

INSTRUMENTAR UTILIZAT ÎN ETL (EXTRACT TRANSFORM LOAD)

Ionuț DELEU

Departamentul Ingineria Software și Automatică, grupa TI-201FR, Facultatea Calculatoare Informatică și
Microelectronică, Universitatea Tehnică a Moldovei, Chișinău, Moldova

Autorul corespondent: Ionuț DELEU, ionut.deleu@isa.utm.md

Îndrumătorul/coordonatorul științific: **Dorian SARANCIUC**, lector universitar

Rezumat. În epoca digitală, unde datele devin resursele cheie pentru inovație și luarea deciziilor, procesul ETL (Extract, Transform, Load) devine un fundament esențial în arhitectura datelor. În această lucrare se analizează noțiunea de ETL, contextul său în gestionarea datelor și modul în care evoluția acestui proces își găsește locul în lumea tehnologiei și afacerilor. Sunt descrise instrumente și tehnologii moderne pentru realizarea procesului ETL. Este arătat rolul ETL în diverse domenii, provocările și bunele practici în procesul ETL.

Cuvinte cheie: ETL, Extract, Transform, Load, tehnologii, Apache NiFi, Talend, AWS Glue, Azure Data Factory, Google Cloud Dataflow.

Introducere

În epoca digitală, unde datele devin resursele cheie pentru inovație și luarea deciziilor, procesul ETL (Extract, Transform, Load) devine un fundament esențial în arhitectura datelor. În această introducere, vom explora definiția ETL, contextul său în gestionarea datelor și modul în care evoluția acestui proces își găsește locul în lumea tehnologiei și afacerilor.

Fundamentele Procesului ETL

Structura Celor Trei Etape: Extract, Transform, Load

Extracția Datelor: Extracția reprezintă prima treaptă în procesul ETL și constă în aducerea datelor din surse diverse către mediul de procesare. Această fază poate implica accesul la baze de date relaționale, preluarea datelor din fișiere CSV sau JSON și, uneori, conectarea la servicii API pentru a obține informații în timp real [1].

Transformarea Datelor: În această etapă complexă, datele sunt supuse unui proces riguros de transformare. Aici, curățarea datelor de eventuale imperfecțiuni, standardizarea formelor și structurilor, și îmbogățirea cu informații adiționale sunt aplicate pentru a pregăti datele pentru analiză. Diverse tipuri de transformări, de la filtre și agregări până la îmbogățirea semantică, joacă un rol crucial în modelarea datelor.

Încărcarea Datelor: În ultima etapă, datele transformate sunt încărcate în destinația finală. Aceasta poate fi un data warehouse tradițional, un depozit de date distribuit sau un serviciu cloud. Gestionarea metadatelor devine esențială pentru a asigura un acces optimizat la date și pentru a menține integritatea informațiilor.

Instrumente și Tehnologii ETL

A. Apache NiFi și Arhitectura Grafică

Apache NiFi este un instrument ETL open-source, cu o arhitectură grafică care facilitează automatizarea fluxului de date între sisteme diverse. Interfața vizuală oferă o modalitate intuitivă de a defini și gestiona fluxurile de date, simplificând procesul ETL.

B. Talend și Soluții ETL Open-Source

Talend reprezintă o platformă ETL care combină abordarea open-source cu opțiuni comerciale. Cu un mediu de dezvoltare grafic, Talend permite proiectarea și implementarea eficientă a proceselor ETL, oferind flexibilitate și scalabilitate.

C. Servicii Cloud ETL: AWS Glue, Azure Data Factory, Google Cloud Dataflow

Trecerea la medii cloud devine o tendință și în domeniul ETL. AWS Glue, Azure Data Factory și Google Cloud Dataflow sunt exemple de servicii cloud ETL gestionate, care aduc scalabilitate și flexibilitate în procesul de gestionare a datelor [2].

D. Tendințe și Inovații în Instrumentele ETL

O serie de inovații au fost introduse în instrumentele ETL, de la funcționalități avansate de gestionare a metadatelor până la integrarea cu instrumente de analiză și business intelligence. Tendințe precum ETL serverless și automatizarea proceselor aduc îmbunătățiri semnificative în eficiența operațională.

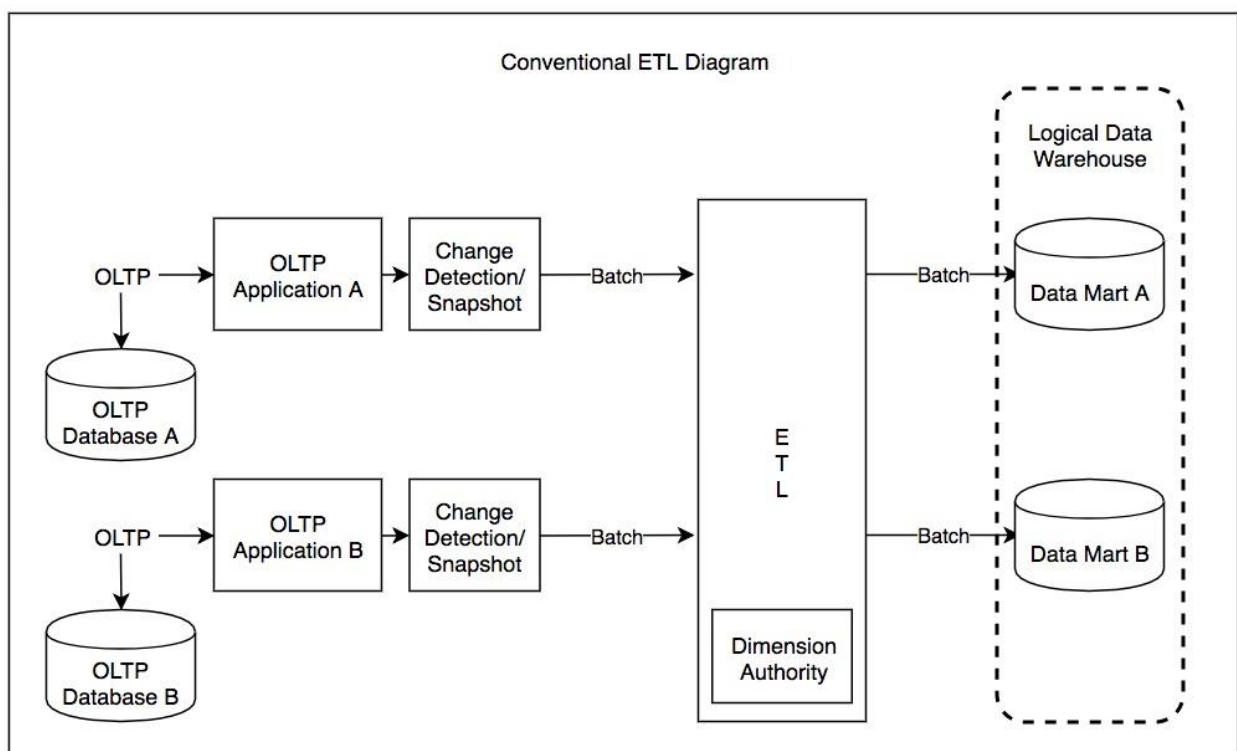


Figura 1. Diagrama ETL convențională

Rolul ETL în Diverse Domenii

A. Business Intelligence și Analiza de Date

În domeniul Business Intelligence, ETL alimentează depozitele de date care stau la baza rapoartelor și analizelor BI. Integrarea datelor din surse multiple permite generarea de insights valoroase pentru procesul decizional [3].

B. Sănătate și Asistență Medicală

În industria medicală, ETL este aplicat pentru integrarea datelor din sistemele de gestionare a pacienților, laboratoare și imagistică medicală. Acesta susține cercetările și diagnozele, asigurând coerența datelor medicale.

C. Retail și Comerț Electronic

În sectorul retail, ETL este esențial pentru analiza comportamentului clienților și gestionarea eficientă a stocurilor. Integrarea datelor din surse precum sistemele de vânzări și platforme de comerț electronic optimizează strategiile comerciale.

D. Telecomunicații și Gestionarea Datelor de Rețea

Companiile de telecomunicații utilizează ETL pentru analiza utilizării rețelei, facturare și optimizarea calității serviciilor. ETL joacă un rol vital în gestionarea masivelor volume de date din industria telecom.

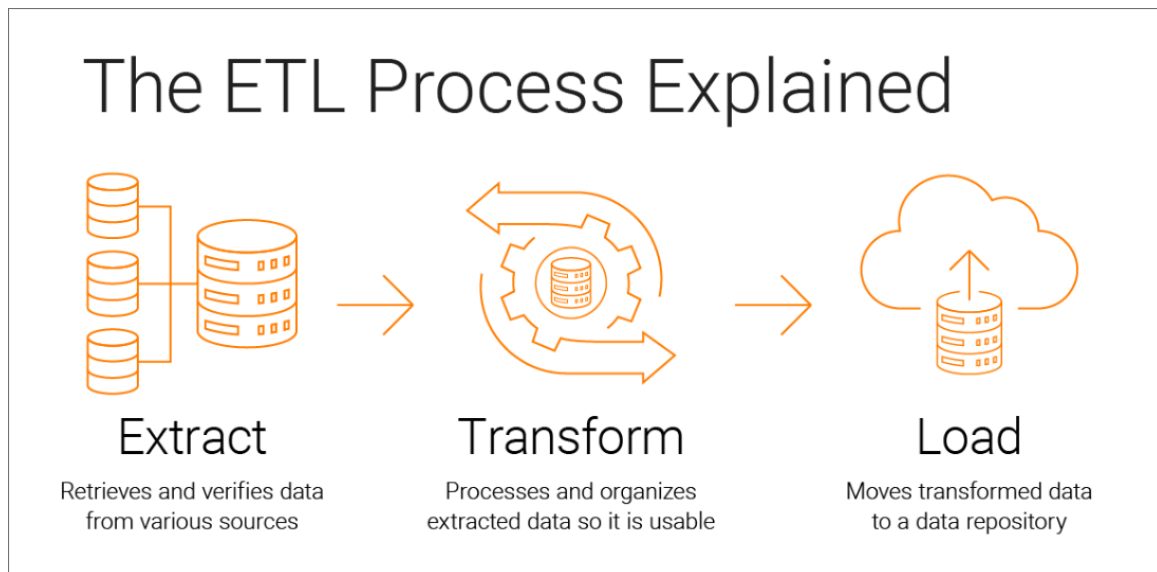


Figura 2. Explicația procesului ETL

Challenges și Best Practices în Procesul ETL

A. Gestionarea Erorilor și Mecanisme de Remediere

În fața provocărilor, gestionarea erorilor în timpul procesului ETL devine critică. Mecanisme eficiente de remediere și monitorizare sunt esențiale pentru menținerea integrității și fiabilității proceselor.

B. Verificări de Calitate a Datelor

Asigurarea calității datelor este un aspect esențial al procesului ETL. Implementarea de verificări și validări a calității datelor contribuie la obținerea rezultatelor analitice precise.

C. Securitatea Datelor în Timpul Procesului ETL

Într-un mediu în care se acordă o atenție deosebită securității datelor, procesul ETL trebuie să implementeze măsuri stricte de securitate. Criptarea datelor, controlul accesului și auditarea sunt elemente cheie [4].

D. Automatizare și Orchestrare Eficientă a Proceselor

Automatizarea proceselor ETL și orchestrarea acestora devin vitale pentru eficiența operațională. Utilizarea unor instrumente dedicate pentru programarea și gestionarea fluxurilor de date asigură o gestionare eficientă a resurselor.

Diferențe între ETL (Extract, Transform, Load) și ELT (Extract, Load, Transform)

Compararea între ETL (Extract, Transform, Load) și ELT (Extract, Load, Transform) evidențiază diferențe semnificative în abordarea proceselor de manipulare a datelor în domeniul depozitării și analizei acestora.

ELT reprezintă o alternativă la ETL, în care datele extrase sunt transferate și încărcate direct în sistemul țintă înainte de a fi transformate. Această abordare necesită o arhitectură de analiză bine concepută, care să includă etape de curățare și îmbogățire a datelor, precum și modalități eficiente de gestionare a dimensiunilor.

În cartea semnificativă a lui Ralph Kimball și Joe Caserta, "The Data Warehouse ETL Toolkit" (Wiley, 2004), utilizată ca resursă de predare pentru procesele ETL în depozitarea datelor, se explorează această problemă specifică [1].

O evoluție notabilă în acest context o reprezintă depozitele de date bazate pe cloud, cum ar fi Amazon Redshift, Google BigQuery, Microsoft Azure Synapse Analytics și Snowflake Inc. Aceste platforme oferă putere de calcul extrem de scalabilă, permițând companiilor să opteze pentru eliminarea transformărilor preîncărcare și să reproducă datele brute în depozitele lor de date. Ulterior, aceste date pot fi transformate folosind limbajul SQL conform necesităților specifice.

După implementarea procesului ELT, datele pot fi ulterior procesate și stocate într-un magazin de date.

Ambele abordări, ETL și ELT, prezintă argumente pro și contra. Majoritatea instrumentelor de integrare a datelor se orientează către ETL, în timp ce ELT este preferat în bazele de date și dispozitivele de depozit de date. Există, de asemenea, posibilitatea de a utiliza abordarea TEL (Transform, Extract, Load), unde datele sunt transformate inițial pe un blockchain (ca o modalitate de înregistrare a modificărilor datelor, de exemplu, arderea jetonelor) înainte de a fi extrase și încărcate într-un alt depozit de date.

Comparativ între "ETL pipeline" și "data pipeline"

Termenii "ETL pipeline" și "data pipeline" sunt uneori utilizați în mod interschimbabil. Cu toate acestea, există diferențe fundamentale între cele două.

Un "data pipeline" este folosit pentru a descrie orice set de procese, instrumente sau acțiuni utilizate pentru a prelua date din diverse surse și a le muta către un depozit țintă. Acest lucru poate declanșa acțiuni suplimentare și fluxuri de procesare în cadrul sistemelor sursă interconectate.

În cazul unui "ETL pipeline", datele transformate sunt stocate într-o bază de date sau depozit de date. Acolo, datele pot fi utilizate pentru analize de afaceri și pentru a obține perspective relevante.

Care sunt diferitele tipuri de "ETL pipelines"?

"ETL pipelines" sunt categorisite în funcție de latența lor. Cele mai obișnuite forme folosesc fie procesarea în lot (batch processing), fie procesarea în timp real (real-time processing).

Procesarea în lot (batch processing)

Procesarea în lot este utilizată în cazul analizelor tradiționale și al utilizărilor de business intelligence, unde datele sunt colectate, transformate și mutate periodic către un depozit de date în cloud [4].

Utilizatorii pot implementa rapid volume mari de date provenind din surse izolate într-un lac de date sau depozit de date în cloud. Pot programa apoi joburi pentru procesarea datelor cu intervenție umană minimă. În cadrul procesării în lot a "ETL", datele sunt colectate și stocate în timpul unui eveniment cunoscut sub numele de "fereastră de lot" ("batch window"). Loturile sunt utilizate pentru gestionarea eficientă a volumelor mari de date și a sarcinilor repetitive.

Procesarea în timp real (real-time processing)

"ETL pipelines" pentru procesarea în timp real permit utilizatorilor să preia date structurate și nestructurate din diverse surse de flux. Acestea includ Internet of Things (IoT), dispozitive conectate, fluxuri de social media, date de senzori și aplicații mobile. Un sistem de mesagerie cu un debit mare asigură capturarea precisă a datelor.

Transformarea datelor se realizează utilizând un motor de procesare în timp real, cum ar fi Spark streaming. Acest lucru alimentează funcționalități ale aplicației, cum ar fi analizele în timp real, urmărirea locației GPS, detectarea fraudelor, întreținerea predictivă, campanii de marketing țintite și îngrijirea proactivă a clienților.

Provocările trecerii de la "ETL" la "ELT"

Capacitățile sporite de procesare ale depozitelor de date în cloud și ale lacurilor de date au schimbat modul în care datele sunt transformate. Acest lucru a determinat multe organizații să treacă de la "ETL" la "ELT". Aceasta nu este întotdeauna o schimbare simplă [5].

Mapările "ETL" s-au dezvoltat suficient pentru a susține complexitatea în tipuri de date, surse de date, frecvență și formate. Conversia cu succes a acestor mapări într-un format care susține "ELT" necesită o platformă de date corporative capabilă să proceseze date și să susțină optimizarea pushdown fără a afecta front-end-ul. Ce se întâmplă dacă platforma nu poate genera codul specific ecosistemului sau depozitului de date necesar? Dezvoltatorii ajung să codeze manual interogările pentru a incorpora transformări avansate. Acest proces intensiv din punct de vedere al muncii este costisitor, complicat și frustrant. De aceea, este important să selectăm o platformă cu o interfață ușor de utilizat care poate gestiona replicarea acelorași mapări și să ruleze într-un model "ELT".

Trends și Viitorul ETL

A. ETL în Timp Real și Importanța Analizei În Timp Real

Tendința către analiza în timp real devine din ce în ce mai pronunțată în domeniul ETL. Procesele ETL în timp real permit organizațiilor să obțină insights imediate, fiind esențiale în medii de afaceri dinamice [6].

B. Integrarea cu Machine Learning și Inteligența Artificială

Integrarea procesului ETL cu tehnologii precum machine learning și inteligența artificială aduce un nivel superior de automatizare și anticipare în gestionarea datelor.

C. Evoluții Tehnologice: ETL în Cloud, Serverless ETL

Trecerea către mediul cloud și adoptarea arhitecturilor serverless în procesul ETL aduc beneficii semnificative în ceea ce privește scalabilitatea, flexibilitatea și costurile operaționale.

D. Rolul ETL în Era Analizei Big Data

Odată cu explozia datelor big data, ETL devine un jucător esențial în procesul de gestionare și analiză a acestor volume masive de informații.

Concluzii

A. Recapitularea Importanței Procesului ETL

Referatul a evidențiat importanța procesului ETL în gestionarea datelor, de la colectarea lor la transformarea și încărcarea în destinațiile finale. Rolul său esențial în furnizarea de informații strategice a fost subliniat în diverse domenii și industrii.

B. Impactul Transformațional al ETL în Organizații

Procesul ETL nu este doar o necesitate tehnică; este o forță transformațională în organizații, contribuind la eficiența operațională, luarea deciziilor informate și inovație.

C. Provocările Viitoare și Adaptabilitatea Procesului ETL în Era Digitală

Provocările viitoare, cum ar fi volumul exponențial al datelor și necesitatea de analize în timp real, cer o adaptabilitate continuă a procesului ETL. Inovațiile tehnologice vor continua să modeleze viitorul ETL într-o lume digitală în continuă schimbare.

Mulțumiri

Mulțumesc coordonatorului științific, Dorian SARANCIUC lect. univ. departamentul ISA pentru dedicare și ghidare științifică.

Surse bibliografice

- [1] Ralph., Kimball (2004). Setul de instrumente ETL pentru depozitul de date: tehnici practice pentru extragerea, curățarea, conformarea și livrarea datelor. Caserta, Joe, 1965-. Indianapolis, IN: Wiley. ISBN 978-0764579233. OCLC 57301227.
- [2] Denney, MJ (2016). „Validarea procesului de extragere, transformare, încărcare utilizat pentru a popula o bază de date mare de cercetare clinică”. Jurnalul Internațional de Informatică Medicală . 94 : 271–4. doi : 10.1016/j.ijmedinf.2016.07.009. PMC 5556907. PMID 27506144.

- [3] Zhao, Shirley (20.10.2017). „Ce este ETL? (Extract, Transform, Load) | Experian” . Calitatea datelor Experian.
- [4] IBM.COM. What is ETL (extract, transform, load)?. [Resursă electronică] [accesat 10.01.2024]. Disponibil: <https://www.ibm.com/topics/etl>
- [5] WIKIPEDIA.ORG. Extract, transform, load. [Resursă electronică] [accesat 10.01.2024]. Disponibil: https://en.wikipedia.org/wiki/Extract,_transform,_load
- [6] INFORMATICA.COM. What is ETL (extract transform load)? [Resursă electronică] [accesat 10.01.2024]. Disponibil: <https://www.informatica.com/nl/resources/articles/what-is-etl.html>