**PLANUL LECTIEI 7-8 (2 ORE), 16\_03\_23**

### LECTIA 7 PROIECTAREA DWH. MAPAREA DATELOR

### PROIECTAREA DWH

***Depozitarea datelor nu se referă la instrumente.*** Mai degrabă, este vorba despre ***crearea unei strategii de planificare, proiectare și construire a unui magazin de date capabil să răspundă la întrebările unei afaceri***.

În acest context o strategie ***bună este un proces*** care nu este niciodată terminat, mai degrabă este un ***ciclu continuu de perfecționări***.

***Instrumentele care există* funcționează bine pentru a ajuta la realizarea proiectului informational pentru business** prin încărcarea tabelelor țintă în depozitul de date, ***nu pot crea un plan pentru depozit****.*

**PRIN URMARE, ESTE IMPORTANT SĂ ÎNȚELEGEM PAȘII CARE TREBUIE REALIZATI MAI ÎNAINTE DE UTILIZAREA INSTRUMENTULUI.**

#### IATĂ ACESTI PASI:

1.  **Clarificăm cerințele sarcinei Dvs**. (Instrumentul principal: document de definire a cerințelor, a rapoartelor pentru procesul decizional), prin examinarea întrebărilor și răspunsurilor adresate utilizatorilor de afaceri în timpul procesului de cercetare. Adițional este important să examinăm orice rapoarte sau analize existente pe subiectul abordat pentru crearea unui DD.

2.  **Clarificăm datele sursă pentru sarcina formulată**. Clarificăm relațiile care există între datele sursă. Încercăm să înțelegem datele sursă și conținutul acestora, efectuînd o analiză a lacunelor cu privire la ceea *ce urmează să facem și nu avem în ceea ce privește datele existente*. Identificăm sursele de găsire a acestor date sau de generare a lor.

3.  **Proiectăm modelul de date pentru depozitul de date*.*** Modelăm DD folosind una din „schemele” existente.

4.  **Definim regulile de mapare,** adică regulile ce definesc de obicei *ce trebuie de făcut în condițiile specifice ale proiectului concret formulat* de sarcină. Creăm reguli noi pentru ceea ce urmează să facem, pornind de la tabelele sursă existente la tabelele țintă. Aceste reguli  nu prezic viitorul, dar urmează să ajute să luăm măsuri semnificative în present pentru a realiza proiectul formulat cu scopul lui țintă pentru activitățile de business.

5.  **Convertim regulile Proiectului sarcinii sursă studiat, în scopul final,**   prin prezentarea regulilor pentru a atinge Scopul descriind metadatele lui (instrument principal al utlizării datelor și rezolvării practice a Proiectului formulat folosind SSAS (sau SAS) împreună cu proiectarea multidimensională și tehnologia cuburilor de date).

***Depozitarea datelor nu se referă la instrumente.*** Mai degrabă, este vorba despre ***crearea unei strategii de planificare, proiectare și construire a unui magazin de date capabil să răspundă la întrebările unei afaceri***.

În acest context o strategie ***bună este un proces*** care nu este niciodată terminat, mai degrabă este un ***ciclu continuu de perfecționări***.

***Instrumentele care există* funcționează bine pentru a ajuta la realizarea proiectului informational pentru business** prin încărcarea tabelelor țintă în depozitul de date, ***nu pot crea un plan pentru depozit****.*

### PAȘI PENTRU PLANIFICAREA DD

***Înțelegerea cerințelor***

***De exemplu,***

***Sistemul ar trebui să imprime rezultatele sau să le prezinte la ecran la cererea utilizatorului.***

**Culegerea datelor**

       Acest pas se începe cu culegerea datelor disponibile suficiente pentru a crea un model bine structurat și a obține răspunsurile dorite. Putem obține aceste date si informație despre ele din:

1. ***cerințele din interviuri cu userul,***
2. ***examinarea rapoartelor existente și***
3. ***evaluarea sistemelor informaționale operaționale existente.***

#### Obiectivul cercetării:

* ***Definim scopul și decidem asupra subiectului (subiectelor) pentru a proiecta un DD***
* ***Identificăm întrebările de interes ce urmează a fi clarificate***

***De exemplu,*** pentru compania ***Northwinds Trading* /vezi Anexa 1/**cu BDO prezentată mai jos, urmează, în timpul utilizării  **modelului unui process de business,** să răspundem la un șir de ***întrebări/interogări*** cum ar fi:

1. ***Cine a cumpărat produsele (clienții și structura acestora)?***
2. ***Cine a vândut produsul (organizația de vânzări etc.)?***
3. ***Ce s-a vândut (structura produsului)?***
4. ***Când a fost vândut produsul (structura timpului)?***
5. ***Care sunt caracteristicile vânzării (reducere etc.)? etc…etc***

***Masurile***, care ar exprima vânzările, nu există izolat, dar mai degrabă în contextul dimensiunilor stratificate, cum ar fi: **Produse, Angajați, Clienți și Timpul.** Împreună aceste dimensiuni definesc ce tip de date sunt disponibile.

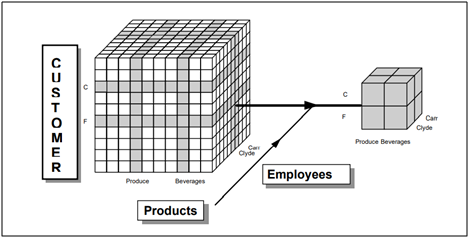
       Prin urmare, ***venitul din vânzări*** este măsură și este calificată/determinată de astfel de dimensiuni cum ar fi ***Produse, Angajați, Clienți și Timpul.***

#### Înțelegerea datelor sursă Modelul logic multidimensional

***Să ne reamintim că în timpul analizei modelului logic, o importanță mare o au cerințele funcționale. Or, este important să găsim răspunsul la întrebarea: Ce dorim cu adevărat de la informațiile pe care le avem despre activitățile stipulate în Proiect cu referință la procesul studiat și cum este ea structurată?***

***Ori, pentru a raspunde la această intrebare, a atinge scopul formulat, este necesar să atingem următoarele obiective:***

* 1. ***Definim cerințele funcționale***
  2. ***Concretizăm subiectele/entitățile cu care vom lucra***
  3. ***Determinăm ce însemnătate/ințeles are dimensiunea Timp pentru Proiectul nostru***
  4. ***Identificăm granularitatea/detaliile (cât de adânc putem coborî) în cercetările noastre pentru subiectele concretizate mai sus.***
  5. ***Creăm fapte și dimensiuni „reale” pornind de la subiectele/entitățile pe care le-m concretizat mai sus***

******

**PAȘII CREĂRII UNUI DWH EFICIENT**

**Pasul 1: Determinarea cerințelor de business/afaceri**

**Pasul 2: Colectarea și analizarea datelor/informațiilor**

**Pasul 3: Identificarea proceselor de bază ale afacerii**

**Pasul 4: Construirea modelului conceptual si selectarea arhitecturii design-ului DWH. Schemei ER a DWH**

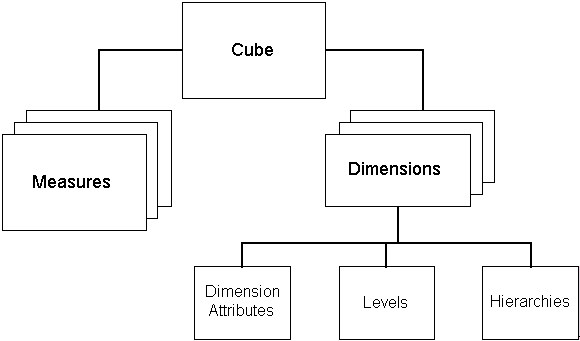
**Pasul 5: Localizarea surselor de date și planificarea transformărilor datelor/ Alegerea soluției ETL**

**Pasul 6: Determinarea politicilor OLAP cuburilor**

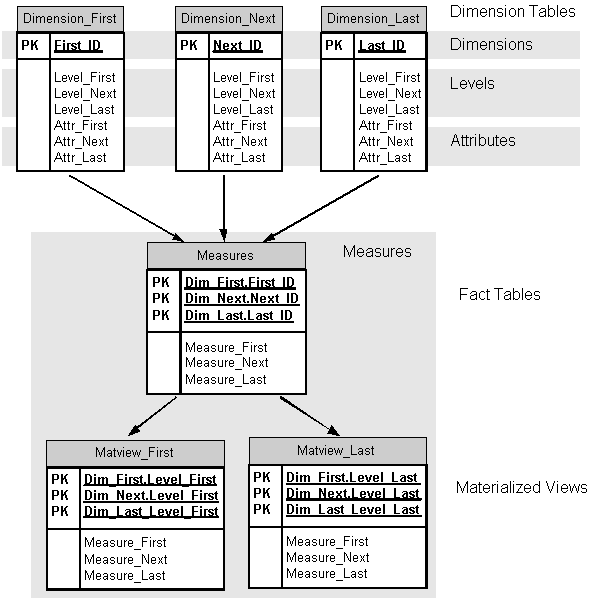
**Pasul 7: Setarea duratei cercetării**

**Pasul 8: Implementarea planului de cercetare.**

**Modelul de date multidimensional este compus din: *cuburi logice, măsuri, dimensiuni, ierarhii, niveluri și atribute*. Simplitatea modelului este inerentă deoarece definește obiecte care reprezintă entități de afaceri din lumea reală. Analiștii știu ce măsuri de afaceri sunt interesați să examineze, ce dimensiuni și atribute fac datele semnificative și cum dimensiunile afacerii lor sunt organizate în niveluri și ierarhii. Schema logică generală a modelului de date multidimensional**

****

**PRIVITĂ PRIN PRISMA MODELULUI DE DATE RELATIONAL**



#### MODELUL FIZIC

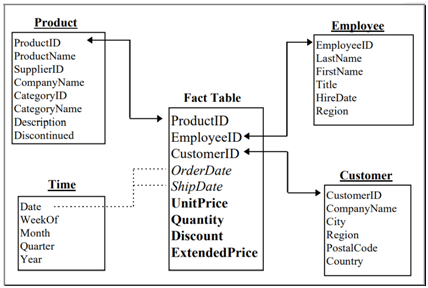
***Modelul fizic in proiectarea DD conține detaliile datelor disponibile și cum urmează ele a fi stocate.***

***Proiectarea fizică a datelor se prezintă printr-o schemă. O schemă pentru DD poate fi reprezentată de unul sau mai multe moduri constructive de proiectare, cum ar fi:***

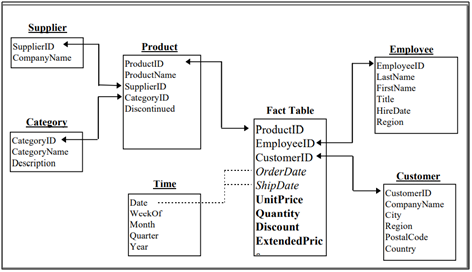
1. ***Modelul entitate-relație***
2. ***Schema de tip stea a faptelor***
3. ***Schemă de tip fulg de zăpadă a faptelor***
4. ***Schemă de tip constelație a faptelor***
5. ***Magazin multidimensional persistent /ce rămîne neschimbat mult timp/***
6. ***Tabelele rezumative***

**DE EXEMPLU**

***Model de schemă de tip Stea pentru sistemul Northwinds Trading Company.***

******

***Schema de tip Fulg de Zăpadă***

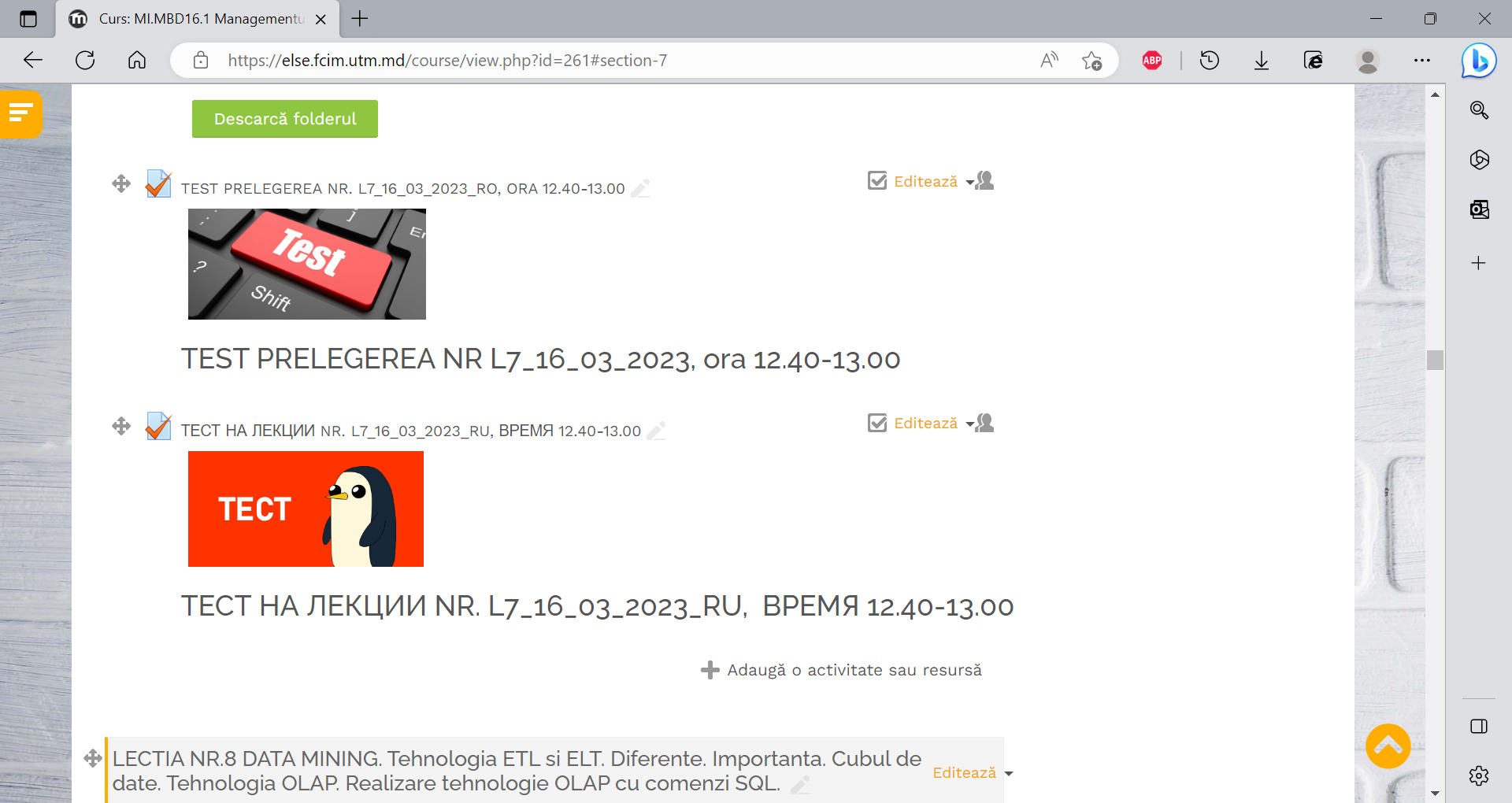
******

* LITERATURA DE LUCRUDosar

|  |  |
| --- | --- |
|  | MATERIALE PENTRU MAPAREA DATELOR SI PROCESUL ETL |
|  |  |

|  |  |  |
| --- | --- | --- |
|  |  | [1__PROIECTARE.docx1\_\_PROIECTARE.docx](https://else.fcim.utm.md/pluginfile.php/112237/mod_folder/content/0/1__PROIECTARE.docx?forcedownload=1) |
|  |  | [3__OLAP Cubes in Datawarehousing_RECOMAND.pdf3\_\_OLAP Cubes in Datawarehousing\_RECOMAND.pdf](https://else.fcim.utm.md/pluginfile.php/112237/mod_folder/content/0/3__OLAP%20Cubes%20in%20Datawarehousing_RECOMAND.pdf?forcedownload=1) |

|  |  |  |
| --- | --- | --- |
|  |  | [CE ESTE ETL_LAST.pptxCE ESTE ETL\_LAST.pptx](https://else.fcim.utm.md/pluginfile.php/112237/mod_folder/content/0/CE%20ESTE%20ETL_LAST.pptx?forcedownload=1) |
|  |  | [CE ESTE MAPAREA DATELOR.docxCE ESTE MAPAREA DATELOR.docx](https://else.fcim.utm.md/pluginfile.php/112237/mod_folder/content/0/CE%20ESTE%20MAPAREA%20DATELOR.docx?forcedownload=1) |



Top of Form

**Bottom of Form**

1. **PROCESUL ETL.**

* **CE ESTE ETL\_LAST.pptx**

1. **DATA MINING.**

* **ARHITECTURA DWH ETL, ELT.pptx**

1. **VEDETI FISIERUL 1\_\_PROIECTARE.docx DE DATA TRECUTA**

**DE LA PAGINA 11**

1. **Fisierul 2\_\_DM\_\_AND STRUCUTRING DAT.pptx DE LA**

**PAGINA 15**

1. **Fisierul 3\_\_LECTIA NR 9\_ARH\_DWH\_CUBURI.pptx**
2. **FISIERUL 4\_\_BI\_\_DATA\_\_MINING.pptx**
3. **Fisierul 5\_\_DM\_BI\_INTEGRAREA DATELOR\_\_1.pptx**

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.

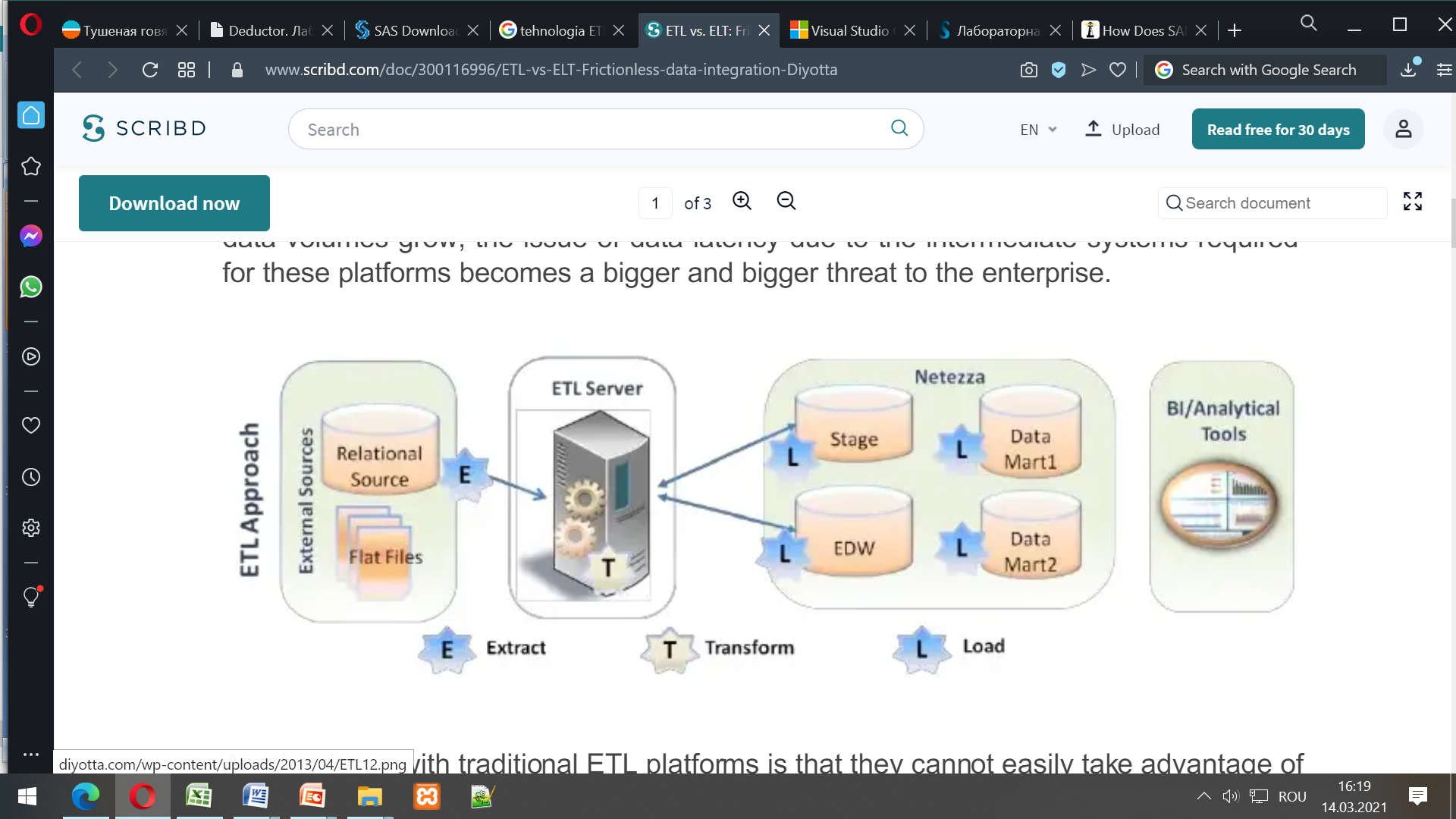
Această tehnologie este folosită în **conjuncţie cu data mining**, astfel încât volumul de informaţie stocat în depozitele de date este utilizat pentru generarea de cunoştinţe, motiv pentru care domeniul care se ocupă cu astfel de probleme, poartă denumirea de **Business Intelligence**.

**DE CE ASA ATENTIE ACESTEI TEHNOLOGII? SI DE CE ACUM?**

**Iată de ce??!!**

<https://www.google.com/search?q=tehnologia+ETL&source=lmns&bih=602&biw=1243&client=opera&hs=kkR&hl=en-US&sa=X&ved=2ahUKEwi057qZ-6_vAhWVNuwKHVchA9YQ_AUoAHoECAEQAA>

[DWH-BI ETL & Big Data Architect (\*) (dbschenker.com)](https://www.dbschenker.com/ro-ro/cariera/posturi-vacante/joboffer-202007100032%23en)

****

**Urmează prezentarea**

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.

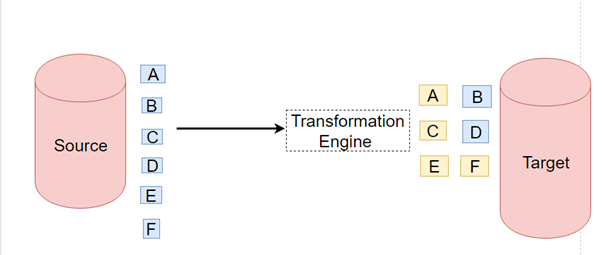
**ETL vs ELT: Must Know Differences**

**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**.

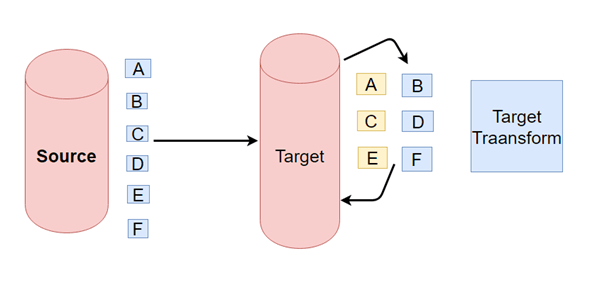
In ETL process transformation engine takes care of any data changes. /SPECIFIC PENTRU DEDUCTOR!!!/

[](https://www.guru99.com/images/1/022218_0954_ETLvsELTMus1.png)

**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.**

[](https://www.guru99.com/images/1/022218_0954_ETLvsELTMus2.png)

**Difference between ETL vs. ELT**

ETL and ELT process are different in following parameters:

|  |  |  |
| --- | --- | --- |
| **Parameters** | **ETL** | **ELT** |
| **Process**  **procesare** | Data is transformed at staging server and then transferred to Datawarehouse DB.  **Datele sunt transformate la serverul de stocare și apoi transferate la Datawarehouse DB.** | Data remains in the DB of the Datawarehouse.  **Datele rămân în baza de date a Datawarehouse** |
| **Code Usage** **Utilizare**  **cod** | Used for  **folosit pentru**   * Compute-intensive Transformations   **Transformări intensive în calcul**   * Small amount of data   **Cantitate mică de date** | Used for High amounts of data  **utilizate pentru cantități mari de date** |
| **Transformation** **Transformări** | Transformations are done in ETL server/staging area.  **Transformările se fac în serverul ETL / zona de stocare** | Transformations are performed in the target system  **Transformările sunt efectuate în sistemul țintă** |
| **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.
  1. ETL înseamnă Extract, Transform and Load, în timp ce ELT înseamnă Extract, Load, Transform
  2. În procesul ETL, fluxurile de date de la sursă la etapizare către țintă.
  3. ELT permite sistemului țintă să facă transformarea. Nu este implicat niciun sistem de stadializare / scenarii.
  4. ELT abordează multe provocări ale ELT, dar este costisitor și necesită abilități de nișă pentru implementare și întreținere.

http://aipi2014.andreirosucojocaru.ro/lib/exe/indexer.php?id=laboratoare%3Alaborator11&1549479854