**MATERIAL PENTRU SEMINARUL 1**

**Schema de lucru**

**CUM PROIECTĂM UN DWH**

1. **DEPOZITE DE DATE - NOTIUNI**

 Un depozit de date este o structură ce conţine mai multe colecţii de date, de provenienţă diferită, având dimensiuni foarte mari, folosite pentru deciziile luate la nivel tactic şi strategic în cadrul unei organizaţii. De regulă, acestea există separat de bazele de date operaţionale (bazate pe modelul relaţional). Înainte de a fi trasferate din sursele de date, informaţiile sunt supuse unor procese **ETL** (Extract, Transform, Load) prin care sunt filtrate şi prelucrate pentru a respecta anumite convenţii stabilite la proiectarea depozitelor de date. Această tehnologie este folosită în conjuncţie cu data mining, astfel încât volumul de informaţie stocat în depozitele de date este utilizat pentru generarea de cunoştinţe, motiv pentru care domeniul care se ocupă cu astfel de probleme, poartă denumirea de **Business Intelligence**.

 În limba română, acest domeniu din tehnologia informaţiei a fost tradus prin *inteligenţa afacerii* (există chiar şi un volum care preia această denumire, avându-i ca autori pe Manole Velicanu şi Gheorghe Matei și apărut Editura Academiei de Studii Economice, Bucureşti, 2010), însă această terminologie nu a beneficiat de asentimentul tuturor specialiştilor în domeniu.

 Frecvent, modelului relaţional folosit în bazele de date (care este reprezentat în 2 dimensiuni: atribute şi înregistrări) i se adaugă şi dimensiunea timp, astfel că informaţiile reţinute au caracter istoric şi sunt orientate pe subiecte.

 Merită reţinute şi **definiţiile** date pentru depozite de date de către cercetătorul american William Harvey Inmon, considerat părintele acestei tehnologii precum şi de către consiliul OLAP. Potrivit lui W. H. Inmon, *depozitul de date este o colecţie de date orientate pe subiecte, integrate, istorice şi nevolatile, destinate sprijinirii procesului de luare a deciziilor manageriale*. Consiliul OLAP defineşte depozitul de date ca fiind *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*.

 Se poate concluziona că depozitul de date reprezintă soluţia optimă pentru organizarea datelor în cazul unor aplicaţii de tip sisteme suport pentru decizii sau sisteme informatice executive. Rolul său constă în a oferi o imagine coerentă asupra datelor relative la activitatea unei organizaţii şi a contextului în care aceasta acţionează.

 Costul dezvoltării unui depozit de date este destul de mare, iar investiţia se amortizează într-un orizont de timp mai mare.

 Frecvent, pentru proiectarea şi implementarea unui depozit de date se cheltuie câteva milioane de dolari. Distribuţia acestei investiţii este destul de echilibrată: 1/3 se cheltuie cu hardware-ul, 1/3 cu software-ul, iar 1/3 cu serviciile profesionale.

 Recuperarea investiţiei nu se poate măsura direct, ci se realizează din profitul operaţional rezultat din deciziile luate pe baza cunoştinţelor sintetizate din depozitul de date care aduc companiei un avantaj competiţional.

 Un indicator al utilităţii depozitului de date în contextul organizaţiei este reprezentat de dimensiunea sa, care ar trebui să se dubleze (ca urmare a operaţiilor ETL) într-un interval de aproximativ 12-18 luni.

 Tipic, dimensiunea unui depozit de date variază între câţiva zeci şi câteva sute de terabytes. Dacă constrângerile de flexibilitate o impun, se poate adopta soluţia unui data mart care conţine numai datele relevante pentru o anumită analiză.

 În prezent, datorită soluţiilor de integrare a datelor, procesul **ETL** tinde să se transforme în **ELT**, (**diferenta vezi in Anexa 1**) adică încărcarea datelor se face înainte de transformarea lor, datorită capabilităţilor (atât hardware, cât şi software) de a realiza procesări la nivelul serverului de date.

 **Principalele caracteristici ale unui depozit de date sunt:**

* asigură accesul la datele organizaţiei într-un timp cât mai scurt; pentru obţinerea unui astfel de deziderat, datele sunt în mod intenţionat denormalizate, astfel că este permisă redundanţa datelor (spre diferenţă de bazele de date relaţionale);
* permite utilizarea informaţiilor conţinute fără alte prelucrări suplimentare căci aceste procese au fost realizate la construcţia depozitului de date (mecanismul ETL – numai după ce s-a asigurat calitatea datelor, acestea pot fi folosite); în acest sens, este de asemenea deosebit de important ca datele să fie consistente;

 Calitatea datelor reprezintă un factor extrem de important în procesul de analiză. Eşecul multor depozite de date a fost cauzat tocmai de utilizarea unor date incomplete sau irelevante pentru domeniul relativ la care a fost construit depozitul de date.

 Într-un depozit de date, consistenţa datelor se manifestă prin aceea că, la interogări identice realizate de utilizatori diferiţi să fie oferite aceleaşi informaţii. Situaţia în care datele nu sunt coerente se datorează faptului că procesul de încărcare nu a fost complet.

* utilizează date istorice pentru identificarea unor tendinţe ce pot fi apoi folosite pentru realizarea de prognoze; de aceea, orice informaţie va avea asociat şi un atribut referitor la timp;

Orizontul de timp al informaţiilor reţinute într-un depozit de date acoperă o perioadă cuprinsă între 5 şi 10 ani.

* foloseşte informaţii care sunt orientate pe subiectele procesului economic (clienţi, furnizori, produse) mai degrabă decât pe aplicaţii, ca în cazul bazelor de date opraţionale;
* realizează operaţii de actualizare extrem de rar, acestea constând exclusiv în adăugarea de informaţii la depozitul de date şi nicidecum la modificarea celor existente;
* oferă acces la date doar pentru citire.

 Într-un depozit de date se disting **datele detaliate**, relativ recente din punct de vedere al încărcării, **date agregate** (sintetizate) – care presupun un grad de prelucrare în prealabil – precum şi metadatele prin care se precizează structura, provenienţa, regulile de agregare ale informaţiilor reţinute.

 Datele agregate sunt supuse unor procese de consolidare, totalizare, împachetare şi determină creşterea redundanţei informaţiilor reţinute la nivelul depozitului de date.

 Metadatele sunt folosite pe tot parcursul ciclului de viaţă al depozitului de date (pentru operaţii de încărcare, consultare, actualizare).

## [SOLUTII DE TIP DATAWAREHOUSE PENTRU ERP, CRM](http://www.bion-team.ro/servicii-mob/solutii-de-tip-datawarehouse-pentru-erp-crm)

## http://www.bion-team.ro/images/Solutii-de-tip-DataWarehouse-pentru-ERP-CRM.jpg

 Utilizarea de pachete de aplicaţii este în creştere rapidă în toate domeniile, în special în sistemele de  front-office (CRM) şi sistemele back-office (ERP). Această creştere are loc nu numai pentru pachete de aplicaţii care se ocupă de procesarea tranzacţiilor de afaceri(BTx) , ci și pentru soluţii de tip business intelligence (BI) şi data warehouse-ing (DW).

 BI nu mai este folosit doar pentru a face raportare strategică / tactică şi penru analize ci și pentru conducerea şi optimizarea proceselor operaționale de afaceri şi fluxurilor de lucru. BI nu mai este doar ceva opțional ci devine esenţial pentru succesul afacerii.
 O tendinţă printre utilizatorii de pachete de aplicaţii  BTx (CRM, ERP) este de a utiliza aplicații BI de tip real-time pentru a îmbunătăţi şi optimiza operaţiunile de afaceri. Exemple de aplicații incluse aici: detectarea fraudelor, up-selling, cross-selling, evaluarea impactului comenzilor întârziate, a stabilirii preţurilor la produse, cote de vânzări, şi aşa mai departe. În unele aplicaţii, acest lucru implică integrarea proceselor de BI în soluțiile de tip BTx, în timp ce în alte situaţii atrage după sine interconectarea BTx cu procesele externe de BI.

 Utilizarea în timp real a aplicațiilor(cererilor) de BI de către BTX este uneori numit

“in-line analytics”.

 Procesele externe BI sau rezultatele lor pot fi folosite în fluxurile de tranzacţii de afaceri. Acest lucru permite utilizarea unui sistem BI  de înaltă performanţă pentru a optimiza procesele de zi cu zi  prin oferirea  suplimentară de informații contextuale (de exemplu probabilitatea unui client de a pleca la concurență).

 Utilizarea BI pentru a furniza informații demografice adiționale, a interacţiunilor istorice sau de predicţie a comportamentului viitor al clientului, adaugă contextul necesar pentru a optimiza procesul de afaceri. Problema este că multe procese de luare a deciziilor nu sunt inventariate și centralizate. Trebuie să se înţeleagă procesele de afaceri  pentru a exploata această abordare. Deşi aplicațiile de tip “in-linie analytics” pot fi folosite pentru multe procese operaţionale, deocamdată ele se folosesc frecvent în sistemele de front-office.
 Business Intelligence nu mai este folosit doar pentru raportare, există o cerere tot mai mare pentru soluții analitice pentru a conduce fluxurile de lucru operaţionale şi acţiunile din departamentele de front-office.

 Inginerii de suport pentru clienţi, de exemplu, au nevoie de perspective relevante despre clienții cu care vorbesc. Ei au nevoie să ştie dacă un client are potențial sau o valoare mare sau este probabil să renunțe la servicii/produse, în scopul de a realiza interacţiunea cea mai benefică pentru companie.

**Tipuri de procese de afaceri si sistemele de procesare a datelor:**

* Tranzacţii de afaceri (BTx) – pentru pentru operațiunile zilnice
* Business Intelligence (BI) - pentru procesele de analiză și optimizarea operaţiunilor
* Procese de colaborare - pentru comunicarea şi schimbul de informaţii cu privire la operaţiunile de afaceri.



Atunci când sunt combinate, aceste trei tipuri de procese permit utilizatorilor gestiunea şi comunicarea eficientă cu privire la toate aspectele afacerii.

1. **ARHITECTURA DEPOZITELOR DE DATE**

 Structural, un depozit de date poate fi descris în funcţie de componentele pe care le deţine, în funcţie de nivelurile pe care este organizat precum şi raportat la arhitectura funcţională.

 Din punctul de vedere al componentelor care îl formează, distingem **sursele de date**, **depozitul de date** şi **instrumentele de acces** (interfeţe de analiză).



 În categoria **surslor de date**, un aport important îl au bazele de date operaţionale ale organizaţiei, la care se adaugă amprenta de timp corespunzătoare, precum şi diferite informaţii arhivate sau date externe, referitoare la domeniul economic în care îşi desfăşoară activitatea organizaţia sau cu privire la clienţii, respectiv partenerii de afaceri ai acesteia.

Întrucât sursele de date au – de cele mai multe ori – caracter eterogen, acestea trebuie supuse unor **procese de transformare** înainte de a fi transferate propriu-zis în depozitele de date urmând a fi folosite fie direct, fie sub forma unor informaţii agregate.

##



**Procesul ETL presupune, aşadar, următoarele etape:**

* extragerea datelor din diferitele surse şi transferarea lor, folosind acelaşi mecanism de reprezentare;
* transformarea datelor: eliminarea datelor incorecte sau incomplete precum şi a celor nerelevante pentru organizaţia respectivă;
* încărcarea datelor în depozitul de date, urmând ca acestea să fie folosite pentru operaţiile de agregare; de regulă, operaţiile de agregare vizează unele criterii ce ţin de modul în care informaţiile vor fi utilizate cel mai probabil în contextul depozitului de date.

Având în vedere aceste aspecte, o problemă importantă este reprezentată de **mecanismul de integrare a datelor**, astfel încât acestea să asigure o viziune consistentă şi coerentă a organizaţiei pentru orizontul de timp vizat. Prin urmare trebuie avute în vedere următoarele aspecte:

* un mecanism de codificare unic (un sistem de coduri pentru fiecare câmp, acelaşi pentru toate înregistrările ce au semnificaţia respectivă, indiferent de provenienţa lor);
* acelaşi sistem de unităţi de măsură;
* un mod de organizare fizică a datelor unic rezultat după stabilizarea specificată în proiectarea depozitului de date;
* convenţii pentru denumirea atributelor.

 Astfel stocate în depozitele de date (deopotrivă date preluate din diferite surse de date şi informaţii agregate), informaţiile pot fi exploatate în mai multe moduri, în funcţie de scopul în care se doresc a fi folosite. Se pot utiliza data mart-uri care sunt constituite dintr-un anumit segment al depozitelor de date, cunoştinţele din acestea fiind supuse prelucrărilor prin instrumentele de acces (interfeţe de analiză) – produse software ce au în vedere analiza acestora. Frecvent, se recurge la tehnologii precum data mining sau OLAP.

 Astfel de instrumente de analiză au fost dezvoltate şi de compania Oracle (Oracle Data Miner, respectiv Oracle Business Intelligence Enterprise Edition).

 Interfeţele data mining asigură transformarea datelor în cunoştinţe prin tehnici ce ţin de analiza statistică superioară sau inteligenţă artificială.

 În contextul depozitelor de date, data mining este echivalent cu descoperirea cunoştinţelor (*eng.* KDD = Knowledge Discovery in Databases). Modelul matematic este foarte riguros şi se bazează pe analiza componentelor principale, prin care informaţii caracterizate prin mai multe atribute sunt sintetizate în cunoştinţe ce pot fi reprezentate prin mult mai puţine componente (denumite şi principale, sintetizate din atribute prin anumite formule, specifice fiecărui caz).

 Instrumentele OLAP au la bază reprezentarea multidimensională a informaţiilor şi permit accesul la date într-un mod interactiv, în funcţie de nivelul la care doresc să le vizualizeze (detaliat sau generic).

 Această arhitectură pe componente poate fi analizată şi în funcţie de nivelurile (*eng.* tier) la care sunt văzute datele sau din punct de vedere funcţional.

 Astfel, **arhitectura pe niveluri** identifică un nivel inferior, format din sursele de date şi depozitul de date propriu-zis precum şi din toate mecanismele pentru transformarea acestora (ETL), un nivel mediu format din instrumentele specializate pentru analiza datelor (data mining sau OLAP) şi un nivel superior care accesează datele care au fost generate.

 Din **punct de vedere funcţional** se disting:

* modulul operaţional – reprezentat de datele organizaţiei (interne sau preluate din surse externe);
* modulul central al depozitelor de date care poate fi implementat fie distribuit, sub formă de data mart-uri organizate în funcţie de diferite domenii de interes (spre exemplu, departamente) fie sub forma unei surse de date unice, la care au acces toţi utilizatorii;
* modulul strategic de afaceri, unde datele sunt pregatite pentru analiză, dezvoltându-se diferite instrumente pentru prezentarea şi interpretarea lor.

O viziune a depozitului de date din punct de vedere funcţional permite proiectarea acestuia raportat la cerinţele de afaceri.

1. **CLASIFICAREA DEPOZITELOR DE DATE**

Clasificarea depozitelor de date se poate face în funcţie de criteriile avute în vedere. Astfel, se pot distinge **aria de cuprindere**, **procesele decizionale** pentru care au fost proiectate precum şi **modelul de date implementat**.

Astfel, în funcţie de **aria de cuprindere** se pot distinge:

1. depozite de date de la nivelul organizaţiei (*eng.* enterprise data warehouses);
2. concentrările de date (*eng.* data marts);
3. depozitele de date virtuale (*eng.* virtual data warehouses).

Dacă depozitele de date de la nivelul organizaţiei conţin toate informaţiile referitoare la organizaţia respectivă (atât date extrase ca atare din diferite surse cât şi informaţii agregate), având dimensiuni foarte mari şi implicând costuri destul de importante atât pentru proiectare cât şi pentru implementare, concentrările de date sunt formate dintr-un subset de informaţii care privesc doar un anumit domeniu (subiect) al organizaţiei conţinând, de regulă, doar date obţinute prin agregare. Astfel, concentrările de date sunt considerate ca fiind un subansamblu al unui depozit de date mai ieftin, mai uşor de construit şi de intreţinut. Mai rar utilizate sunt depozitele de date virtuale, în fapt nişte vederi ale informaţiilor reţinute în bazele de date relaţionale, soluţia fiind adoptată doar atunci când numărul de înregistrări este mai mic, întrucât operaţiile (interogările propriu-zise, eventual agregările) sunt realizate de către sistemele de gestiune pentru baze de date, a căror scalabilitate este redusă.

Raportat la **procesele decizionale** pentru care sunt proiectate, se pot distinge:

1. **depozite de date**
	* *organizaţionale*
	* *departamentale;*
	* *orientate pe diferite procese de afaceri;*
2. **concentrări de date**
	* departamentale;
	* orientate pe diferite procese de afaceri.

 Depozitul de date de tip organizațional (*eng.* Galatic Data Warehouse) cuprinde toate informaţiile de la nivelul companiei, având ca obiectiv integrarea şi prelucrarea tuturor datelor cu relevanţă pentru aceasta.

 Depozitele de date departamentale sau orientate pe procese de afaceri reprezintă un subset al depozitului de date organizaţional, orientat pe diferite subiecte (departament sau proces de afaceri), fiind proiectate în scopul asigurării unei concurenţe a utilizatorilor şi scalabilităţi cât mai bune, exprimate sub forma timpului de răspuns.

 În aceeaşi categorie se înscriu şi concentrările de date specializate (având acelaşi tip de subiecte), diferenţa faţă de depozitele de date fiind faptul că nu conţin date detaliate, la nivelul acestora fiind folosite numai informaţii agregate pentru generarea unor rapoarte.

Din punctul de vedere al **modelului de date** adoptat, depozitele de date sunt împărţite în:

1. **depozite de date relaţionale;**
2. **depozite de date multidimensionale;**
3. **depozite de date hibride.**

 Un depozit de date relaţional se foloseşte în cazul când datele provin dintr-o bază de date care îl şi gestionează sau dacă se folosesc depozite de date virtuale.

 Un depozit de date multidimensional defineşte dimensiuni şi ierarhii ale datelor (conform modelului multidimensional), aceasta reprezentând varianta optimă de folosire a acestei tehnologii, putându-se procesa un volum mare de date prin intermediul serverului multidimensional care dispune de diferite optimizări pentru procesul de analiză.

 Un depozit de date hibrid foloseşte modelul multidimensional pentru stocarea datelor istorice şi modelul relaţional pentru datele curente, oferind o perspectivă completă pentru toate informaţiile existente la nivelul organizaţiei.

1. **MODELUL MULTIDIMENSIONAL ȘI ADOPTAREA LUI ÎN PROIECTAREA UNUI DEPOZIT DE DATE**

 Întrucât în depozitele de date operaţiile mai frecvente sunt cele de analiză a informaţiilor reţinute (aceasta reprezintând o caracteristică ce distinge depozitele de date faţă de bazele de date relaţionale unde cele mai frecvente operaţii erau cele de actualizare), precum şi obţinerea unor perspective diferite asupra lor în funcţie de nivelul de agregare dorit, a fost necesară dezvoltarea unui model care să permită un mod de organizare a datelor care să permită astfel de prelucrări.

 Iniţial s-a pornit de la diferite extensii ale modelului relaţional (Gray, Li-Wang, Gyssens-Lakshmanan) ajungându-se la modelul lui Ralph Limball care defineşte cubul n-dimensional prin intermediul unei scheme de tip stea. De aici, s-a ajung la modele de tip fulg de nea sau galaxie, din care s-au inspirat apoi şi modelele propuse de Gupta şi Sarawagi sau Cabibbo şi Torlone.

 Modelul multidimensional implică definirea obiectelor referitoare la tabele de fapte cu atribute de tip măsuri (metrici), a obiectelor de tip dimensiuni pentru care se definesc ierarhii precum şi diferite alte atribute (identificare, descriere). Modelul multidimensional al datelor poate fi descris la nivel conceptual (al înţelegerii utilizatorilor, fără legătură cu modul de implementare propriu-zis), la nivel logic (conceptele pot fi înţelese în continuare de utilizatori, însă sunt adaptate sistemelor de gestiune pentru care acestea sunt implementate) şi la nivel fizic (care descriu modul în care sunt stocate datele fizic).

 **Dimensiunile** sunt structuri formate din atribute grupate pe mai multe niveluri ale unei ierarhii, astfel încât informaţiile pot fi vizualizate mai detaliat sau mai sintetic, în funcţie de cerinţele operaţionale. O definiţie formulată de consiliul OLAP consideră dimensiunea ca fiind *un atribut structural al unui cub care constă dintr-o listă de membri, percepuţi ca fiind de acelaşi tip (de exemplu toate lunile, trimestrele, anii formează dimensiunea timp); dimensiunile reprezintă un mod foarte concis, intuitiv de organizare şi selectare a datelor pentru explorare şi analiză*.

 Datorită faptului că exprimă informaţiile într-un mod concis, dimensiunile sunt considerate drept tabele secundare.

 **Ierarhiile** reprezintă structurile în care se organizează atributele unei dimensiuni, pe baza relaţiei părinte-copil, semnificaţia acesteia fiind că părintele este obţinut prin agregarea mai multor copii. Acestea sunt utilizate pentru navigarea în cadrul dimensiunilor în funcţie de nivelul de detaliere / sinteză dorit.

 **Nivelurile** definesc poziţiile în cadrul ierarhiilor, relaţiile dintre acestea raportându-se şi la noţiunea de generaţie, conform definiţiei consiliului OLAP: *doi membri ai unei ierarhii sunt de aceeaşi generaţie dacă au acelaşi număr de strămoşi. Termenii de generaţie şi nivel sunt necesari pentru a descrie subgrupuri de membri întrucât, de exemplu, deşi doi fraţi membri au acelaşi părinte şi sunt de aceeaşi generaţie, ei ar putea să nu fie la acelaşi nivel, dacă unul dintre fraţi are copil şi celălalt nu*.

 **Atributele** reprezintă caracteristicile unei dimensiuni care se pot găsi doar la un singur nivel al unei ierarhii. Se disting atributele de identificare (pentru dimensiune în sine sau pentru un nivel al ierarhiei), respectiv atributele de descriere care realizează o clasificare a datelor în cadrul ierarhiei.

 **Faptele** sunt tabele caracterizate prin legături către dimensiuni, având şi atribute proprii care sunt denumite măsuri (metrici), conţinând de regulă date care pot fi sintetizate (agregate) pentru fiecare nivel din ierarhiile definite pentru dimensiunile referite permiţând o analiză cu un grad de detaliere diferit.

 **Măsurile** (metricile) au relevanţă doar în contextul unor anumite dimensiuni şi pot fi clasificate în funcţie de modalitatea de calcul, tipul de funcţii agregate utilizate, modul de calcul raportat la dimensiuni. Astfel, în funcţie de modalitatea de calcul, se definesc măsuri de bază (existente în sursele de date fiind preluate ca atare) şi măsuri derivate, obţinute prin combinarea măsurilor de bază, fiind precizate prin formula utilizată la obţinerea lor. Raportat la tipul de funcţii agregate se disting măsurile distributive, măsurile algebrice şi măsurile holistice. În privinţa modului de calcul raportat la dimensiuni există indicatori aditivi, indicatori semiaditivi şi indicatori neaditivi.

 Funcţiile agregate distributive (corespunzătoare măsurilor distributive) sunt caracterizate prin faptul că împărţind setul de date în subgrupe şi aplicând respectiva metodă pentru fiecare dintre acestea, iar apoi funcţia pe rezultatele agregate se obţine acelaşi rezultat cu aplicarea sa pe întreg setul de date neîmpărţit. Un astfel de exemplu e funcţia COUNT(). Funcţiile algebrice sunt obţinute prin aplicarea unei funcţii agregate distributivă (AVG() = SUM() / COUNT(); la fel, alte exemple din această categorie sunt MIN(), MAX(), STDEV()). Funcţiile holistice nu sunt limitate constant pe spaţiul de stocare cerut de deschiderea subagregării (exemple în acest sens sunt MEDIAN(), MODE(), RANK()).

Indicatorii aditivi sunt cei care pot fi însumaţi pe toate dimensiunile, indicatorii semiaditivi cei care pot fi însumaţi doar pe anumite dimensiuni, iar indicatorii neaditivi cei care nu pot fi însumaţi pe nici o dimensiune.

 **Metadatele** sunt considerate cea mai importantă componentă a unui depozit de date întrucât ele nu doar descriu structura de date respectivă ci reţin şi provenienţa datelor precum şi algoritmii utilizaţi pentru agregarea în funcţie de nivelurile definite la nivelul dimensiunii.

 Astfel de metadate se numesc operaţionale şi includ informaţii cu privire la evoluţia în timp si circulaţia datelor, reţinând chiar şi elemente legate de jurnalizare.

 Aspecte legate de algoritmii folosiţi sunt şi granularitatea, partiţiile, ariile de subiecte, rapoartele şi diferite filtre.

 Tot aici sunt reţinute şi transformările pe care le suportă informaţiile de la preluarea lor din sursele de date (relaţionale) până la încărcarea lor în depozitul de date. De asemenea, metadatele reţin şi anumite moduri de vizualizare, în condiţiile în care unele dintre acestea sunt mai utilizate decât altele.

 Acestea asigură şi calitatea datelor definind valorile valide pentru fiecare atribut din depozit, astfel încât înainte de încărcarea lor pot fi aplicate anumite reguli de corecţie. Prin metadate se asigură şi gestiunea versiunilor, întrucât regulile care se aplică asupra datelor pot fi schimbate odată cu trecerea timpului, având în vedere dependenţa lor şi de acest atribut.

 Modelele de reprezentare ale depozitelor de date (sub formă de schemă) fie utilizează extensii ale modelului relaţional, fie structurează obiectele acestuia sub forma unei structuri de tip cub multidimensional.

 Ca extensii ale modelului relaţional se folosesc schemele pentru depozite de date propuse de Ralph Kimball:

1. **stea;**
2. **fulg de nea;**
3. **galaxie;**
4. **constelaţie.**

 În **schema de tip stea**, în centru se află una sau mai multe tabele de fapte care sunt în relaţie cu dimensiunile, suportând două tipuri de interogări: consultare – care se realizează doar pe tabela de fapte – şi joncţiune în care restricţiile cu privire la atribute se fac între tabela de fapte şi dimensiuni. Dimensiunile sunt denormalizate, întrucât nu există alte legături între tabele cu excepţia celor dintre tabela de fapte şi dimensiuni.

 Astfel, schema depozitului de date este formată dintr-un set de atribute {A1, A2, …, An}, o tabelă de fapte f şi m dimensiuni d1, d2, …, dm. Atât tabela de fapte f cât şi dimensiunile d1, d2, …, dm sunt formate dintr-un subset distinct de atribute {Ak | k = 1..n}.

Atributele {Ak | k = 1..n} care nu fac parte din dimensiunile d1, d2, …, dmse numesc măsuri (metrici) întrucât caracterizează tabela de fapte.

 Deşi acest tip de schemă are avantajul unei regăsiri foarte rapide a informaţiilor, totuşi implică o inflexibilitate destul de mare.

 **Sintetizând, caracteristicile schemei de tip stea sunt:**

* cheile primare ale dimensiunilor fac parte din atributele cheii primare (compuse) ale tabelei de fapte şi reprezintă elementele prin intermediul cărora se realizează legăturile între acestea;
* atributele măsură ale tabelei de fapte pot fi agregate.

 O variantă a schemei de tip stea este **schema de tip fulg de nea** unde unele dintre dimensiuni sunt normalizate, datele fiind distribuite în alte tabele. Astfel, se menţine redundanţa controlată din cadrul bazelor de date relaţionale, întreţinerea depozitului de date fiind mai facilă (întrucât se evită anomaliile care pot apărea în cazul unor operaţii de actualizare), economisindu-se totodată şi spaţiul de stocare. Totusi, operaţiile de actualizare în cazul unui depozit de date nu sunt foarte frecvente, iar economia spaţiului de stocare este neglijabilă în comparaţie cu volumul de date reţinut. Dezavantajul acestei structuri e legată de timpul de regăsire care creşte în cazul unei interogări întrucât presupune joncţiuni între mai multe tabele.

 O **schemă de tip galaxie** reprezintă o asociere între mai multe scheme de tip stea, conţinând tabele suplimentare pentru datele agregate. Legătura se face prin intermediul unei dimensiuni. În cadrul galaxiei există o stea principală (centrală) care conţine măsurile de bază în timp ce în celelalte stele sunt distribuite măsurile derivate. Această schema de proiectare a depozitelor de date prezintă atât avantaje cât şi dezavantaje: pe de o parte măsurile de bază (atomice) sunt menţinute independent de măsurile derivate, iar pe de altă parte determinarea măsurilor derivate implică asigurarea unor mecanisme pentru realizarea coerenţei cu măsurile de bază.

 În **schema de tip constelaţie** există mai multe legături între dimensiunile depozitului de date, acestea fiind complet denormalizate. Legăturile dintre tabelele de fapte sunt realizate tot prin intermediul dimensiunilor, neexistând o legătură directă între ele. O astfel de abordare este urmată dacă depozitul de date este construit dintr-o bază de date relaţională, aceasta contribuind la întreţinerea acestei structuri.

 O altă metodă de reprezentare a datelor care respectă modelul multidimensional este **cubul de date** (n-dimensional), definit într-un sistem cartezian ce are drept coordonate atributele depozitului de date. Desigur, cubul de date reprezintă un spaţiu de date logic, nicidecum fizic.

De regulă întrucât de obicei numărul de dimensiuni depăşeşte cifra 3, se foloseşte noţiunea de hipercub, definită ca un grup de celule de date aranjate după dimensiunile datelor; o matrice tridimensională poate fi vizualizată ca un cub unde fiecare dimensiune formează o faţă a cubului. De obicei, coordonatele cele mai frecvent întâlnite în reprezentarea depozitelor de date sunt timpul, produsele, regiunile geografice, canalele de distribuţie.

 Prin cuburi se pot realiza diferite secţiuni care poartă denumirea de tablouri.

Un cub cu 3 sau 4 dimensiuni poate fi reprezentat facil în plan, însă pentru mai multe dimensiuni trebuie apelat la anumite modele, ca de exemplu Agrawal-Gupta (care definesc un hypercub ale cărui elemente pot fi 0, 1 respectiv un n-tuplu, propunând şi un set de operatori care acţionează asupra acestuia) sau Golfarelli-Maio-Rizzi care folosesc un graf bazat pe o schemă de fapte.

1. **EXEMPLU**

Se doreşte proiectarea unui depozit de date pentru gestiunea tranzacţiilor dintr-o librărie, prin preluarea informaţiilor din baza de date relaţională.



↓



 Se observă că s-a obţinut un depozit de date de tip constelaţie în care tabelele de fapte sunt comenzi aprovizionare şi facturi, legăturile intre acestea fiind realizate prin intermediul dimensiunilor timp şi carte.

 Fiecare dintre ele mai este legată de o dimensiune (comenzi aprovizionare de dimensiunea furnizor, iar facturi de dimensiunea utilizatori). Întrucât nu s-a procedat la denormalizarea bazei de date , dimensiunea carte prezintă în continuare legături către alte patru dimensiuni. Metricile din cadrul tabelelor de fapte sunt cantitate (indicator aditiv) şi stare (indicator neaditiv).

 Oricare tabelă de fapte împreună cu dimensiunile sale poate fi considerată stea centrală.

 O excepţie de la regula denormalizării s-a făcut totuşi în cazul tabelelor de fapte (supply\_order conţine şi supply\_order\_detail, iar invoice include şi invoice\_detail).

 De asemenea, pentru dimensiuni au fost definite ierarhii, astfel:

pentru dimensiunea D\_TIME 🡪 nivelurile day, month, quarter şi year;

pentru dimensiunea D\_BOOK 🡪nivelurile book (cu atributele din tabela corespunzătoare), series şi genre,

pentru dimensiunea D\_SUPPLIER 🡪nivelurile supplier, town, region, country iar

pentru dimensiunea D\_USER 🡪nivelurile user, type şi role.

**În construirea ierarhiilor, de remarcat sunt relaţiile de incluziune:**

day < month < quarter < year

book < series < genre

supplier < town < region < country

user < type { (regular administrator, super admininstrator), (client, supplier) } < role { administrator, third party }

**În acest mod, se pot obţine rapoarte cum ar fi:**

* volumul achiziţiilor pentru o anumită lună dintr-o anumită colecţie realizate de la un anumit furnizor;
* volumul vânzărilor dintr-un anumit trimestru pentru cărţi aparţinând unui anumit domeniu către un anumit tip de utilizatori.

În funcţie de nivelul la care se situează în ierarhia specifică unei anumite dimensiuni, rapoartele pot fi mai detaliate sau mai sintetice.

1. **OPERAȚII ÎN CADRUL DEPOZITELOR DE DATE**

Indiferent de modelul adoptat în proiectarea unui depozit de date (relaţional sau multidimensional), operaţiile care pot fi realizate asupra sa sunt aceleaşi, de vreme ce reprezentarea datelor este similară. Din acest motiv, şi transformarea datelor se face foarte uşor.

 În literatura de specialitate se vorbeşte – în funcţie de modelul adoptat în proiectarea depozitului de date – de analize:

* ROLAP (Relational OLAP)
* MOLAP (Multidimensional OLAP)
* HOLAP (Hybrid OLAP)

unde OLAP = OnLine Analytical Processing.

 Operaţiile pot fi folosite de utilizator pentru a avea diferite perspective asupra datelor, fie având un grad de detaliere mai mare sau mai mic (prin agregare), fie pentru a observa anumite dependenţe prin eliminarea unor coordonate (atribute), respectiv extrăgând un anumit subset de date.

 Operaţii de **navigare pe diferite niveluri ale ierarhiei unei anumite dimensiuni** sunt de:

* ***drill down (detaliere pe nivelurile de la baza ierarhiei);***
* ***roll up (agregare pe nivelurile superioare).***

 Deşi cele mai multe dintre instrumentele de analiză asociate depozitelor de date pot realiza astfel de operaţii în mod dinamic, se obişnuieşte ca unele valori globale să fie precalculate această tehnică purtând denumirea de consolidare, însumare sau agregare după aspectele care sunt luate în considerare (conceptual, procedural, structural). Aceste valori sunt măsurile tabelelor de fapte, iar consolidarea se face după dimensiunile corespunzătoare ei, nivel cu nivel.

 Operaţiile propriu-zise urmărite în cadrul consolidării sunt de cele mai multe ori totaluri, însă pot fi folosiţi şi alţi operatori, în cele mai multe cazuri statistici. Nivelul pentru care se realizează consolidarea poartă numele de granularitate. Rezultatele astfel obţinute sunt reţinute în depozitul de date, crescându-i performanţa, fără ca dimensiunea acestuia să crească semnificativ, de vreme ce numărul de valori scade (exponenţial) pe măsură ce se avansează în ierarhie.

 Spre exemplu, volumul vânzărilor pentru o carte poate fi calculat doar pentru o localitate, pentru o regiune sau chiar pentru întreaga ţară.

 Cele mai frecvente tipuri de operaţii folosite în cazul depozitelor de date sunt **rotaţiile**, oferind utilizatorului mai multe posibilităţi de vizualizare a informaţiilor reţinute. Pentru un cub n-dimensional există Pn = 1 x 2 x … x n posibilităţi de rotaţie, aducând în prim-plan o anumită faţetă (bidimensională) motiv pentru care o astfel de operaţie mai poartă denumirea de data **slicing**.

 Spre deosebire de cazul bazelor de date relaţionale în care o astfel de operaţie implică modificarea interogării respective, pentru depozitele de date rotaţiile echivalează doar cu schimbarea modului de reprezentare.

 Astfel, poate fi vizualizat volumul de vânzări pentru o carte în funcţie de regiuni, respectiv distribuţia volumului de vânzări a diferitelor volume dintr-o anumită regiune.

 De asemenea, pot fi realizate **secţiuni** prin cubul de date acestea presupunând limitarea unor anumite atribute la anumite valori, astfel obţinându-se un cub de date redus, motiv pentru care o astfel de operaţie mai poartă denumirea de **dicing**. Din punct de vedere geometric, nu este necesar ca rezultatul obţinut să aibă structura unui cub (chiar de dimensiuni mai mici decât cubul de date iniţial), acesta putând avea orice forme de la dreptunghi (se consideră o singură valoare pentru un atribut) la paralelipiped dreptunghic (în cazul cuburilor de date cu 3 dimensiuni).

 Operaţia de secţionare a cubului de date este echivalentă cu anumite vizualizări (*eng.* views) care se realizau pentru bazele de date relaţionale.

 De exemplu, poate fi vizualizat volumul vânzărilor doar pentru anumite cărţi, doar pentru un interval de timp (o lună, un trimestru) şi doar pentru anumite regiuni.

 Pentru depozitele de date care sunt reţinute sub forma unor baze de date relaţionale, există unele extensii ale unor operatori relaţionali ce permit realizarea unor operaţii de consolidare. Aceștia au fost propuşi în anul 1998 de către cercetătorul american Paul Gray și sunt implementate de către limbajul SQL.

 Operatorul CUBE permite generarea tuturor combinaţiilor posibile pentru operaţiile de consolidare din cadrul unui cub de date. El suportă clauza GROUPING prin care se face distincţia între valorile NULL ale tabelei de fapte şi informaţiile obţinute prin agregare.

 Operatorul ROLLUP permite realizarea unor operaţii de consolidare pentru anumite atribute (specificate) prin intermediul operatorilor definiţi de către sistemul de gestiune al bazei de date, dintre care cel mai frecvent utilizaţi sunt SUM(), COUNT(), AVG(), STDEV(). Întrucât acest operator implică folosirea unor funcţii de grup, el va fi utilizat în conjuncţie cu clauza GROUP BY.

Având în vedere definiţia acestor operatori, CUBE(ROLLUP)=CUBE.

1. **PRINCIPII DE PROIECTARE A UNUI DEPOZIT DE DATE**

**Un depozit de date poate fi proiectat pornind de la unele elemente care stau la baza implementării sale ulterioare, existând şi în acest caz mai multe abordări:**

* abordarea top-down ce porneşte de la informaţiile relevante pentru depozitul de date, utilizate de către managerii executivi pentru stabilirea unor atitudini la nivel tactic sau strategic;
* abordarea care porneşte de la datele sursă unde sunt reprezentate informaţiile stocate în bazele de date operaţionale, ele putând fi diseminate pe mai multe niveluri (în funcţie de gradul de detaliere) conducând la tabele în care datele sunt integrate;
* abordarea în care se stabileşte schema depozitului de date prin specificarea tabelelor de fapte şi a tabelelor de dimensiuni precum şi a mecanismelor de realizare a consolidării sau a surselor de date care vor fi folosite;
* abordarea bazată pe interogări unde se stabileşte funcţionalitatea pe care trebuie să o îndeplinească depozitul de date.

**Ca mod de lucru se pot folosi:**

* **metoda în cascadă, soluţie sistemică care porneşte cu proiectarea completă, dar care poate fi aplicată doar dacă toate problemele care trebuie rezolvate sunt foarte bine înţelese;**

 Această soluţie este foarte scumpă şi implică dezvoltarea depozitului de date într-o perioadă de timp foarte îndelungată. Are avantajul că minimizează problemele legate de integrare, însa se caracterizează şi prin dezavantajul că este lipsită de flexibilitate, constrângere determinată de faptul că trebuie modelate datele pentru întreaga organizaţie.

* **metoda în spirală** **care porneşte de la prototipuri de dimensiuni reduse (concentrări de date);**

 O astfel de abordare este mai ieftină şi permite organizaţiei să evalueze beneficiile investiţiei înainte de a lua decizia de a continua dezvoltarea proiectului. Totodată, se bucură de flexibilitate întrucât concentrările de date conţin date independente. Problemele apar însă atunci când datele trebuie integrate.

* **o metodă mixtă care îmbină caracterul strategic al abordării în cascadă şi rapiditatea ce caracterizază soluţia în spirală.**

**Etapele dezvoltării unui depozit de date pentru o instituţie sunt:**

1. **determinarea contextului organizational;**
2. **realizarea unui studiu preliminar privind cerinţele pe care depozitul de date trebuie să le îndeplinească;**
3. **realizarea auditului preliminar referitor la sistemele sursă;**
4. **identificarea surselor de date externe;**
5. **transformarea câmpurilor sursă în câmpuri destinaţie;**
6. **definirea versiunilor depozitului de date;**
7. **definirea arhitecturii preliminare a depozitului de date;**
8. **construirea metadatelor;**
9. **încărcarea datelor istorice;**
10. **evaluarea mediilor de dezvoltare;**
11. **crearea prototipului pentru versiunea curentă.**

 Implementarea propriu-zisă a depozitului de date implică dezvoltarea unor subsisteme care vor ajuta la exploatarea sa ulterioară şi anume:

* **subsistemul pentru extragerea şi transformarea datelor;**
* **subsistemul pentru calitatea datelor;**
* **subsistemul pentru încărcarea depozitului de date.**

 Ulterior procesului de implementare şi de încărcare cu date va urma etapa de testare, după care depozitul de date va putea fi adoptat la nivelul organizaţiei.

 Procesul de întreţinere vizează încărcarea periodică a depozitului de date, consultarea de statistici cu privire la utilizarea lui, menţinerea calităţii datelor, evaluarea dimensiunii sale şi refacerea datelor în caz de eroare.

**Anexa 1**

**ETL vs ELT: Must Know Differences**

**What is ETL?**

ETL is an abbreviation of Extract, Transform and Load. In this process, an ETL tool extracts the data from different RDBMS source systems then transforms the data like applying calculations, concatenations, etc. and then load the data into the Data Warehouse system.

In ETL data is flows from the source to the target. In ETL process transformation engine takes care of any data changes.



**What is ELT?**

ELT is a different method of looking at the tool approach to data movement. Instead of transforming the data before it's written, ELT lets the target system to do the transformation. The data first copied to the target and then transformed in place.

ELT usually used with no-Sql databases like Hadoop cluster, data appliance or cloud installation.



**Difference between ETL vs. ELT**

ETL and ELT process are different in following parameters:

| **Parameters** | **ETL** | **ELT** |
| --- | --- | --- |
| **Process** | Data is transformed at staging server and then transferred to Datawarehouse DB. | Data remains in the DB of the Datawarehouse. |
| **Code Usage** | Used for* Compute-intensive Transformations
* Small amount of data
 | Used for High amounts of data |
| **Transformation** | Transformations are done in ETL server/staging area. | Transformations are performed in the target system |
| **Time-Load** | Data first loaded into staging and later loaded into target system. Time intensive. | Data loaded into target system only once. Faster. |
| **Time-Transformation** | ETL process needs to wait for transformation to complete. As data size grows, transformation time increases. | In ELT process, speed is never dependant on the size of the data. |
| **Time- Maintenance** | It needs highs maintenance as you need to select data to load and transform. | Low maintenance as data is always available. |
| **Implementation Complexity** | At an early stage, easier to implement. | To implement ELT process organization should have deep knowledge of tools and expert skills. |
| **Support for Data warehouse** | ETL model used for on-premises, relational and structured data. | Used in scalable cloud infrastructure which supports structured, unstructured data sources. |
| **Data Lake Support** | Does not support. | Allows use of Data lake with unstructured data. |
| **Complexity** | The ETL process loads only the important data, as identified at design time. | This process involves development from the output-backward and loading only relevant data. |
| **Cost** | High costs for small and medium businesses. | Low entry costs using online Software as a Service Platforms. |
| **Lookups** | In the ETL process, both facts and dimensions need to be available in staging area. | All data will be available because Extract and load occur in one single action. |
| **Aggregations** | Complexity increase with the additional amount of data in the dataset. | Power of the target platform can process significant amount of data quickly. |
| **Calculations** | Overwrites existing column or Need to append the dataset and push to the target platform. | Easily add the calculated column to the existing table. |
| **Maturity** | The process is used for over two decades. It is well documented and best practices easily available. | Relatively new concept and complex to implement. |
| **Hardware** | Most tools have unique hardware requirements that are expensive. | Being Saas hardware cost is not an issue. |
| **Support for Unstructured Data** | Mostly supports relational data | Support for unstructured data readily available. |

**Summary:**

* ETL stands for Extract, Transform and Load while ELT stands for Extract, Load, Transform
* In ETL process data flows from the source to staging to the target.
* ELT lets the target system to do the transformation. No staging system involved.
* ELT address many a challenge of ELT but is expensive and requires niche skills to implement and maintain.

