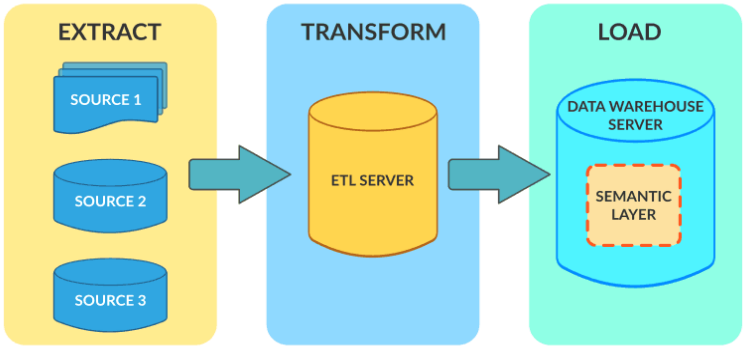
Data Science

Francesco Pugliese, PhD

neural1977@gmail.com

ETL – Extract, Transform and Load

- ✓ **ETL**, è un processo di Data Integration (Integrazione Dati) che combina i dati provenienti da diverse sorgenti di dati all'interno di una singola data store consistente che è in genere caricato in un data warehouse o un sistema Target.
- ✓ Man mano che i database sono cresciuti in popolarità intorno al 1970, **l'ETL** fu introdotto come processo di integrazione e caricamento dati per elaborazione ed analisi, e alla fine è divenuto il metodo primario per processare dati per i progetti di data **warehousing**.
- ✓ Un Enterprise Data Warehouse (EDW) è un sistema che aggrega dati provenienti da differenti sorgenti in un singolo data store che supporti processi come: data analysis, data mining or Artificial Intelligence (AI, ML)



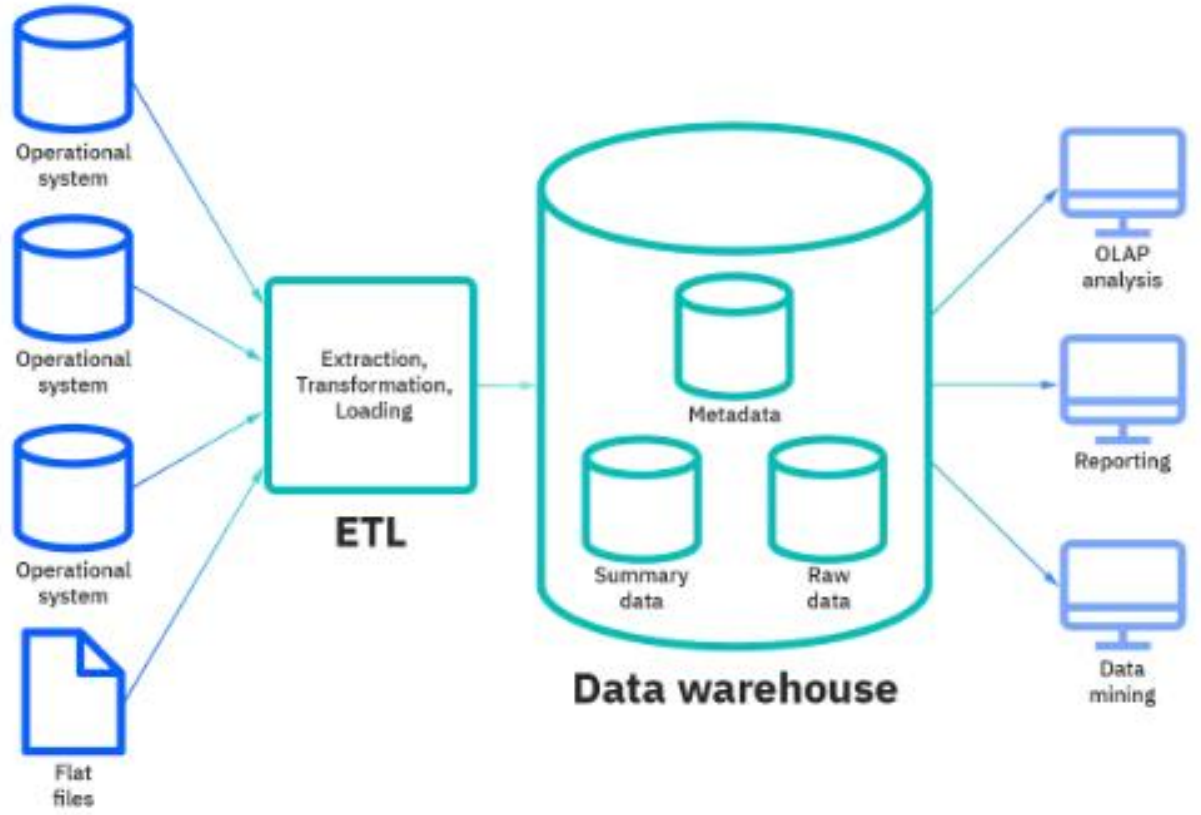
ETL – Data Warehouse

- Un **Data Warehouse** è un sistema che permette ad un'organizzazione di eseguire potenti analisi su elevati volumi (petabytes e petabytes) di dati in modalità che un database standard non è in grado di eseguire.
- I sistemi di **Data Warehouse** sono stati una parte dei sistemi di Business Intelligence per oltre 3 decenni, ma si sono evoluti solo di recente mediante nuovi titpi di dati e metodi di hosting.
- Originariamente un Data Warehouse veniva ospitato on-premises su un computer mainframe, e le sue funzionalità si focalizzavano sull'estrazione dei dati da varie sorgenti, pulizia e preparazione dei dati, caricamento e immagazzinamento dei dati all'interno di un database relazionale.
- Più recentemente, un Data Warehouse può essere ospitato su un dispositivo dedicato o su un cloud, e alla maggior parte dei sistemi di Data Warehouse sono state aggiunte capacità analitiche, di visualizzazione dati e tool di presentazione.

ETL – Architettura di un Data Warehouse

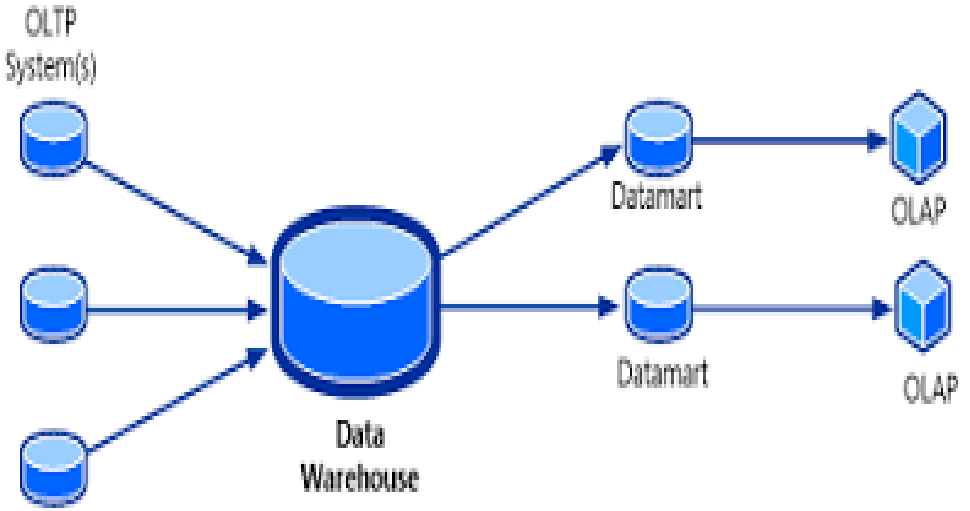
- ✓ Generalmente parlando ha una architettura three-tier (3 livelli):

Bottom tier: E' costituito da una data warehouse server, di solito si tratta di un sistema database relazionale, il quale colleziona, ripulisce e trasforma i dati provenienti da sorgenti di dati multiple attraverso un processo conosciuto come ETL (Extract, Transform and Load) o un processo conosciuto come Extract, Load and Transform (ELT).



ETL – Architettura di un Data Warehouse

Middle tier: Questo è costituito da un **OLAP** (ossia un **OnLine Analytical Processing**) Server che abilita l'utente ad avere delle velocità di query elevate. Esistono 3 tipi di modelli OLAP che possono essere usati in questo tier conosciuti come: **ROLAP**, **MOLAP** e **HOLAP**. Il tipo di modello OLAP usato è dipendente dal tipo di sistema database che esiste.

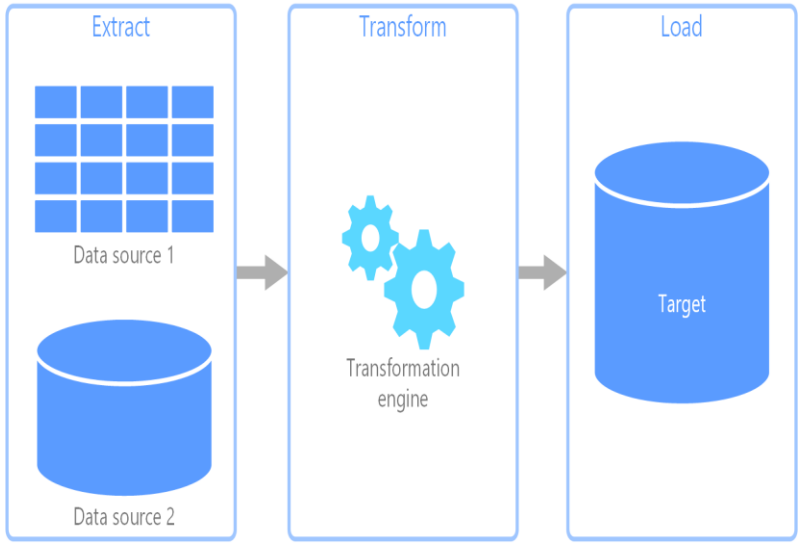


Top tier: Questo livello è rappresentato da qualcosa del tipo interfaccia **front-end user** o vari tool di reportistica, che abilita l'utente finale a condurre analisi ad-hoc sui loro dati di business.

ETL – Extract, Transform and Load

- ✓ **ETL** fornisce le fondamenta per la data analytics e i workstream di machine learning. Attraverso una serie di regole, l'ETL purifica e organizza i dati in un modo che incontra specifici bisogni di business intelligence, come report mensili ma piuò anche migliorare i processi di back-end o l'esperienza dell'utente finale.
- ✓ In genere **l'ETL** è utilizzato dalle organizzazioni per:

1) Estrarre dati da sistemi legacy
2) Ripulire i dati per miglirarne la qualità e renderli consistenti
3) Caricare i dati all'interno di un database target



ETL versus ELT

- ✓ La più semplice differenza tra **ETL** e **ELT** è in termini di operazioni. **ELT** copia ed esporta i dati dalle sorgenti, ma invece di caricarli in su un'area per la trasformazione successiva, **l'ELT** carica i dati grezzi direttamente sullo store di target dei dati per poter essere trasformati alla bisogna.
- ✓ Mentre entrambi **ETL** e **ELT** fanno leva su una varietà di repository di dati, quali database, data warehouse e data lake, ciascuno dei due processi possiede i suoi vantaggi e svantaggi.

1. **ETL** è particolarmente utile per dataset ad alto volume non strutturati dal momento che il caricamento può avvenire direttamente dalla sorgente. Questo processo richiede più definizione all'inizio, le regole di business per la data transformation hanno bisogno di essere costruite.
2. **ELT** è più ideale per nel mondo dei Big Data dal momento che non richiede una progettazione anticipata per la data extraction e lo storage dei dati. **ELT** è divenuto più popolare con l'adozione dei database su cloud, anche se non ci sono ancora molte best practices su **ELT**.

Trasformazione dei Dati (Data Transformation)

- ✓ L'Analisi dell'informazione richiede di solito dati accessibili e ben strutturati per ottenere i migliori risultati possibili. La Data Transformation rende alle organizziazioni possibile l'alterazione della struttura e del formato dei dati grezzi secondo le necessità. La Data Analytics più efficiente deriva anche dal modo in cui l'impresa trasforma i suoi dati.
- ✓ La Data Transformation

- ✓ Data transformation is the process of changing the format, structure, or values of data. For data analytics projects, data may be transformed at two stages of the data pipeline. Organizations that use on-premises data warehouses generally use an ETL (**extract, transform, load**) process, in which **data transformation is the middle step**. Today, most organizations use cloud-based data warehouses, which can scale compute and storage resources with latency measured in seconds or minutes. The

Bibliografia

https://www.stitchdata.com/resources/data-transformation