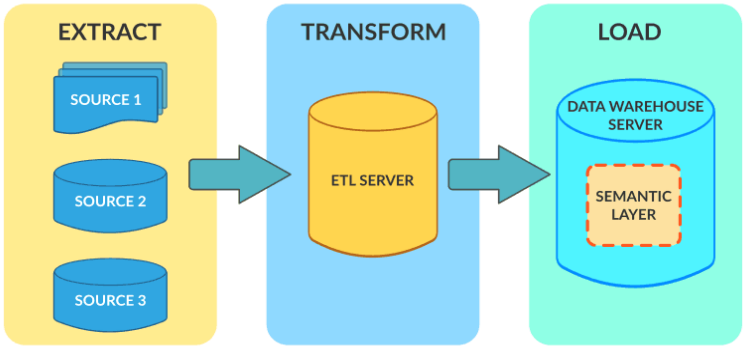
Data Science

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)



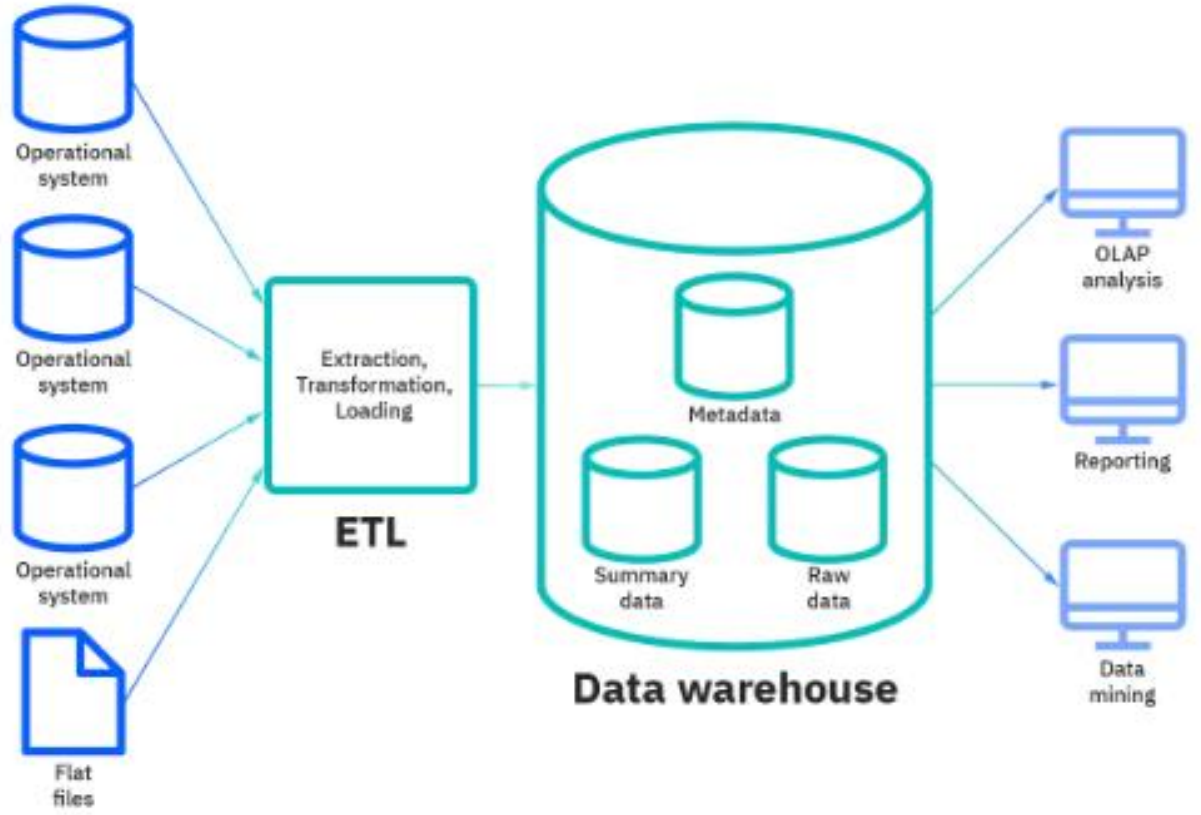
ETL – Data Warehouse

- 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

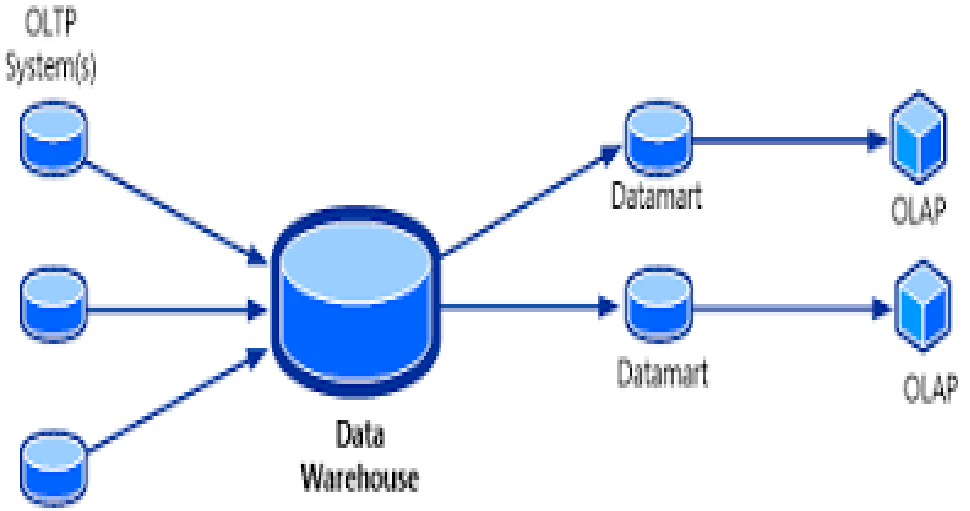
- ✓ Generalmente parlando ha una architettura three-tier (3 livelli):

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

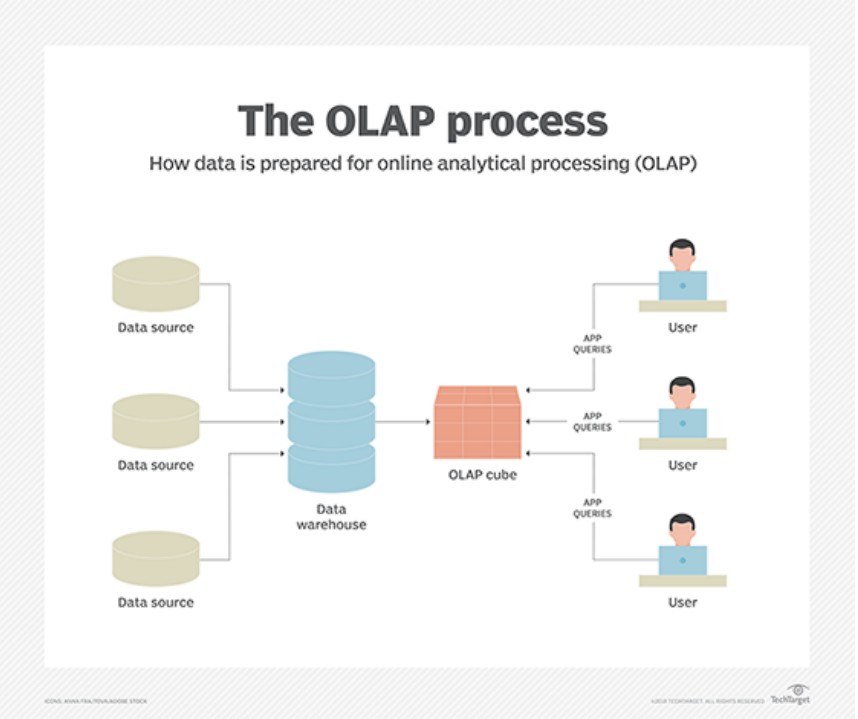
Middle tier: Questo è costituito da un **OLAP** (ossia un **OnLine Analytical Processing**) Server che abilita l'utente ad avere delle velocità di query elevate. Esistono 3 tipi di modelli OLAP che possono essere usati in questo tier conosciuti come: **ROLAP**, **MOLAP** e **HOLAP**. Il tipo di modello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

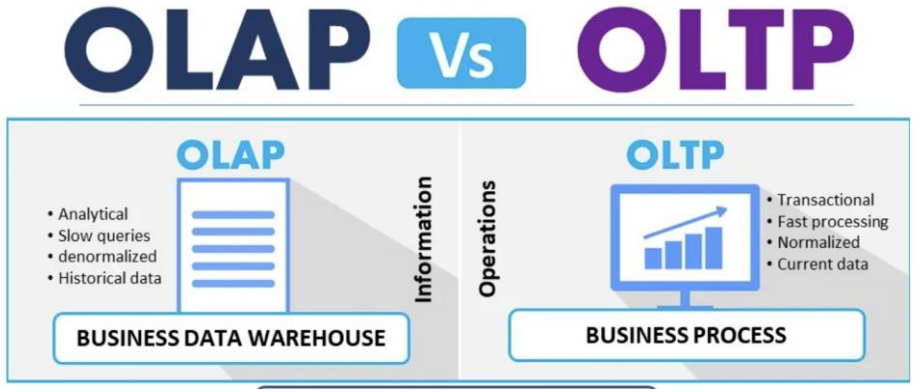
Data Warehouse – Comprendere OLAP e OLTP

- **OLAP (Online Analytical Processing)** è un software per eseguire analisi multidimensionali ad alta velocità su grandi volumi di dati provenienti da data store unificati e centralizzati come i **Data Warehouse**. **OLTP (Online Transactional Processing)** abilita l'utente ad avere un'esecuzione in tempo reale su grandi numeri di transazioni su database effettuate da un gran numero di persone, tipicamente su Internet.
- La principale differenza tra **OLAP** e **OLTP** è il nome: **OLAP** è analitico per natura mentre **OLTP** è transazionale.



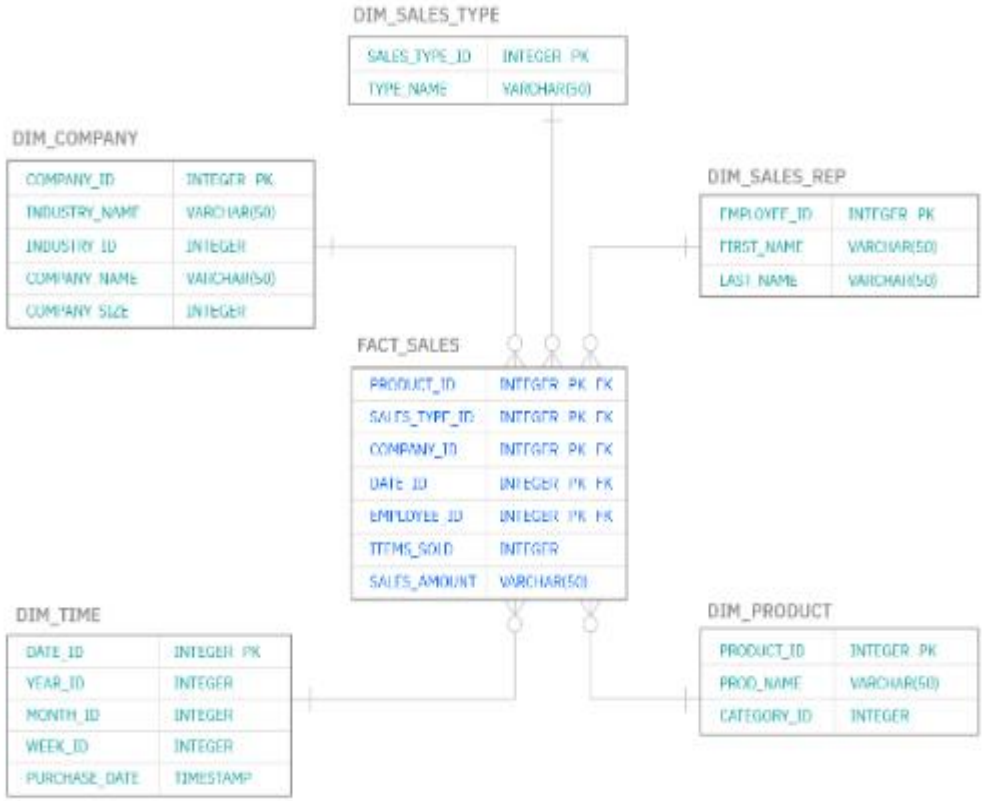
Data Warehouse – Comprendere OLAP e OLTP

- ✓ I tool **OLAP** sono progettati per l'analisi multidimensionale dei dati all'interno di un Data Warehouse, il quale contiene sia dati storici che transazionali. I comuni usi di OLAP sono il **Data Mining** ed altre applicazioni di Business Intelligence, calcoli analitici complessi, scenari predittivi, come anche funzioni di reportistica di business come analisi finanziaria, budgeting e forecast plannig.
- ✓ **OLTP** è progettato per supportare applicaziomni orientate alle transazioni elaborando transazioni recenti il più rapidamente e accurato possibile. Comuni usi di OLTP includono ATMs, software e-commerce, elaboraziojni di pagamenti di carte di credito, prenotazioni online, sistemi di prenotazioni, strumenti di record-keeping, ecc.



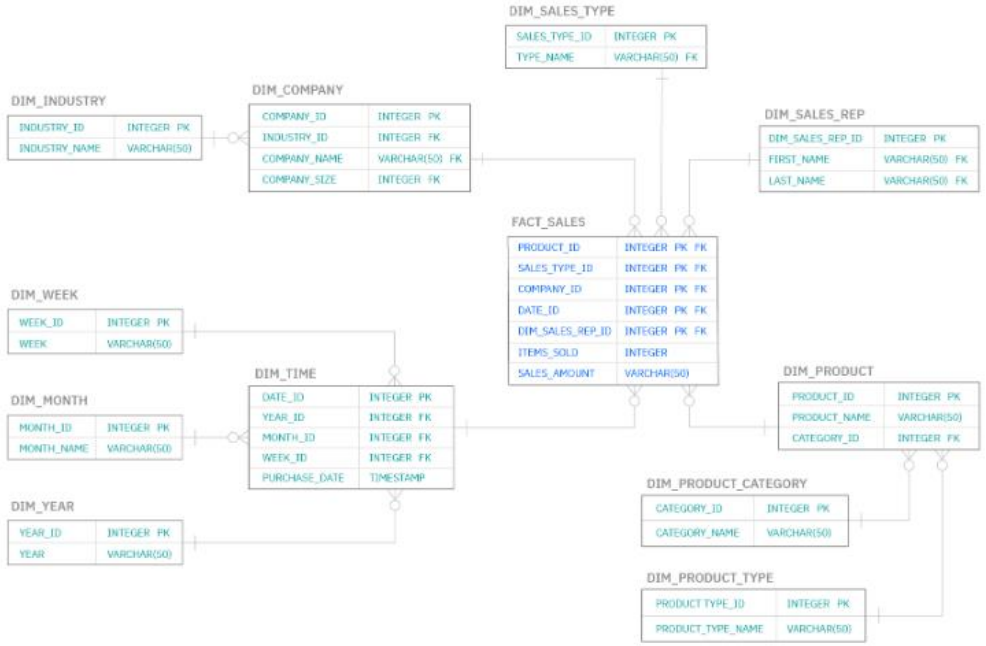
Data Warehouse –Schema

- Gli schemi sono i modi in cui i dati sono organizzati all'interno dei database o ddata ware house. Ci sono due principali strutture degli schemi: lo **star schema** e lo **snowflake**, che influenza il progetto del modello dei dati.
- **Star Schema:** Questo schema è costituito da una «fact table» che può essere unita ad un certo numero di «dimension table» denormalizzate. Esso è considerato il più comune tipo di schema, e i suoi utenti beneficiano delle sue rapide velocità durante il processo di query.



Data Warehouse –Schema

- **DSnowflake Schema:** Anche se non è ancora ampiamente usato, lo snowflake schema è un'altra organizzazione della struttura in un data warehouse. In questo caso la fact table è connessa a un numero normalizzato di dimension table. Questo schema è costituito da una «fact table» che può essere unita ad un certo numero di «dimension table» denormalizzate. Esso è considerato il più comune tipo di schema, e i suoi utenti beneficiano delle sue rapide velocità durante il processo di query.

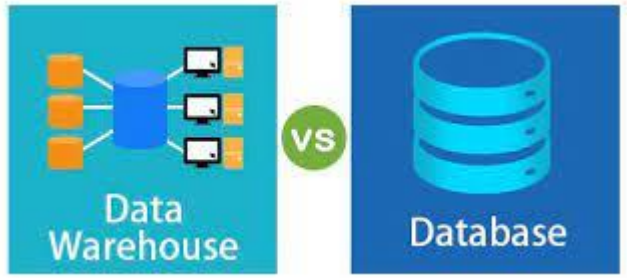


Data Warehouse vs Database, Data Lake, Data Mart

- **Data Warehouse vs Data Lake:** Un **Data Warehouse** colleziona e riunisce dati grezzi da sorgenti multiple di dati verso un repository centrale, strutturato usando uno schema predefinito di dati espressamente progettato per l'analisi dei dati. Mentre un **Data Lake** è un data warehouse senza uno schema predefinito. Come risultato, il data lake permette più tipi di analisi che un data warehouse semplice. I data lake sono comunemente costruiti su piattaforme di Big Data come Apache Hadoop.
- **Data Warehouse vs Data Mart:** Un **Data Mart** è un sottoinsieme di un data warehouse che contiene dati specifici per una particolare linea di business o dipartimento. Dal momento che il data mart contiene un più piccolo sottoinsieme di dati, i data mart permettono ai dipartimenti o linee di business di scoprire informazioni di valore più focalizzati e più rapidamente di quanto sia possibile lavorando con un dataset più ampio presente in un data warehouse.

Data Warehouse vs Database, Data Lake, Data Mart

- ✓ **Data Warehouse vs Database:** Un database è costruito principalmente per soddisfare rapide query ed elaborazioni di transazioni, non per fare analytics.
- ✓ Un database tipicamente viene utilizzato come **data store** focalizzato ad una specifica applicazione, mentre un data warehouse immagazzina dati provenienti da qualsiasi applicazione o persino tutte quelle appartenenti in una organizzazione.
- ✓ Un database si focalizza su aggiornamento dati in tempo reale mentre un data warehouse ha un obiettivo più ampio, catturando serie storiche di dati o correnti per effettuare analisi predittiva, machine learning e altri tipi di analisi avanzate.



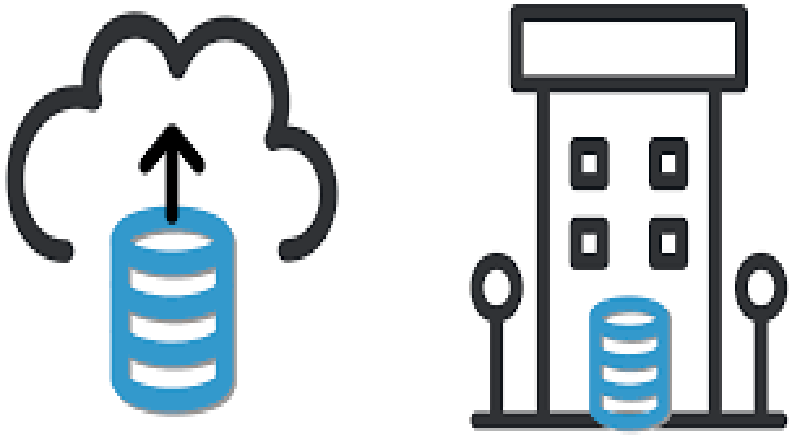
Tipi di Data Warehouse

- **Data Warehouse su Cloud:** Un Cloud Data Watehouse è un data warehouse specificatamente costruito per essere eseguito su cloud, e viene offerto ai clienti come un servizio gestito dal cloud. I data warehouse basati su cloud sono divenuti sempre più popolare negli ultimi 5 anni dal momento che molte compagnie usano servizi cloud per cercare di ridurre sempre più l'impatto dei loro **data center** on-premises.

- Con il termine **software on premise** (od on premises, come sarebbe più corretto) si fa riferimento alla fornitura di programmi informatici installati e gestiti attraverso computer locali. Deriva dall'inglese “on the premises”: nelle sedi, nei locali (del titolare della licenza).

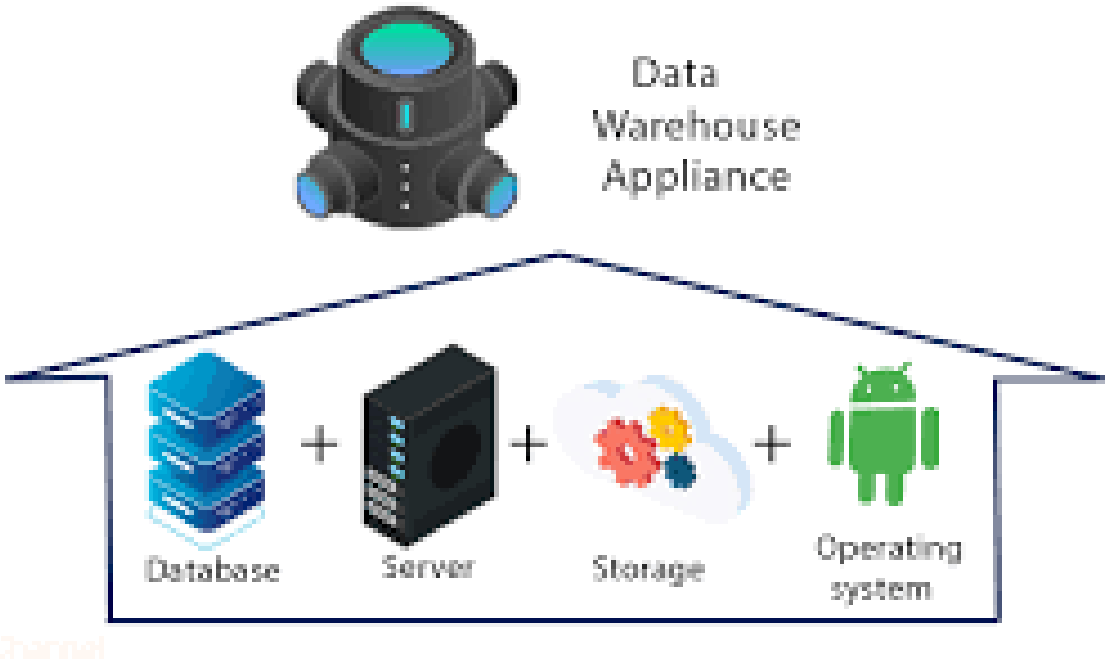
Tipi di Data Warehouse

- **Data Warehouse Software (on-premises / su licenza):** Una organizzazione può acquistare un data warehouse sotto licenza e poi deployare il data warehouse sulla propria infrastruttura on-premises. Sebbene questo sia tipicamente più costoso di un servizio di data warehouse su cloud, può essere una scelta migliore per entità governative (come **l'ISTAT**), istituzione finanziarie o altre organizzazioni che vogliono avere più controllo sui loro dati o anno la necessità di soddisfare rigide norme di sicurezza o standard di privacy dei dati o regolamentazioni varie.



Tipi di Data Warehouse

- **Apparati di Data Warehouse (Data Warehouse Appliance):** Un'apparato di Data Warehouse è un insieme di sistemi hardware e software come CPU storage, sistema operativo e data warehouse software che un'organizzaione può connettere alla su rete e usarla come parte di essa. Un data warehouse appliance si colloca tra cloud e le implementazioni on-premises in termini di costi, velocità di deployment, scalabilità, e controllo di gestione.



Benefici di un Data WarehouseTipi di Data Warehouse

1. **Migliore qualità dei dati:** Un data warehouse centralizza i dati da una varietà di sorgenti di dati, come sistemi transazionali, database operazionali, e file piatti. Dunque, ripulisce i dati, elimina i duplicati e li standardizzato per creare un unica sorgente di dati.
2. **Più veloce e informazioni di business:** I dati provenienti da disparate sorgenti limitano il potere decisionale dei decision makers per avviare strategie di business con una certa affidabilità. I data warehouse permettono la **Data Integration** (Integrazione Dati), permettendo agli utenti del business di estrarre tutte le informazione necessarie dai dati della compagnia durante ciascuna decisione di business.

Benefici di un Data WarehouseTipi di Data Warehouse

3. **Decision-making più intelligente:** Un data warehouse supporta funzioni di Business Intelligence ad ampia scala come il **data mining** (che cerca pattern e relazioni nei dati), intelligenza artificiale e machine learning. I professionisti e i leader di business possono usare i dati per prendere decisioni smart in virtualmente ogni area dell'organizzazione, dai processi di business al management finanziario all'inventory management.

4. **Guadagnare e far crescere un vantaggio competitivo:** Tutti i benefici visti si combinano per aiutare un organizzazione a trovare più opportunità nei dati, più rapidamente di quanto sia possibile con data store displocati in luoghi diversi e disparati.

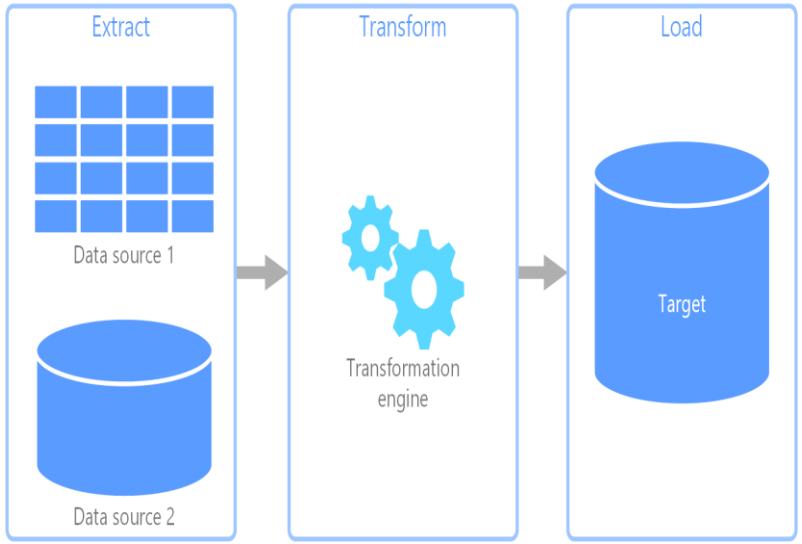
Database Transazionali

- I **Database Transazionali** sono ottimizzati per l'esecuzione di sistemi di produzione, dai siti web alle banche fino ai negozi al dettaglio. Questi database si distinguono per la rapidità di lettura e scrittura di singole righe di dati senza alterarne l'integrità. Questi database transazionali sono archivi di righe, dunque i dati sono archiviati su disco come righe anzichè colonne. Gli archivi di righe sono molto utili quando è necessario sapere tutto di un cliente nella tabella utente, per esempio. Tuttavia non sono ottimali quando per esempio si cerca di conteggiare i clienti di un determinato codice postale, in quanto è necessario caricare in memoria nonsolo la colonna codice postale ma anche le colonne nome, indirizzo, ecc. Dunque i DB transazionali non sono creati per l'analisi.

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 – Sistemi Legacy

- Un sistema **legacy**, in informatica, è un sistema informatico, un'applicazione o un componente obsoleto, che continua ad essere usato poiché l'utente (di solito un'organizzazione) non intende o non può rimpiazzarlo. Legacy equivale a versione "retrodatata" (rispetto ai sistemi/tecnologie correnti). Un esempio sono il **Cobol** o i **Mainframes** dei sistemi bancari.

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.
- ✓ 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 scalability of the cloud platform lets organizations skip preload transformations and load raw data into the data warehouse, then transform it at query time — a model called ELT (**extract, load, transform)**.

RDF

- ✓ 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.
- ✓ Cosa è l'IRI in un RDF

- ✓ Un **URI** può essere classificato come qualcosa che definisce posizioni (**URL**) o nomi (URN) o entrambi. Un **URL** (Uniform Resource Locator) **è** un **URI** che identifica una risorsa tramite la sua "collocazione" ("location") in un grafo. Di fatto, non identifica la risorsa per nome, ma con il modo con cui la si può reperire.

TurtleDB e Triplestore

turtleDB is a framework for developers to build offline-first, collaborative web apps. It provides a user-friendly API for developers, empowering them with the ability to create apps with in-browser storage, effective server synchronization, document versioning, and flexible conflict resolution for any document data.

Web applications will work seamlessly online or offline, and developers can leave the backend to turtleDB - it will handle all data synchronization and conflict resolution between users. Works with MongoDB out of the box!

Bibliografia

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

https://www.ibm.com/cloud/learn/data-warehouse#:~:text=A%20data%20warehouse%2C%20or%20enterprise,AI)%2C%20and%20machine%20learning.