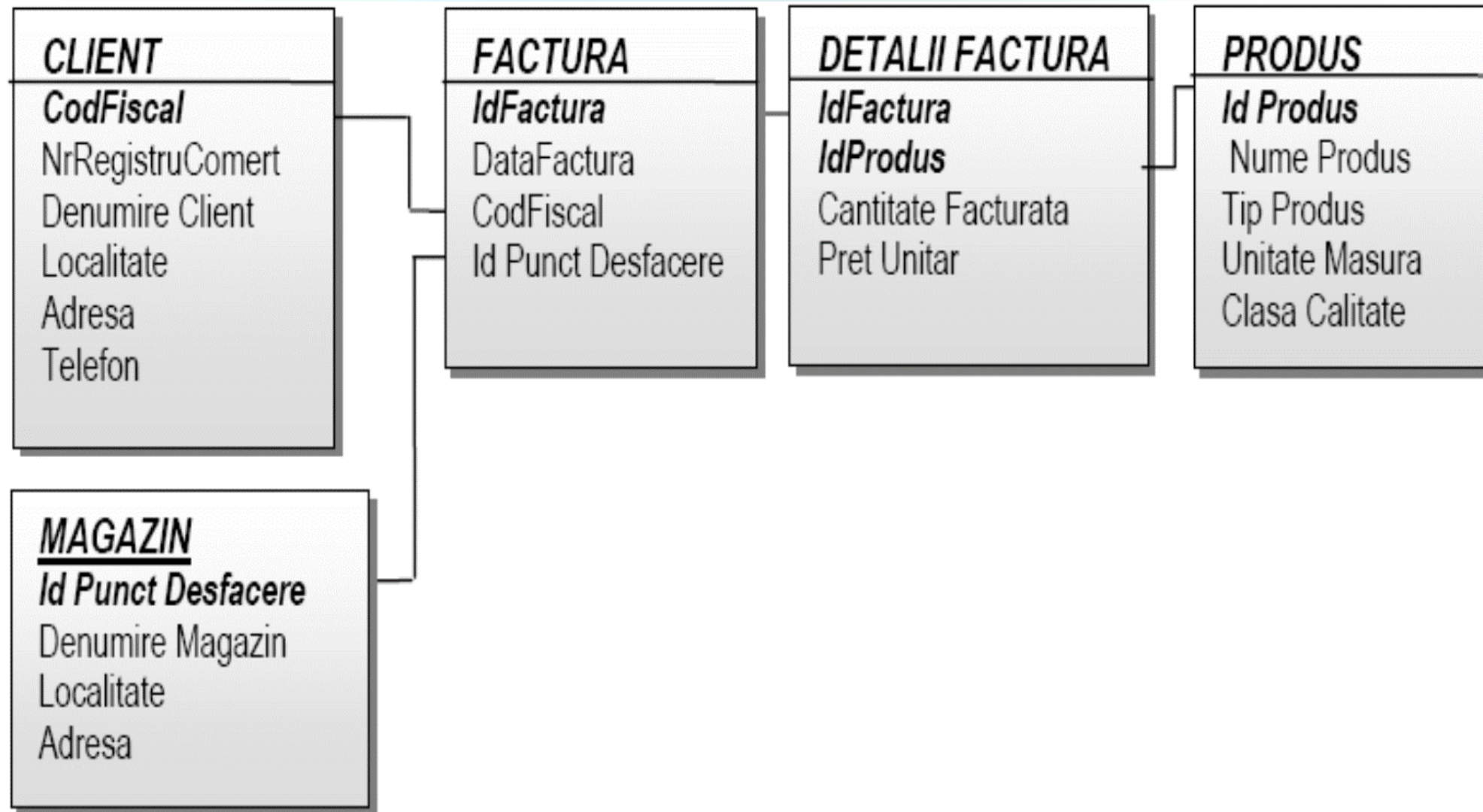
Tipurile de structuri

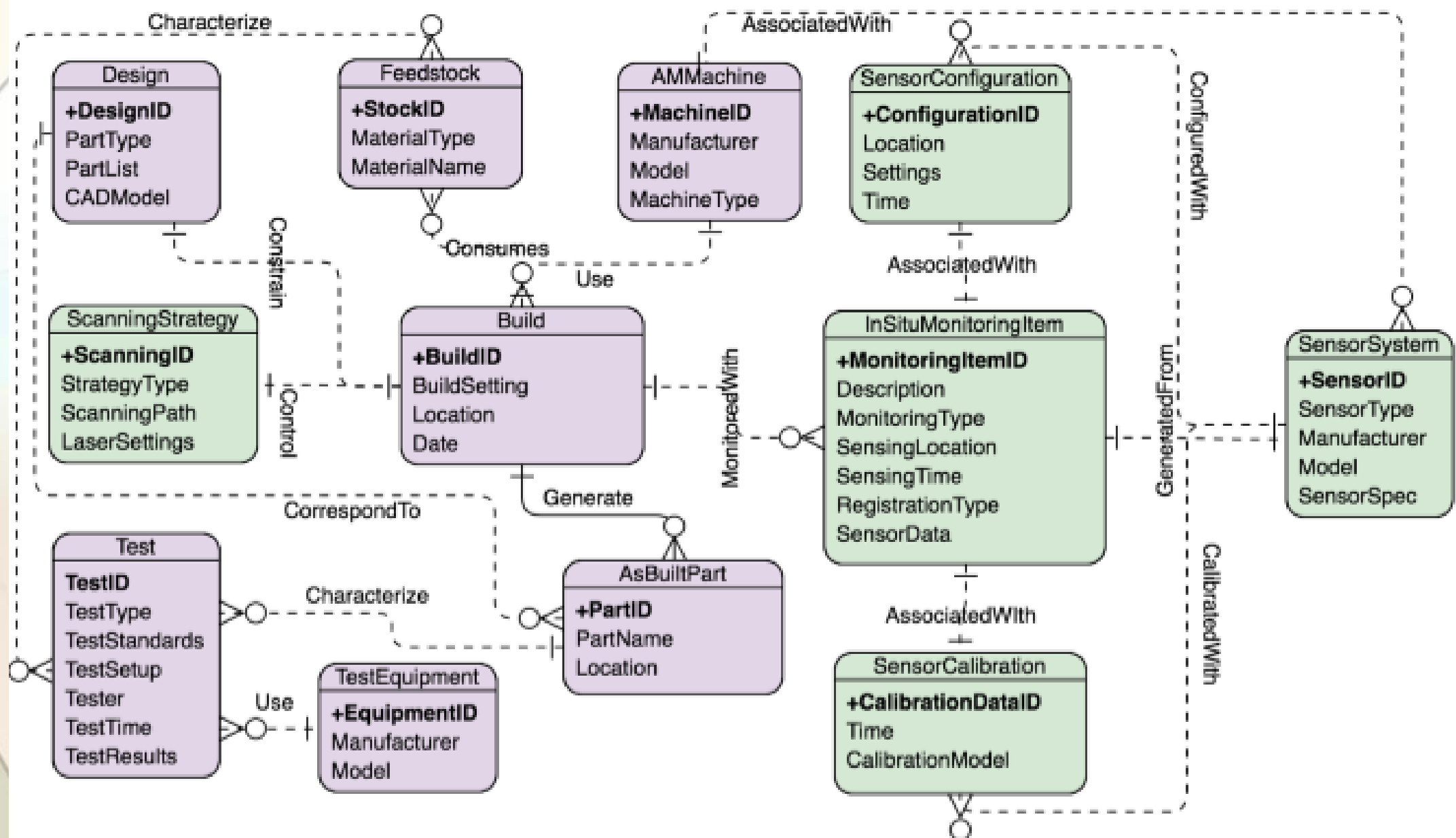
- În funcție de cerințele de analiză, datele se pot păstra la nivel mai detaliat sau mai agregat (depinde de nivelul de detaliere a dimensiunilor).
- Tabelele de fapte sunt legate de tabelele dimensiune prin chei străine.
- Această relație creează o structură multidimensională, care permite analize complexe și flexibile.

Structuri comune:

- **Schema stea**: O tabelă de fapte este în centrul schemei și este conectată la mai multe tabele dimensiune.
- **Schema Fulg de Nea**: unele tabele dimensiune sunt normalizate pentru a reduce redundanța.

Schema Facturi – Model Relațional

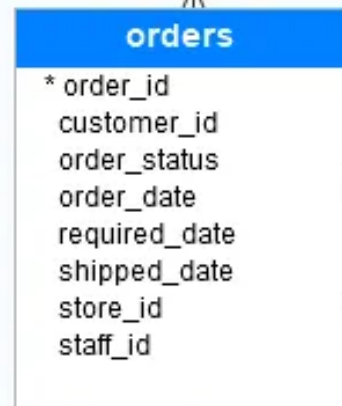
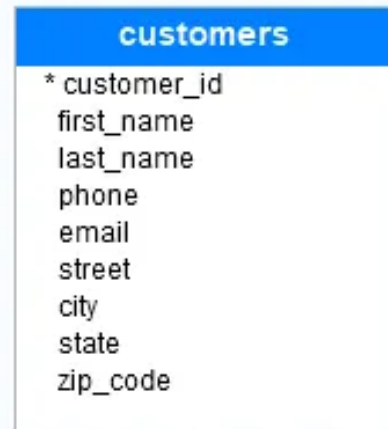




BD BikeStores

Sales

Production



Structura DW „STEA” (Star Scheme)

- Cel mai des utilizat model de organizare al DD.
- Tabela de fapte cuprinde, fără redundanțe, marea parte a datelor.
- Tabela de fapte este conectată la tabelele dimensiune pe baza cheilor externe pe care acestea le conțin.
- **Star Join** = legătura stabilită între un tabel de fapte și tabelele dimensiune.
- **Star Query** = jonctiunea dintre un tabel de fapte și mai multe tabele dimensiune.
- Avantaj: performanțe optime pentru interogările dintr-un depozit de date.

Caracteristici „STEa”

- Structură specifică pentru modelarea datelor multidimensionale.
- Este una dintre cele mai simple și intuitive structuri, fiind eficientă pentru interogări OLAP.
- Componentele principale ale structurii stea:

1. Tabelul fapt (Fact Table):

- Poziția centrală:** În schema stea, tabelul fapt se află în centru și conține datele numerice principale care sunt analizate.
- Coloane:** Acesta include coloane care reprezintă cheile străine către tabelele dimensiune și măsurătorile de interes (de exemplu, vânzări, cantități, profit).
- Măsurători:** Sunt valori numerice care pot fi agregate, cum ar fi Σ vânzărilor, nr. produse vândute etc. 

2. Tabelele dimensiune (Dimension Tables)

- **Poziția periferică**: tabelele sunt situate în jurul tabelului fapt și conțin attribute descriptive care oferă context măsurătorilor din tabelul fapt.
- **Attribute**: Fiecare tabel dimensiune are un set de coloane descriptive (attribute), cum ar fi numele produsului, categoria produsului, numele clientului, data etc.
- **Chei primare**: Fiecare tabel dimensiune are o cheie primară unică, care este legată de cheia străină din tabelul fapt.
- Fiecare **cheie străină** din tabelul fapt corespunde unei chei primare dintr-un tabel dimensiune.

Caracteristici stea (continuare)

- **Simplitate:** Structura stea este simplă și ușor de înțeles, ceea ce face interogările SQL mai directe și ușor de scris.
- **Performanță:** Este optimizată pentru interogările de tip OLAP, în special pentru agregările și sumarizările pe dimensiuni multiple.
- **Redundanță redusă:** Tabelele dimensiune nu sunt normalizate, ceea ce poate introduce un grad de redundanță, dar crește performanța interogărilor.

Exemplu de structură stea

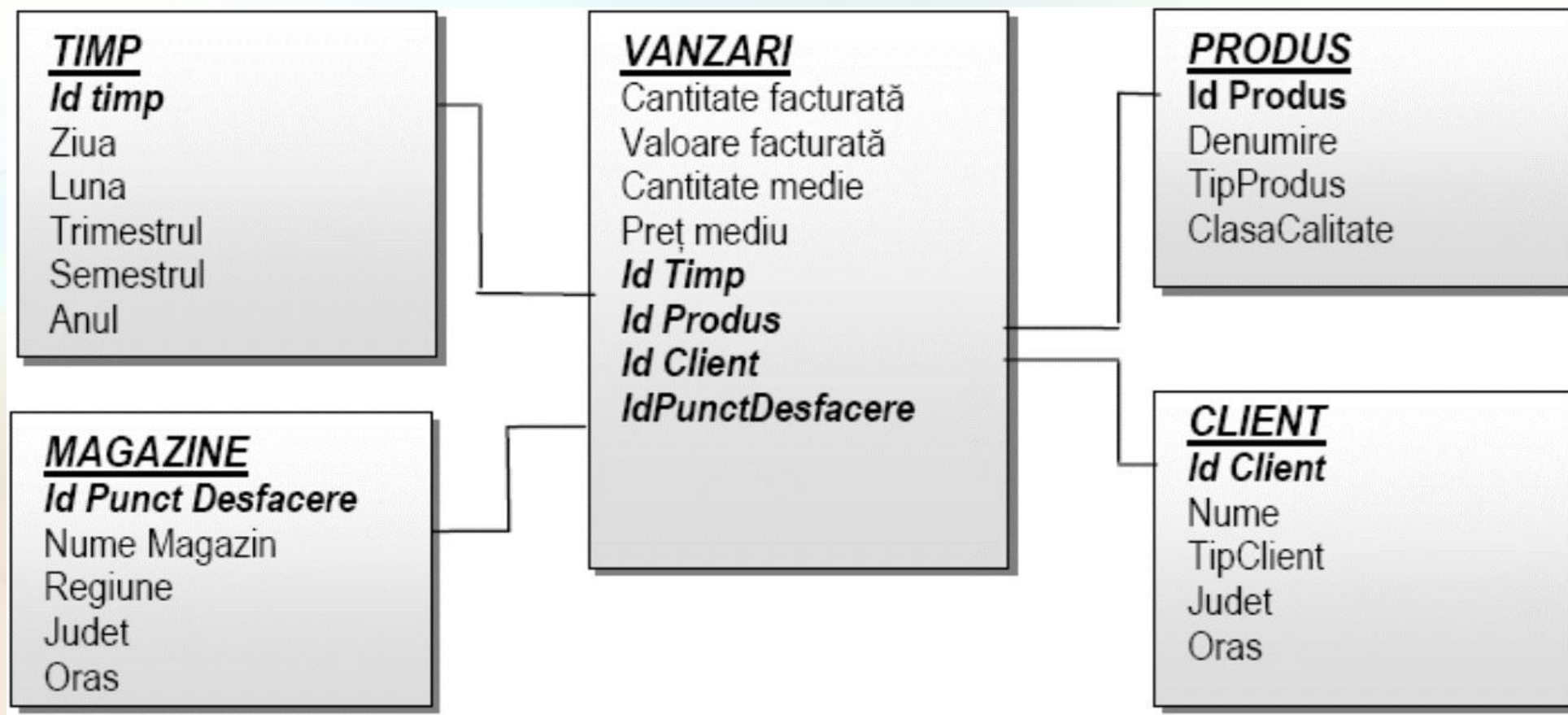
Depozit de date pentru o companie de retail

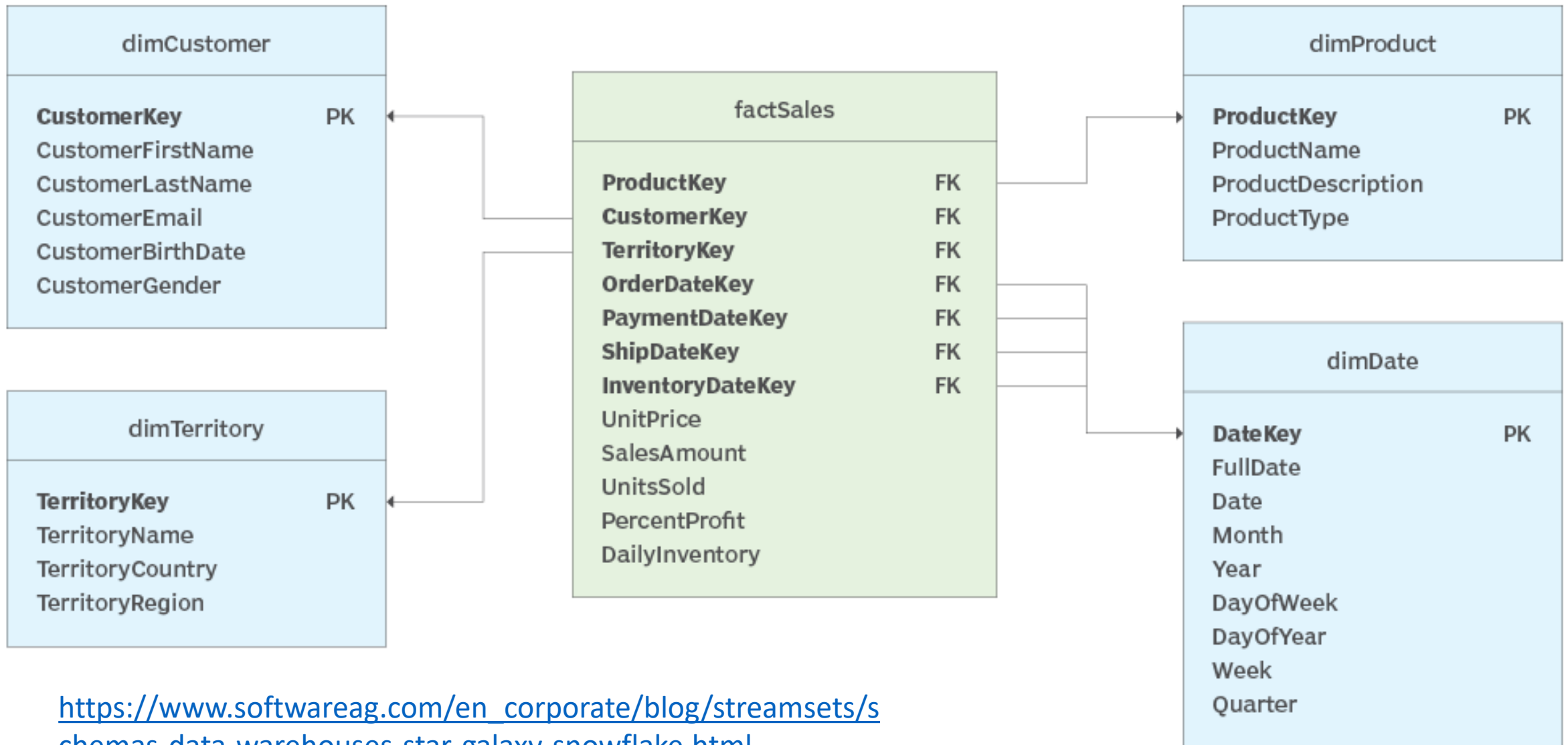
(1). **Tabelul fapt**: **Sales** (ID produs, ID client, ID timp, cantitate_vândută, suma_totală_vânzări).

(2). **Tabele dimensiune**:

- **Product** (ID produs, nume produs, categorie, preț).
- **Customer** (ID client, nume client, locație).
- **Time** (ID timp, zi, lună, an, trimestru).
- Tabelul Sales este Legat prin chei străine de tabelele Product, Customer și Time, formând structura stea.
- Această structură permite analiza vânzărilor pe produse, clienți și perioade de timp, oferind o flexibilitate mare pentru interogări de tip OLAP.

Exemplu „Stea”





https://www.softwareag.com/en_corporate/blog/streamsets/schemas-data-warehouses-star-galaxy-snowflake.html

PRODUCT
id
name
price
discount
category
brand

TIME
id
day
day_of_week
month
quarter
year

SALES
id
order_id
customer_id
product_id
employee_id
time_id
quantity
total_amount
total_discount
payable_amount
amount_paid
due
payment_mode

CUSTOMER
id
name
mob_num
email_id

EMPLOYEE
id
name
mob_num
email_id
address
department
salary
joining_date
last_working_date

Structura DW „Fulg de Nea” (Snowflake Schema)

- **Variantă mai complexă a structurii stea.**
- **Deosebirea: tabelele dimensiune sunt normalizate, fragmentându-le în sub-tabele suplimentare.**
- **Fiecare sub-tabel stochează un grup de attribute și este legat prin chei străine.**
- **Normalizare: Această fragmentare reduce redundanța datelor, respectând principiile normalizării (de obicei, până la FN3).**
- **Structurile de date sunt mai eficiente din punct de vedere al stocării.**

Fulg de Nea (continuare)

- **Complexitate**: Este mai complexă decât structura stea din cauza multiplelor tabele și relațiilor suplimentare.
 - Interogările pot fi mai complicate și pot necesita mai multe joncțiuni.
- **Performanță**: Deși normalizarea poate reduce redundanțele și poate economisi spațiu de stocare, interogările pot fi mai lente datorită numărului mai mare de joncțiuni necesare pentru a obține datele.
- **Scalabilitate**: Este mai scalabilă din punct de vedere al modelării datelor, deoarece este mai flexibilă și permite adăugarea de noi attribute și relații fără a crește semnificativ redundanța.

„Fulg de Nea” (continuare)

- **„Seminormalizat”**, avantajele modelului relațional.
- Tabelele dimensiune respecta regulile de normalizare din modelul relațional => economie de spațiu,
- nu va conduce la reducerea spațiului pt tabela de fapte
- **Avantaje:**
 - Redundanta redusă
 - Usor de întreținut
- **Dezavantaje:** La cereri de interogare complexe (join) => crește timpul de răspuns

Tabelele Dimensiune și Sub-Dimensiune

- ***Fragmentare: De exemplu, într-o structură stea, un tabel dimensiune Product (ProductID, ProductName, Category, Supplier).***
- ***Într-o structură fulg de nea, acest tabel ar putea fi fragmentat în trei tabele:***
- ***Product (ProductID, ProductName, CategoryID)***
- ***Category (CategoryID, CategoryName)***
- ***Supplier (SupplierID, SupplierName)***
- ***Relații: Tabelele dimensiune sunt legate între ele prin chei străine, creând o rețea de tabele interconectate.***

Avantaje și Dezavantaje

■ Avantaje:

- **Redundanță redusă**: Prin normalizarea dimensiunilor, schema fulg de nea minimizează redundanța datelor și îmbunătățește consistența acestora.
- **Spațiu de stocare**: Utilizează mai eficient spațiul de stocare, deoarece elimină duplicarea datelor.

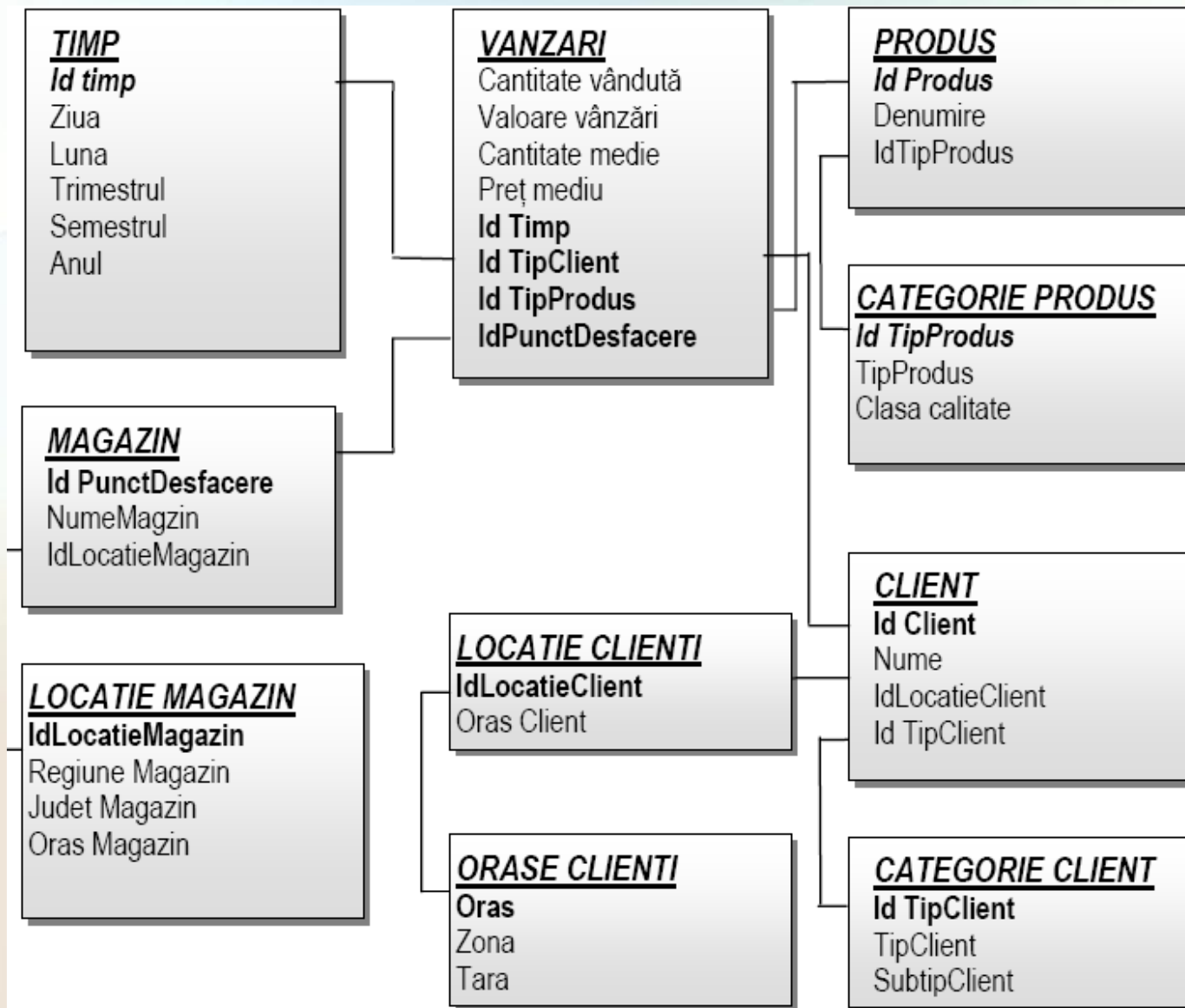
■ Dezavantaje:

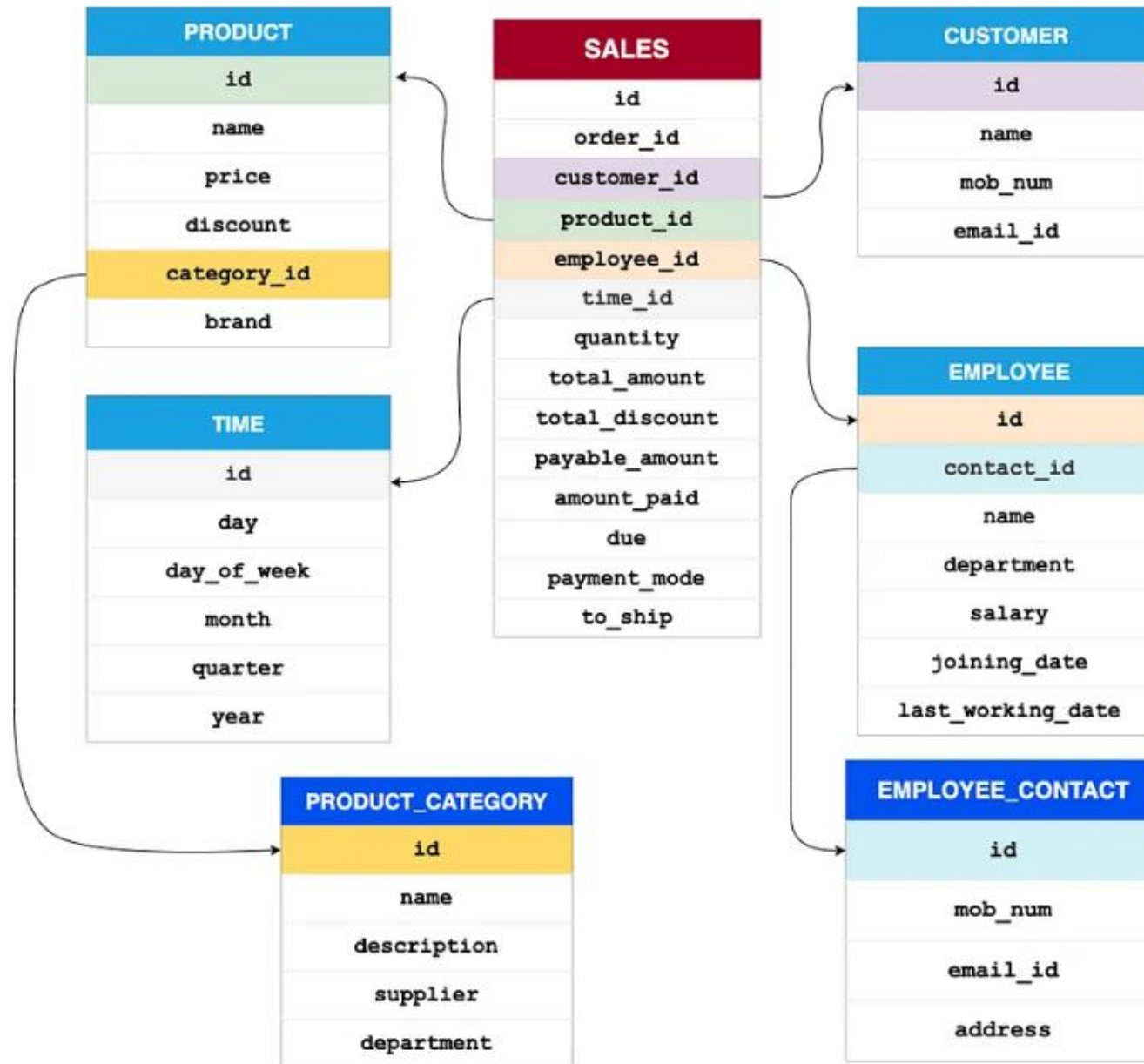
- **Complexitate**: Interogările devin mai complexe și pot necesita mai multe alăturări (joins), ceea ce poate afecta performanța, mai ales în cazul unor interogări complexe.
- **Dificultăți în utilizare**: Designul este mai dificil de înțeles și de utilizat de către utilizatorii non-tehnici.

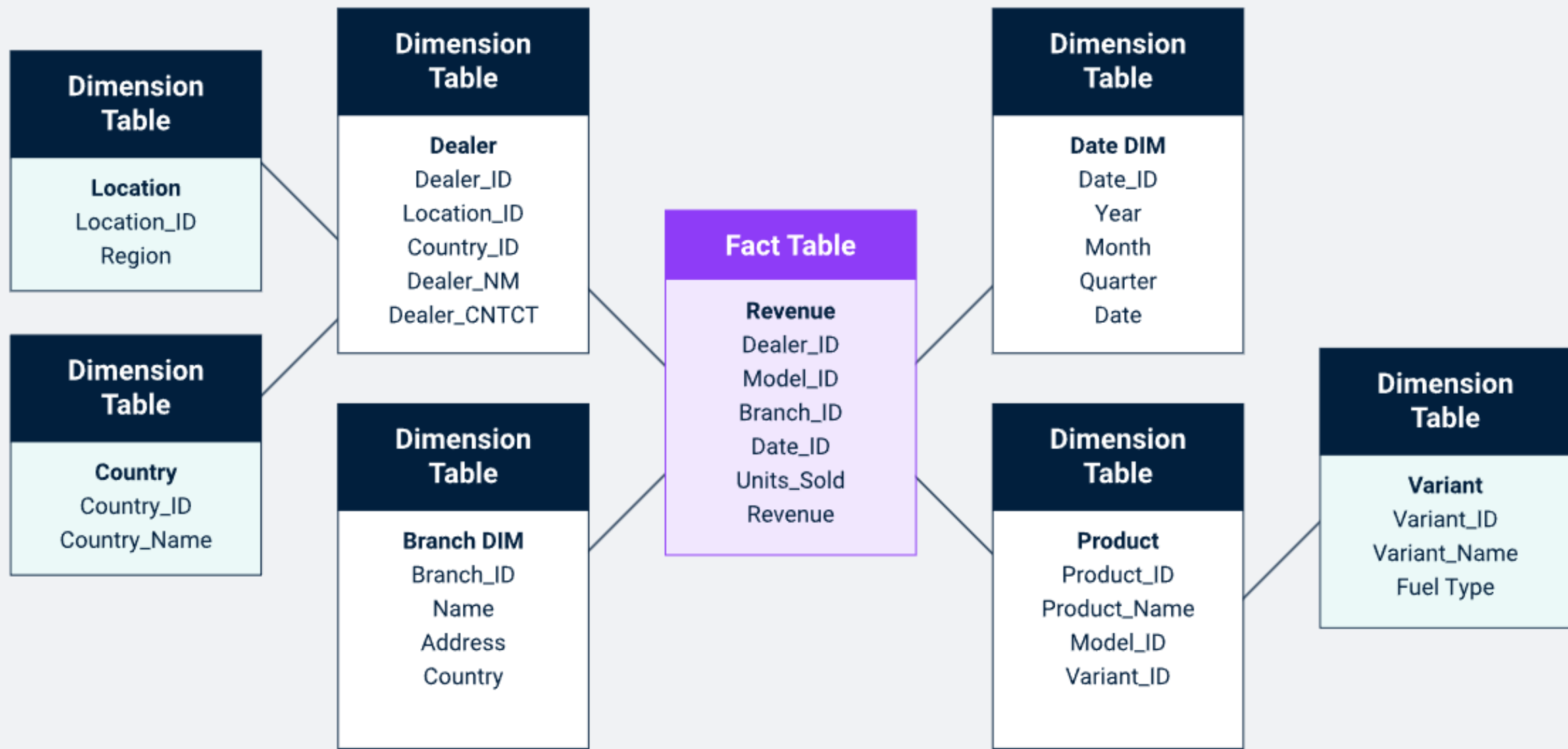
Exemplu de Structură Fulg de Nea

- o companie de retail, structura fulg de nea pentru produsele vândute:
- Tabel Fapt:
- Sales (ProductID, CustomerID, TimeID, Quantity Sold, Sales Amount).
- Tabele Dimensiune:
 - Product (ProductID, ProductName, CategoryID, SupplierID)
 - Category (CategoryID, CategoryName)
 - Supplier (SupplierID, SupplierName)
 - Customer (CustomerID, CustomerName, RegionID)
 - Region (RegionID, RegionName)
 - Time (TimeID, Date, Day, MonthID)
 - Month (MonthID, MonthName, YearID)
 - Year (YearID, Year)

Exemplu „Fulg de zăpadă”





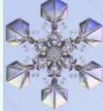


Structura DW – Schema Constelație de fapte

- **Schema galaxie**
- **mai multe tabele de fapte, conectate ce utilizează aceleași tabele-dimensiune**
- **pe lângă tabela de fapte Vânzări, o tabelă suplimentară de fapte Aprovizionări, legată de dimensiuni**

Structura Galaxie (Galaxy Schema), SAU Constelație de fapte (Fact Constellation),

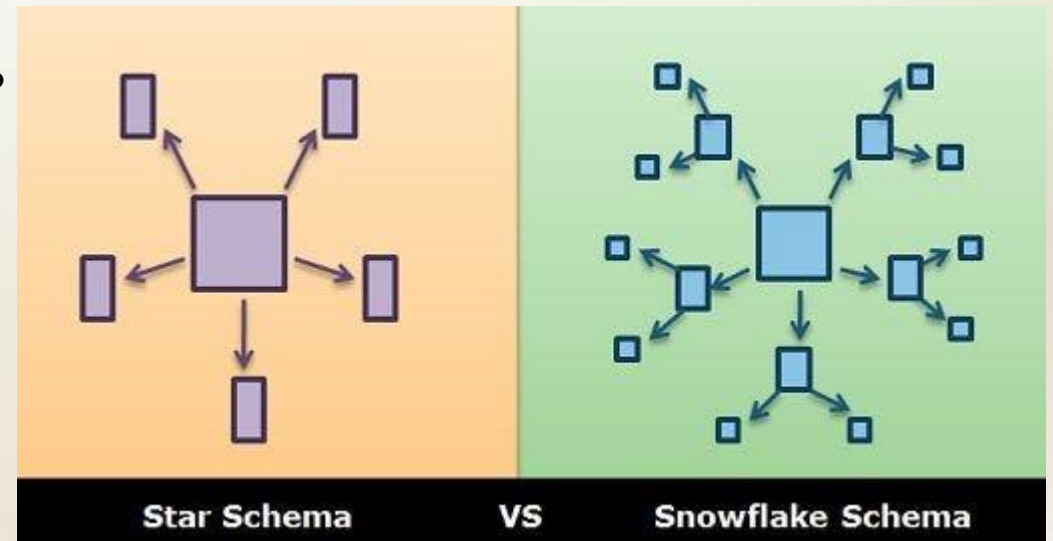
- Model complex utilizat în proiectarea DD.

- Extinde structurile  + 

combină mai multor tabele fapt
care utilizează tabele dimensiune comune,

- Formează o rețea de relații asemănătoare unei constelații sau unei galaxii.

SURSA: <https://nidhiq631.medium.com/star-schema-vs-snowflake-schema-78dc9424a8a2>



Componentele Galaxie

Tabelele fapt:

- ***Există mai multe tabele fapt, fiecare reprezentând un set distinct de măsurători sau evenimente care trebuie analizate.***
- ***Aceste tabele pot conține măsurători diferite, dar împărtășesc una sau mai multe tabele dimensiune comune.***
- ***De exemplu, un tabel fapt pentru vânzări și unul pentru retururi ar putea împărtăși tabelele dimensiune pentru produse și timp.***

Componentele Galaxie (continuare)

Tabelele dimensiune:

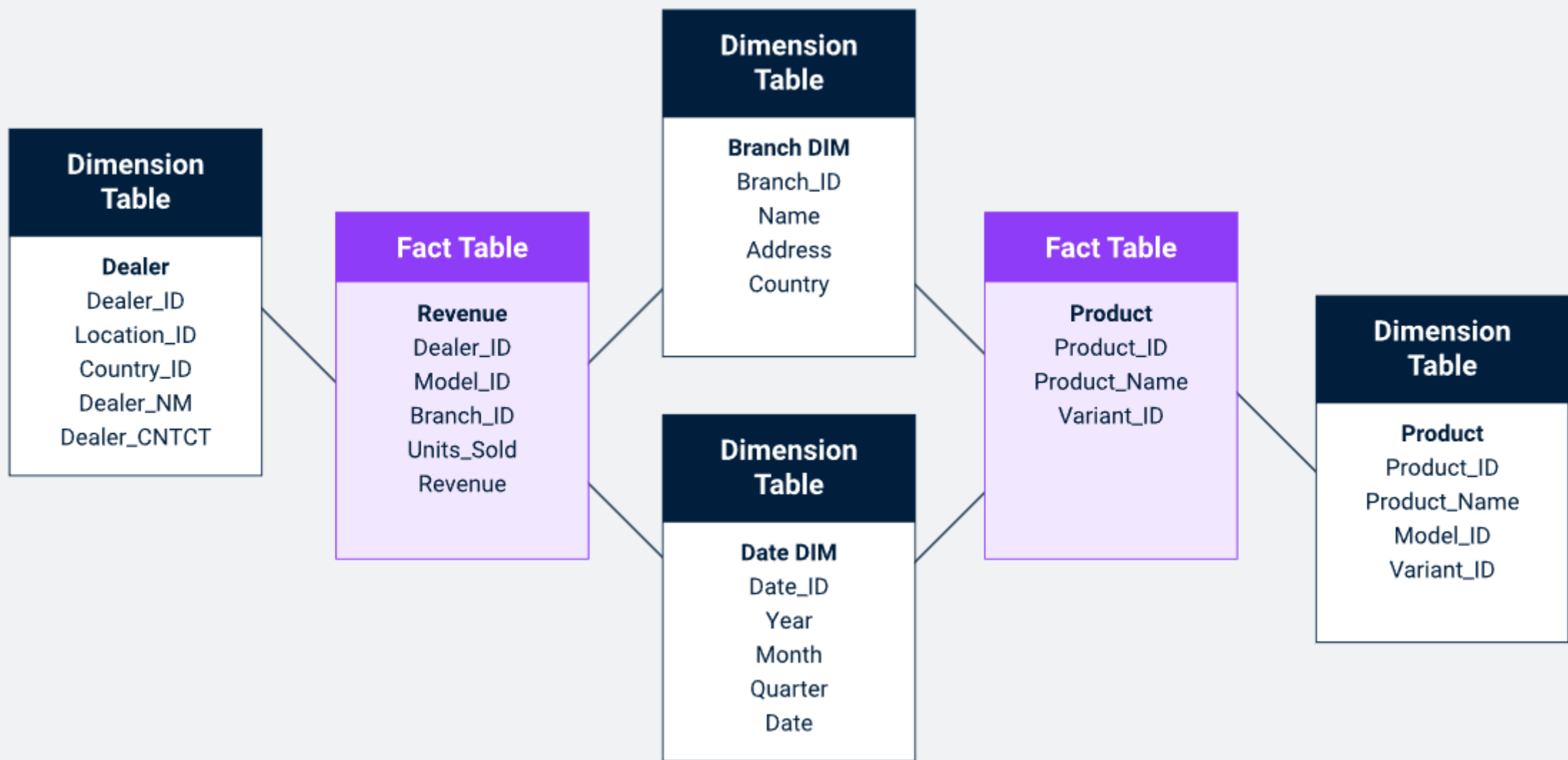
- *Sunt similare cu cele din structurile stea și fulg de nea și sunt folosite pentru a descrie diferitele attribute ale datelor.*
- *În structura galaxie, aceleași tabele dimensiune pot fi utilizate de mai multe tabele fapt.*
- *De exemplu, o tabelă dimensiune Product (Produs) poate fi comună atât pentru tabelul fapt Sales (Vânzări), cât și pentru tabelul fapt Inventory (Inventar).*

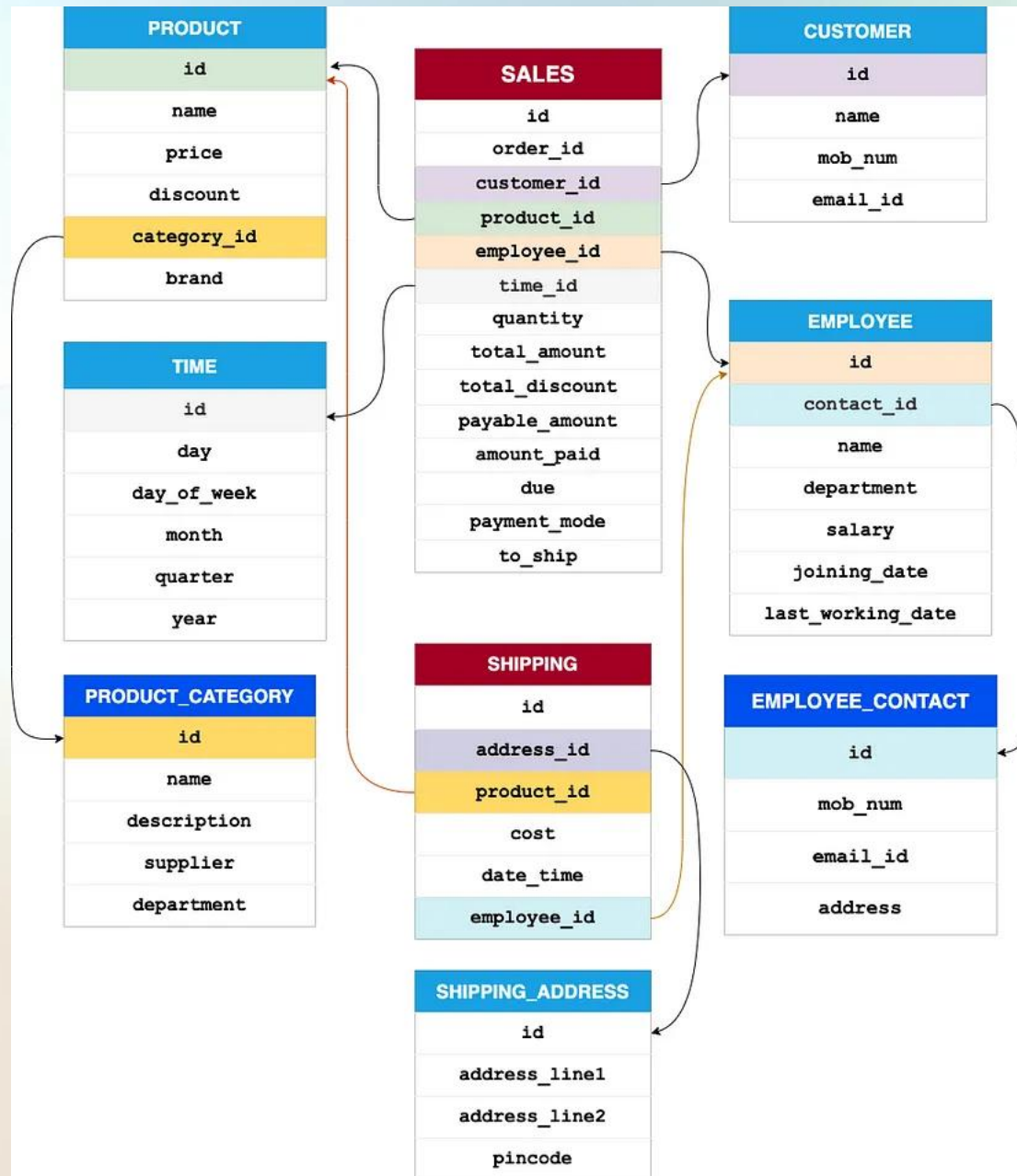
Relații în Galaxie

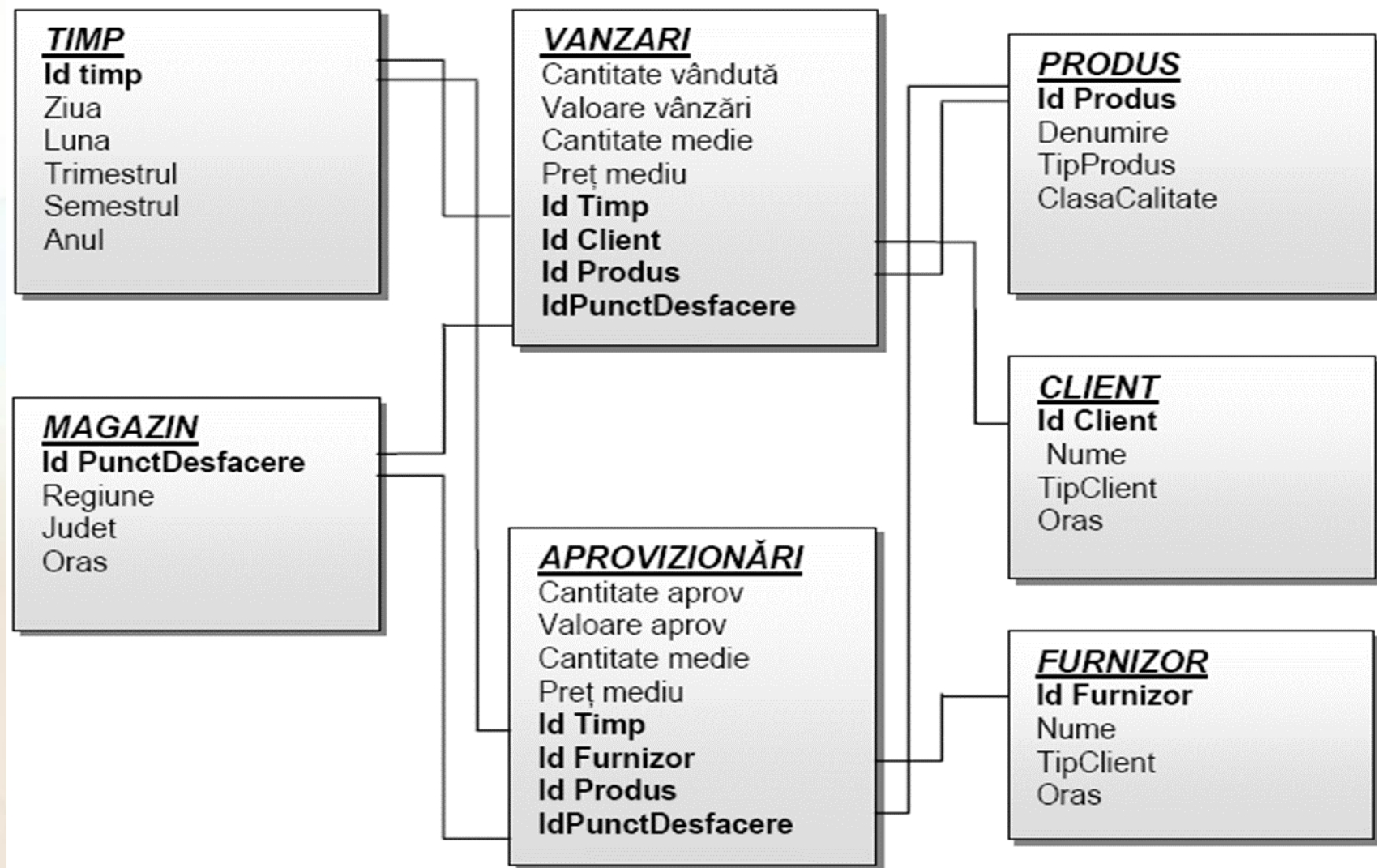
- ***Legătura dintre tabele:***
- ***Fiecare tabel fapt este legat prin chei străine de tabelele dimensiune relevante,***
- ***aceleași tabele dimensiune pot fi utilizate în relații multiple, legându-se la diferite tabele fapt.***

Caracteristici Galaxie

- **Complexitate ridicată**: Este un model complex datorită multiplelor tabele fapt și relațiilor multiple între tabele fapt și tabelele dimensiune.
- **Flexibilitate**: Structura galaxie este foarte flexibilă și scalabilă, permițând adăugarea de noi tabele fapt și dimensiuni fără a modifica semnificativ structura existentă.
- **Redundanță redusă**: Partajarea tabelelor dimensiune între mai multe tabele fapt reduce redundanța datelor și permite o gestionare mai eficientă a acestora.
- **Analyze complexe**: Această structură este ideală pentru scenarii complexe în care este necesară analizarea datelor din perspective multiple și interdependente.







	Schema stea	Schema fulg de nea	Schema Galaxy
Elemente	Tabel de fapte unic conectat la tabele cu mai multe dimensiuni, fără subdimensiuni	Single Fact Table se conectează la tabele cu mai multe dimensiuni care se conectează la mai multe tabele subdimensiuni	Multiple Fact Tables se conectează la tabele cu mai multe dimensiuni care se conectează la mai multe tabele subdimensiuni
Normalizare	Denormalizat	Normalizat	Normalizat
Număr de dimensiuni	Tabelele cu mai multe dimensiuni se mapează la un singur tabel de fapte	Tabelele cu mai multe dimensiuni sunt asociate cu tabelele cu mai multe dimensiuni	Tabelele cu mai multe dimensiuni se mapează cu mai multe tabele de fapte
Redundanța datelor	Ridicat	Scăzut	Scăzut
Performanță	Mai puține chei străine, crește performanța	Performanță scăzută în comparație cu Stea de la un număr mai mare de chei externe	Performanță mai scăzută. Folosit pentru agregarea de date complexe.
Complexitate	Simplu, ușor de înțeles	Mai complicat, mai dificil de înțeles	Cel mai complicat de înțeles. Rezervat pentru structurile de date foarte complexe
Utilizarea depozitării	Spațiu de stocare mai mare datorită redundanței datelor	Spațiu de stocare redus din cauza redundanței limitate a datelor	Utilizarea redusă a spațiului de stocare în comparație cu nivelul de sofisticare din cauza redundanței limitate a datelor
Limitări de proiectare	Un singur tabel de fapte, fără subdimensiuni	Un singur tabel de fapte, sunt permise mai multe subdimensiuni	Sunt permise mai multe tabele de fapte, sunt permise doar dimensiunile de la I nivel

ETL - Extract, Transform, Load

- *Proces fundamental în domeniul integrării datelor,*
- *utilizat pentru a colecta date din diverse surse,*
- *transformarea într-un format unificat și consistent, și*
- *încărcarea într-o destinație specifică (BD sau DD).*
- *Procesul ETL este esențial pentru orice organizație care dorește să utilizeze datele sale pentru a obține informații valoroase și a îmbunătăți procesele de afaceri.*
- *Prin extragerea, transformarea și încărcarea datelor într-un mod eficient, organizațiile pot crea o bază solidă pentru analiza datelor și luarea deciziilor strategice.*

Etapele procesului ETL

(1). Extract (Extragerea): *Identificarea surselor de date.*

(2). Transform (Transformarea): *Curățarea datelor: Se elimină datele duplicate, incomplete sau incorecte.*

- *Integrarea datelor: Datele din diferite surse sunt integrate într-un singur set de date, asigurându-se că există o coerență între diferitele scheme de date.*

(3). Load (Încărcarea): *Definirea destinației-baza de date / depozitul de date în care vor fi încărcate datele transformate.*

Extract (Extragerea)

- **Identificarea surselor** de date: Se identifică toate sursele de date relevante, care pot fi fișiere text, baze de date, aplicații, sau chiar fluxuri de date în timp real.
- **Conectarea la surse**: Se stabilesc conexiunile necesare pentru a accesa datele din fiecare sursă.
- Aceasta poate implica utilizarea de drivere ODBC, JDBC sau API-uri specifice.
- **Extragerea datelor**: Datele sunt extrase din fiecare sursă și stocate într-un format intermediar, de obicei într-o zonă de staging.

Transform (Transformarea datelor)

- **Curățarea datelor**: Se elimină datele duplicate, incomplete sau incorecte.
- Se corectează erorile de formatare și se standardizează valorile.
- **Integrarea datelor**: Datele din diferite surse sunt combinate într-un singur set de date, asigurându-se că există o coerență între diferitele scheme de date.
- **Calcularea valorilor derivate**: Se pot calcula noi valori pe baza datelor existente, cum ar fi totaluri, medii sau procente.
- **Enriching**: Se pot adăuga informații suplimentare datelor existente, cum ar fi date geografice sau informații despre clienți.

Load (Încărcarea)

- **Definirea destinației**: Se specifică baza de date sau depozitul de date în care vor fi încărcate datele transformate.
- **Încărcarea datelor**: Datele transformate sunt încărcate în destinația specificată, respectând structura și formatul predefinit.
- **Optimizarea performanței**: Se pot utiliza tehnici de încărcare în bloc sau de încărcare incrementală pentru a îmbunătăți performanța procesului.

Importanța procesului ETL

- **Integrarea datelor**: Permite combinarea datelor din diferite surse într-o singură sursă de adevăr.
- **Calitatea datelor**: Asigură că datele sunt curate, consistente și complete.
- **Analiza datelor**: Oferă date de înaltă calitate pentru a fi utilizate în procesele de analiză și raportare.
- **Suport pentru decizii**: Permite luarea de decizii mai informate pe baza datelor integrate și analizate.

Inteligența Artificială în SSD



- *Rolul AI în sistemele de suport decizii.*
- *Implementarea AI în procesele de decizie și optimizare.*
- *Algoritmi de decizie bazate pe AI.*

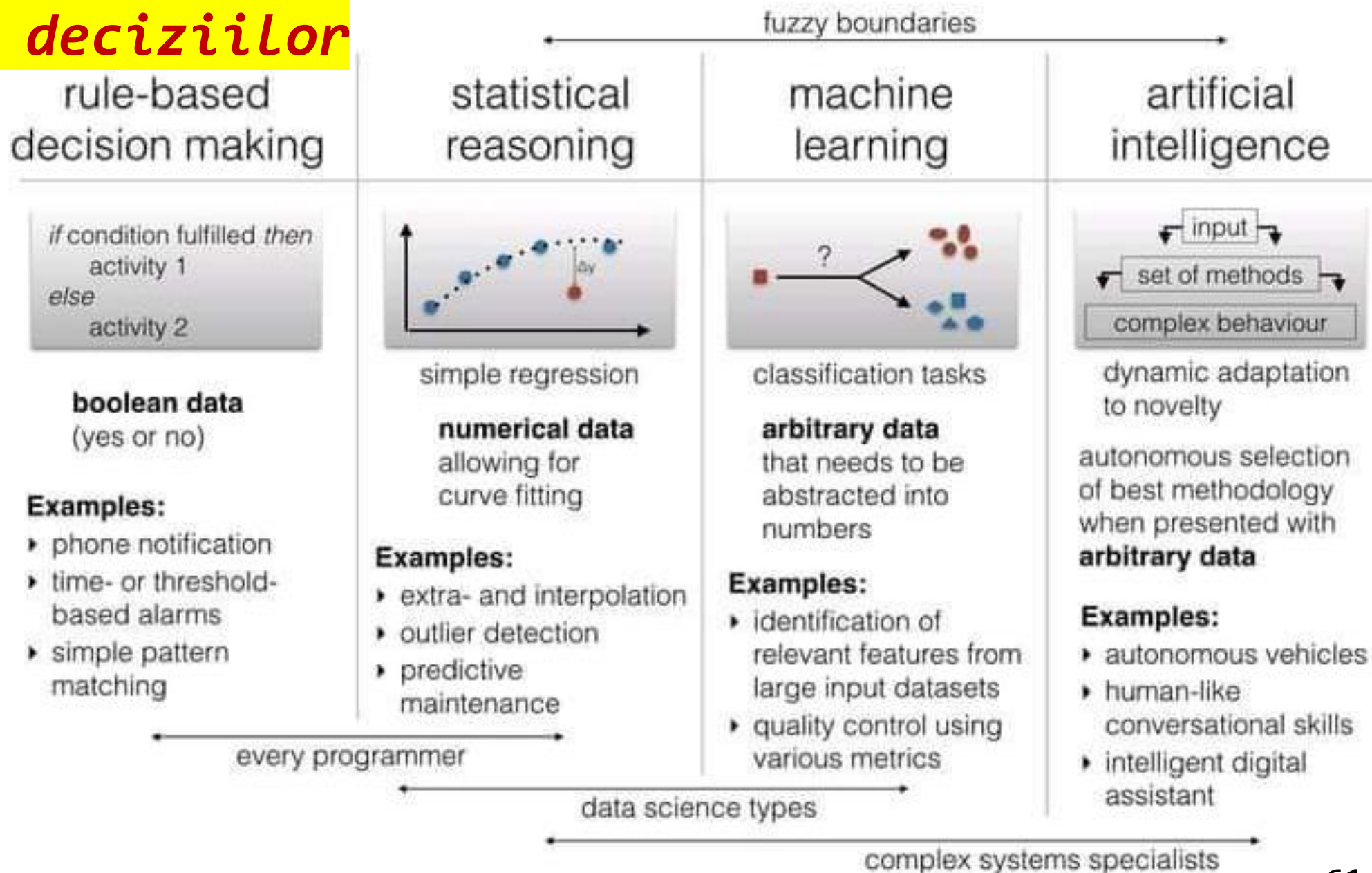


Sursa: <https://www.aifire.co/p/top-8-vector-databases-for-ai-data-machine-learning-projects-in-2024>

MACHINE LEARNING

- *Aplicarea ML în crearea de modele predictive pentru decizii.*
- *Algoritmi de învățare automată orientați spre suportul deciziilor.*
- *Integrarea ML în SSD.*

Algoritmi de luare a deciziilor



DEEP LEARNING



DATA MINING



- *Aplicarea Data Mining pentru a extrage cunoștințe relevante pentru decizii.*
- *Tehnici de Data Mining pentru îmbunătățirea proceselor decizionale.*
- *Studiu de caz: utilizarea Data Mining-ului în deciziile de afaceri.*





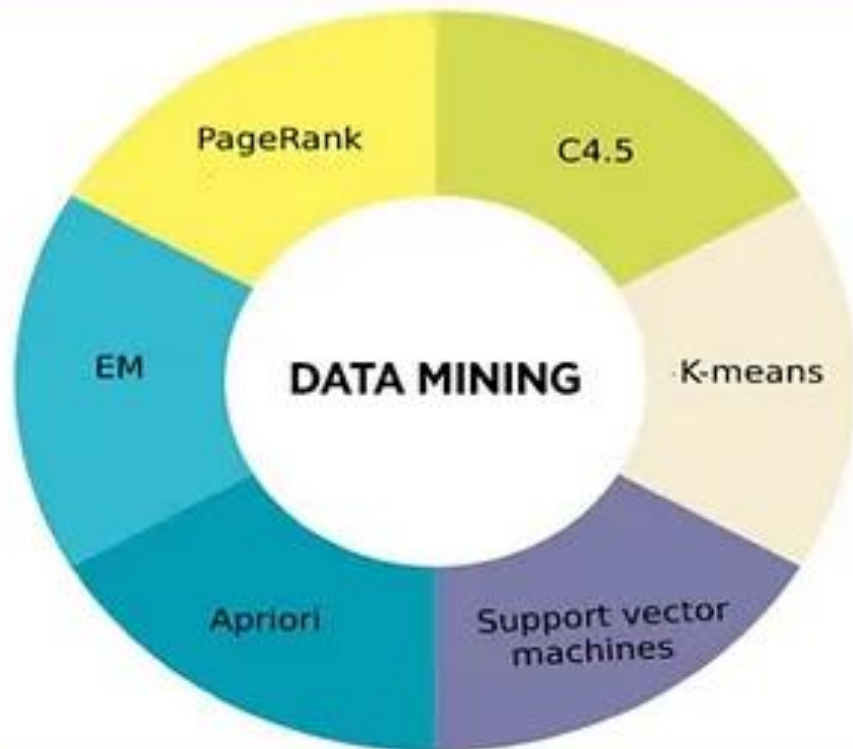




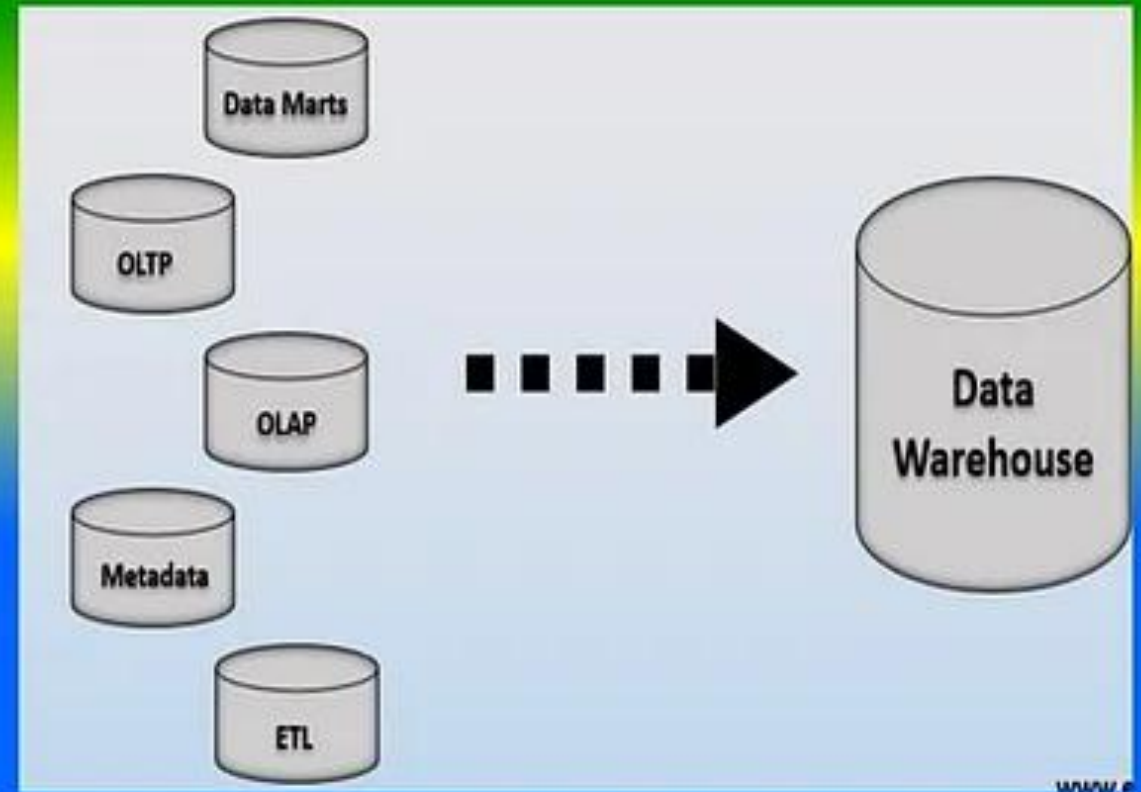


DM vs DW

Data Mining



Data Warehousing



TEXT MINING

- *Utilizarea TM pentru a extrage informații utile pentru decizii.*
- *Tehnici de TM aplicate în analizarea documentelor pentru suportul deciziilor.*
- *Aplicații în identificarea și analiza informațiilor critice din date textuale pentru decizii.*

- Ramură a inteligenței artificiale și a lingvisticii computaționale care proces de analiză automată a unor cantități mari de date textuale și extragerea informațiilor utile. modele și a dobândirii de cunoștințe relevante ascunse.
- Scopul principal este de a transforma datele textuale neorganizate și nestructurate într-o formă structurată, ușor de analizat și interpretat.
- TM este utilizat pe scară largă în diverse domenii, inclusiv marketing, sănătate, securitate și știință, pentru a transforma volume mari de date textuale în informații acționabile.

Aplicații ale Text Mining

- ***Marketing***: Analiza opiniilor clienților, segmentarea pieței, descoperirea de noi tendințe.
- ***Resurse umane***: Screening-ul CV-urilor, analiza feedback-ului angajaților.
- ***Sănătate***: Extragerea informațiilor din dosarele medicale, descoperirea de noi tratamente.
- ***Finanțe***: Analiza rapoartelor financiare, detectarea fraudelor.
- ***Drept***: Analiza documentelor juridice, descoperirea de precedente.

Pașii principali ai text mining

- ***Preprocesare;***
- ***Tokenizare;***
- ***Lematizare și stemming;***
- ***Transformarea datelor;***
- ***Extragerea caracteristicilor;***
- ***Analiza și interpretarea.***

1. Preprocesare

- **Curățarea textului**: Eliminarea elementelor irelevante, cum ar fi punctuația, cuvintele comune (stopwords), și corectarea erorilor de scriere.
- **Tokenizare**: Împărțirea textului în unități de bază (cuvinte, fraze).
- **Lematizare și stemming**: Reducerea cuvintelor la forma lor de bază sau rădăcină pentru a unifica variațiile gramaticale.

2. Transformarea datelor

- **Vectorizarea textului**: Conversia textului în reprezentări numerice utilizând metode precum bag-of-words, TF-IDF (Term Frequency-Inverse Document Frequency), sau embeddings de cuvinte (Word2Vec, GloVe).
- **Reducerea dimensiunii**: Reducerea numărului de caracteristici pentru a simplifica analiza, utilizând tehnici precum PCA (Principal Component Analysis).

3. Extragerea caracteristicilor

- **Identificarea tiparelor și relațiilor:**

Găsirea și extragerea relațiilor semantice, tiparelor sau tendințelor din text.

- **Clasificare și categorizare:**

Etichetarea automată a textului pe baza caracteristicilor extrase.

4. Analiza și interpretarea

- **Clasificarea textului**: Atribuirea textului la categorii predefinite utilizând algoritmi de învățare automată.
- **Detecția emoțiilor și sentimentelor**: Analiza sentimentelor pentru a determina atitudinea autorului față de un subiect (pozitiv, negativ, neutru).
- **Clustering și topic modeling**: Gruparea documentelor în funcție de subiecte sau teme similare utilizând algoritmi de clustering sau topic modeling (de exemplu, LDA - Latent Dirichlet Allocation).

Aplicații comune

- 1) Analiza sentimentelor***: Utilizată în special în rețelele sociale, recenzii de produse, și sondaje pentru a înțelege percepția publică.
- 2) Detectia fraudelor***: Examinarea documentelor pentru a identifica anomalii sau activități frauduloase.
- 3) Recomandări personalizate***: Analistii de marketing pot utiliza text mining pentru a personaliza recomandările de produse pe baza intereselor utilizatorilor.
- 4) Clasificarea documentelor***: Automatează organizarea și clasificarea documentelor în funcție de conținutul lor.

Text Mining în SSD

- *TM permite companiilor să valorifice cantități mari de date nestructurate, cum ar fi e-mailuri, rapoarte, recenzii ale clienților, și postări pe rețelele sociale.*
- *Prin analiza acestor date, organizațiile pot lua decizii mai informate și pot optimiza diverse procese de afaceri.*

1. Analiza sentimentelor și a reputației

- **Monitorizarea brandului**: Text mining-ul poate analiza postările de pe rețelele sociale, recenziile online și articolele de presă pentru a determina cum percepe publicul un brand.
- **Analiza sentimentelor** poate identifica tendințele pozitive sau negative, permițând managerilor să reacționeze rapid la orice problemă de imagine.
- **Feedback-ul clienților**: Companiile pot analiza comentariile și recenziile clienților pentru a identifica probleme comune cu produsele sau serviciile, îmbunătățind astfel calitatea și satisfacția clienților.

2. Suport pentru Luarea deciziilor în marketing și vânzări

- **Segmentarea pieței**: Text mining-ul poate analiza datele textuale pentru a identifica segmente de piață pe baza comportamentului și preferințelor clienților. Acest lucru permite personalizarea ofertelor și campaniilor de marketing.
- **Predictive analytics**: Prin analiza textelor din surse precum e-mailuri de vânzări sau conversații de suport, se pot construi modele predictive care să estimeze probabilitatea unui client de a achiziționa un produs sau de a părăsi compania.

3. Managementul riscului și conformității

- **Detectarea fraudei**: Text mining-ul poate fi utilizat pentru a analiza documente și e-mailuri interne pentru a identifica anomalii care ar putea indica activități frauduloase sau încălcări ale politicilor de conformitate.
- **Monitorizarea reglementărilor**: În industriile reglementate, cum ar fi financiar-bancară sau farmaceutică, text mining-ul poate ajuta la monitorizarea și interpretarea modificărilor în reglementările legale și la asigurarea conformității.

4. Îmbunătățirea relațiilor cu clienții

- **Suportul automatizat al clienților:** Utilizarea text mining-ului în chatbots și sisteme de asistență automată permite recunoașterea rapidă a problemelor comune ale clienților și oferirea de soluții adecvate, îmbunătățind astfel eficiența serviciului de suport.
- **Analiza cererilor de servicii:** Prin analiza cererilor textuale de suport tehnic, companiile pot identifica tiparele de probleme, ceea ce poate duce la optimizarea produselor sau a serviciilor oferite.

5. Optimizarea operațiunilor interne

- **Analiza documentelor și automatizarea proceselor:** Companiile pot folosi TM pentru a analiza documente interne, cum ar fi rapoarte de performanță, contracte, și politici, pentru a identifica informațiile critice și pentru a automatiza fluxurile de lucru repetitive.
- **Managementul cunoștințelor:** TM poate fi utilizat pentru a organiza și structura volume mari de informații din baze de date textuale, facilitând accesul rapid la cunoștințele organizaționale.

6. Previziuni economice și de piață

- **Analiza trendurilor**: Prin extragerea și analizarea datelor textuale din știri, rapoarte economice și alte surse relevante, companiile pot anticipa schimbările de piață și pot ajusta strategiile de afaceri.
- **Sentimentul economic**: TM poate ajuta la evaluarea sentimentului pieței sau al investitorilor prin analiza comentariilor din surse financiare, ajutând astfel la luarea deciziilor strategice de investiție.

Integrarea Text Mining în SSD

- ***Prin utilizarea TM, companiile pot obține o înțelegere mai profundă a datelor nestructurate și pot lua decizii mai bine informate, care contribuie la creșterea competitivității și eficienței în afaceri.***
- ***Pentru a integra text mining-ul în aplicațiile informatice de management al afacerilor și asistarea deciziilor, companiile pot urma următoarele diverse strategii. ***

Integrarea TM în SSD (continuare)

(1). **Dezvoltarea de soluții personalizate**: Crearea de sisteme personalizate care să folosească text mining-ul pentru a extrage date din sursele relevante și a le integra în fluxurile de lucru existente (ERP, CRM).

(2). **Utilizarea platformelor existente**: Integrarea funcționalităților de text mining oferite de platforme existente, cum ar fi IBM Watson, Microsoft Azure Text Analytics, sau Google Cloud Natural Language, pentru a analiza și interpreta datele textuale din cadrul organizației.

Integrarea TM în SSD (continuare)

- **Automatizarea proceselor**: Integrarea text mining-ului în RPA (Robotic Process Automation) pentru automatizarea sarcinilor repetitive care implică analiza textului, cum ar fi procesarea facturilor sau revizuirea contractelor.
- **Formarea echipelor de analiză a datelor**: Pregătirea specialiștilor în text mining și analiză de date pentru a dezvolta și implementa soluții care să sprijine procesul decizional în afaceri.

Instrumente pentru Text Mining

Există o varietate de instrumente și biblioteci open-source și comerciale pentru text mining, precum:

- **NLTK** (Natural Language Toolkit): *O bibliotecă Python pentru procesarea limbajului natural.*
- **spaCy**: *O bibliotecă Python pentru procesarea limbajului natural industrial.*
- **Stanford NLP**: *Un set de instrumente Java pentru procesarea limbajului natural.*
- **RapidMiner**: *O platformă de data mining cu module pentru text mining.*

	Data Warehouse	Data Lake	Data Lakehouse
Storage Data Type	Works well with structured data	Works well with semi-structured and unstructured data	Can handle structured, semi-structured, and unstructured data
Purpose	Optimal for data analytics and business intelligence (BI) use-cases	Suitable for machine learning (ML) and artificial intelligence (AI) workloads	Suitable for both data analytics and machine learning workloads
Cost	Storage is costly and time-consuming	Storage is cost-effective, fast, and flexible	Storage is cost-effective, fast, and flexible
ACID Compliance	Records data in an ACID-compliant manner to ensure the highest levels of integrity	Non-ACID compliance: updates and deletes are complex operations	ACID-compliant to ensure consistency as multiple parties concurrently read or write data



DATA
LOGO TAGLINE



DATA
LOGO TAGLINE



DATA
LOGO TAGLINE



DATA
LOGO TAGLINE