Data Warehousing (traduzione da Lembo)

Francesco Pugliese, PhD

neural1977@gmail.com

Business Intelligence

- ✓ Al fine di permetterte ai maganer di realizzare strumenti potenti di analisi, è necessario definire l'infrastruttura software e hardware appropriata che può essere costituita da:

1. **Hardware Dedicato**
2. **Infrastrutture di rete**
3. **DBMS**
4. **Software di Back-end**
5. **Software di Front-end**

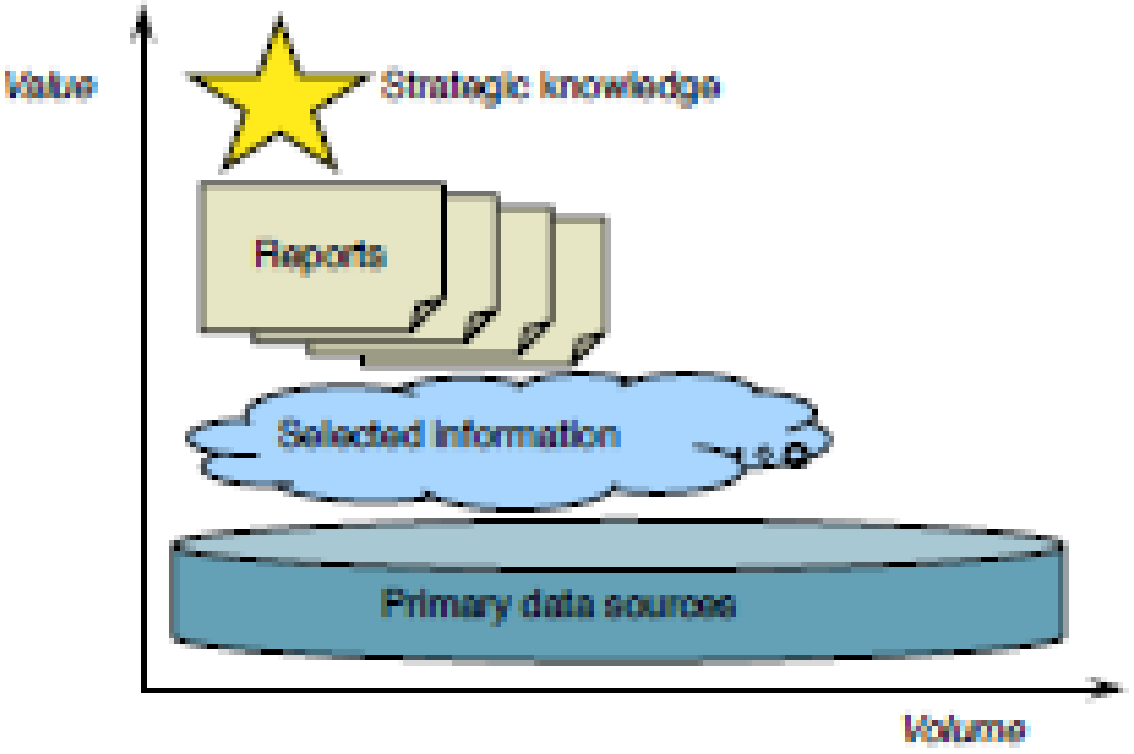
- ✓ Il ruolo chiave di una piattaforma di Business Intelligence è di trasformare i dati di Business in Informazione sfruttabile a diversi livelli di dettaglio

Dai Dati all'Informazione

- ✓ L'informazione è un importante asset (risorsa) di un'azienda o organizzazione in generale, che deve essere necessariamente controllata e pianificata dalle attività di business
- ✓ L'Informazione è il materiale grezzo che è trasformato a partire dai sistemi informativi, come prodotti i semi-finiti sono trasformati a partire dai sistemi di produzione
- ✓ **I dati sono diversi dall'Informazione**
- ✓ Spesso la disponibità di troppi dati rende difficile, se non impossibile, il compito di estrapolare informazione dai dati

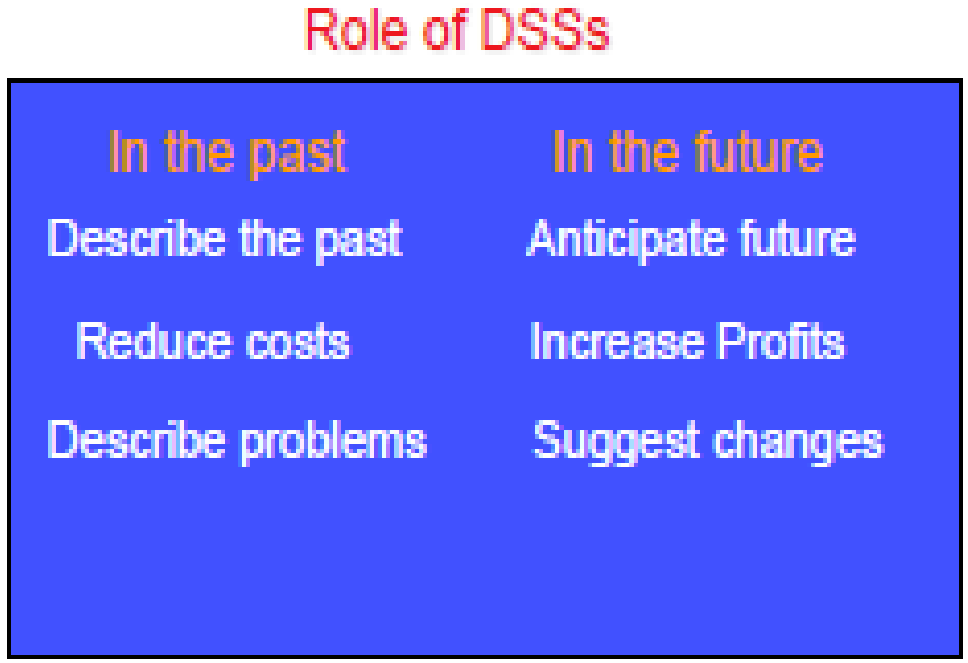
Dai Dati all'Informazione

- ✓ Ogni azienda deve havere accesso a informazione **rapida e completa,** in quanto viene richiesta in questo modo da sistemi di **decision making.**
- ✓ Questa informazione strategica viene estratta principalmente da un'elevata quantità di data operazionali immagazzinati in database enterprise per mezzo di una progressiva selezione ed un processo di aggregazione



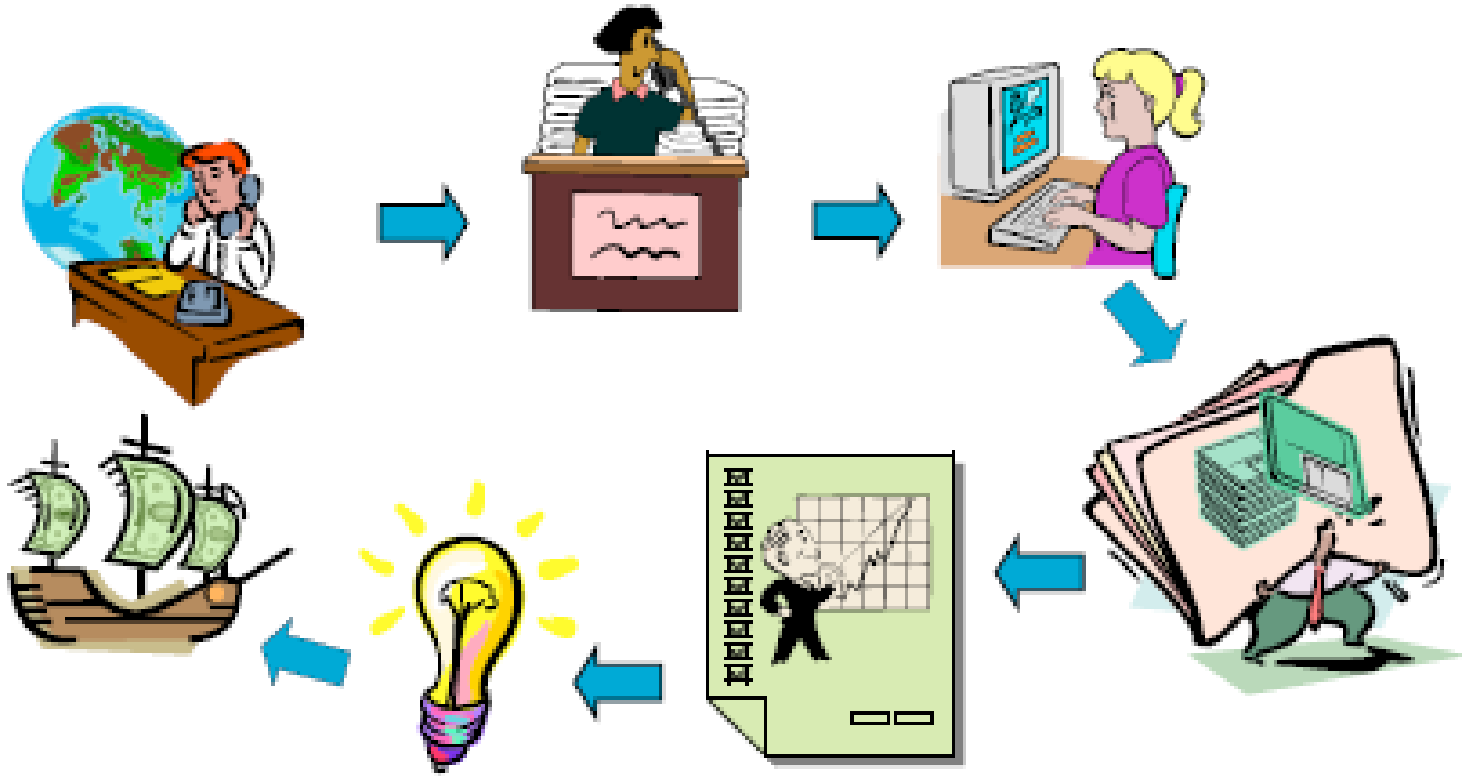
Sistemi di Supporto alle Decisioni

- I **Sistemi di Supporto alle Decisioni (Decision Support Systems - DSSs)** hanno iniziato ad essere popolari negli anni '80.
- Un **DSS** è un insieme di strumenti espandibili e interattivi progettati per elaborare e analizzare dati e per supportare i manager nel prendere delle decisioni.
- I sistemi di Data Warehouse hanno gestito i back end di dati dei DSSs a partire dal 1990.



Un tipico Scenario

- Un **tipico scenario** è che una grande compagnia, con numerosi settori, i cui manager vogliono valutare il contributo di ciascun settore alla performance del business totale della compagnia.



OLTP e OLAP

- ✓ Mescolare insieme query **analitiche** e **transazionali** conduce ad un inevitabile ritardo che rende gli utenti non soddisfatti di entrambe le categorie.
- ✓ Uno dei principali scopi del **Data Warehousing** è mantenere separato **OLAP (On-Line Analyticial Processing)** da OLTP **(On-Line Transactional Processing)**.

Alcune aree dove le tecnologie DW sono normalmente adottate

- ✓ **Commercio:** analisi di vendite, spedizioni, controllo di inventario, customer care
- ✓ **Manifatturiero:** controllo dei costi di produzione, supporto dei fornitori e ordini
- ✓ **Servizi finanziari:** analisi del rischio, individuazione di frodi
- ✓ **Trasporto:** gestione della flotta
- ✓ **Telecomunicazioni:** analisi dei dati di call, profilazione cliente
- ✓ **Healtcare:** analisi delle ammissioni e dimissioni, bilancio dei centri di costo

Caratteristiche di un Data Warehouse

- ✓ **Accessibilità:** per gli utenti con non molte competenze in IT e strutture dati
- ✓ **Integrazione:** di dati sulla base di un modello di enterprise standard
- ✓ **Flessibilità delle Query:** per massimizzare i vantaggi ottenuti a partire dall'informazione esistente
- ✓ **Rappresentazione Multidimensionale:** fornisce agli utenti un intuitivo e gestibile punto di vista dell'informazione
- ✓ **correttezza e completezza:** dei dati integrati

Data Warehouse

- ✓ **Un Data Warehouse** è una collezione di dati che supportano i processi di decision-making.
- ✓ **Un DW fornisce** le seguenti proprietà (Inmon, 2005):
- ✓ **E' orientato al soggetto:** enfasi sui soggetti e non sulle applicazioni come nei sistemi operazionali
- ✓ **E' integrato:** prende vantaggio da mutiple sorgenti di dati
- ✓ **E' consistente:** dovrebbe fornire un punto di vista unificato e riconciliato dei dati

un Data Warehouse mostra la sua evoluzione nel tempo

- ✓ **I Dati operazionali** di solito coprono periodi di tempo molto brevi dal momento che le transazioni riguardano solo i dati recenti. Non esistono dati storici: i dati una volta aggiornati ricevono una cancellazionde dei vecchi valori. Il tempo non è un elemento chiave per i DB operazionali

- ✓ **Un Data Warehouse** rende possibili analisi che coprono anche alcuni anni. I DW vengono regolarmente aggiornati e crescono continuamente. Il tempo è una componente chiave nei DW.

un Data Warehouse è non volatile

- ✓ **I dati non sono mai cancellati** dal Data Warehouse e gli aggiornamenti sono normalmente eseguiti quando i data warehouse sono offline.
- ✓ **Questo significa** che i Data Warehouse possono essere essenzialmente visti come Database di sola lettura.
- ✓ In un **DW** non c'è nessun bisogno di tecniche di gestione avanzata delle transazioni invece richieste dalle applicazioni operazionali.
- ✓ I problemi sono la capacità effettiva delle query e la resilienza.

Query

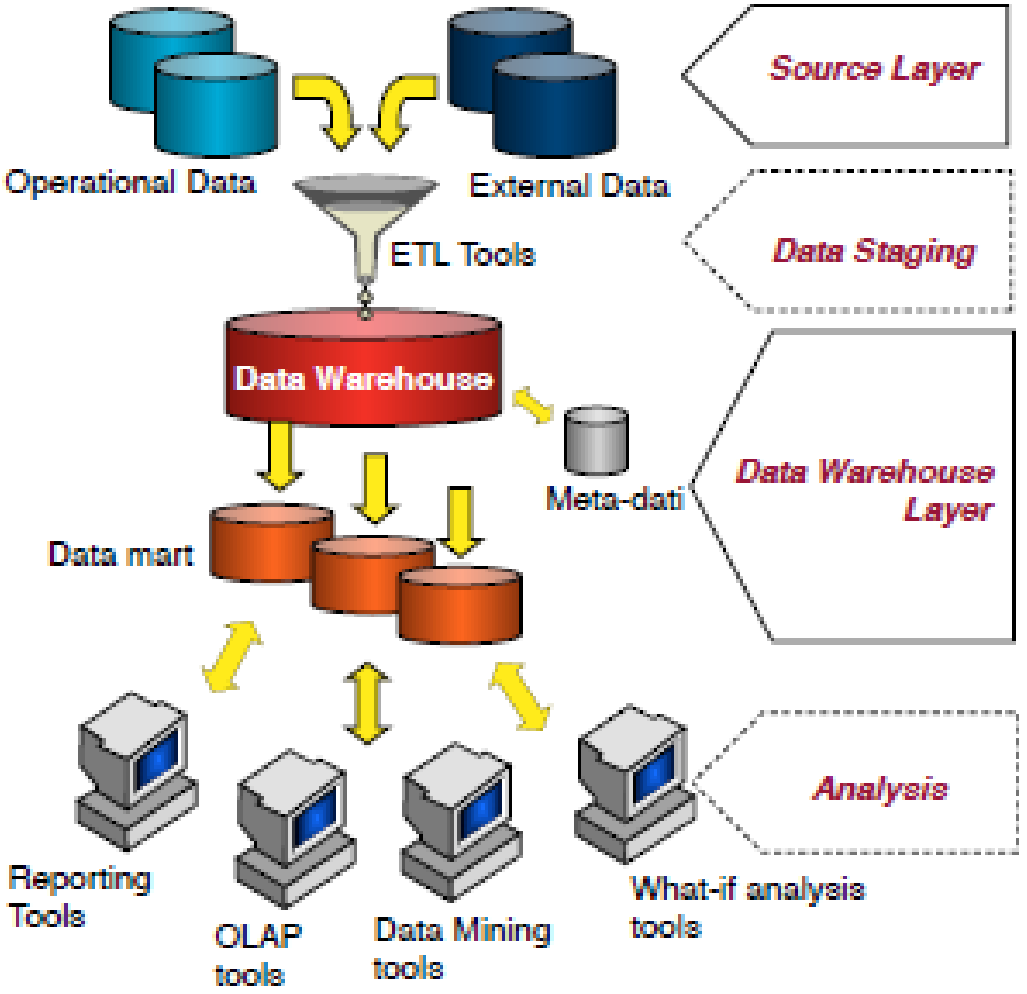
- ✓ **OLTP:** Le query operazionali eseguono transazioni che generalmente leggono e scrivono un piccolo numero di tuple da e per molte tabelle connesse da semplici relazioni. Per esempio, questo si applica se vuoi cercare i dati di un cliente al fine di inserire un nuovo ordine cliente. Il carico di lavoro core è spesso "congelato" in applicazioni (le query di dati ad hoc sono occasionali).
- ✓ **OLAP:** Le query eseguono analisi dinamiche e multidimensionali che hanno bisogno di scansionare un enorme quantità di record per processare un insieme di dati numerici che riassumono le performance di un'enterprise. Il **DW** ha una proprietà chiamata interattività (interactivity) che è essenziale per le sessioni di analisi, in questo modo il carico di lavoro varia costantemente al variare del tempo.

Requisiti per le Architetture di DW

- ✓ **Separazione:** L'elaborazione analitica e quella transazionale dovrebbero essere tenute separate il più possibile.
- ✓ **Scalabilità:** Le architetture hardware e software dovrebbero essere facili da aggiornare all'aumentare del volume dei dati e del numero di requisiti degli utenti
- ✓ **Estensibilità:** L'architettura dovrebbe essere capace di ospitare nuove applicazioni e tecnologie senza riprogettare l'intero sistema.
- ✓ **Sicurezza:** Il monitoraggio degli accessi è essenziale in seguito ai dati strategici immagazzinati nei data warehouse.
- ✓ **Amministrabilità:** La gestione dei DW non dovrebbe essere eccessivamente difficile.

Architettura Two-Layer

- **Data Mart:** Sottoinsieme o aggregazione dei dati immagazzinati in un data warehouse primario. Esso include un insieme di pezzi di informazione rilevanti per una specifica area di business, dipartimenti corporate, o categoria di utenti.
- Il nome evidenzia una separazione tra sorgenti disponibili fisicamente e data warehouse, ma infatti esso consiste di 4 fasi di flussi di dati successivi.



Architettura Two-Layer

- ✓ **I Data Mart** popolati da un data warehouse primario sono spesso chiamati **dipendenti**. Essi sono utili per sistemi di data warehouse da media dimensione a grandi enerprise in quanto:
- ✓ sono usati come builting block durante lo sviluppo incrementale dei data warehouse
- ✓ essi segnano l'informazione richiesta da uno specifico gruppo di utenti per risolvere le query
- ✓ possono distribuire una migliore performance dal momento che sono più piccoli di data warehouse primari.

Architettura Two-Layer

- ✓ **A volte** principalmente per scopi di organizzazione e policy, una soluzione differente può essere adottata in cui le sorgenti sono usati per popolare direttamente i data mart.
- ✓ **Questi Data Mart** sono chiamati indipendenti.
- ✓ Se non c'è nessun data warehouse primario, questo snellisce il progetto ma può portare al riuschio di inconsistenza tra data mart.

Architettura Two-Layer

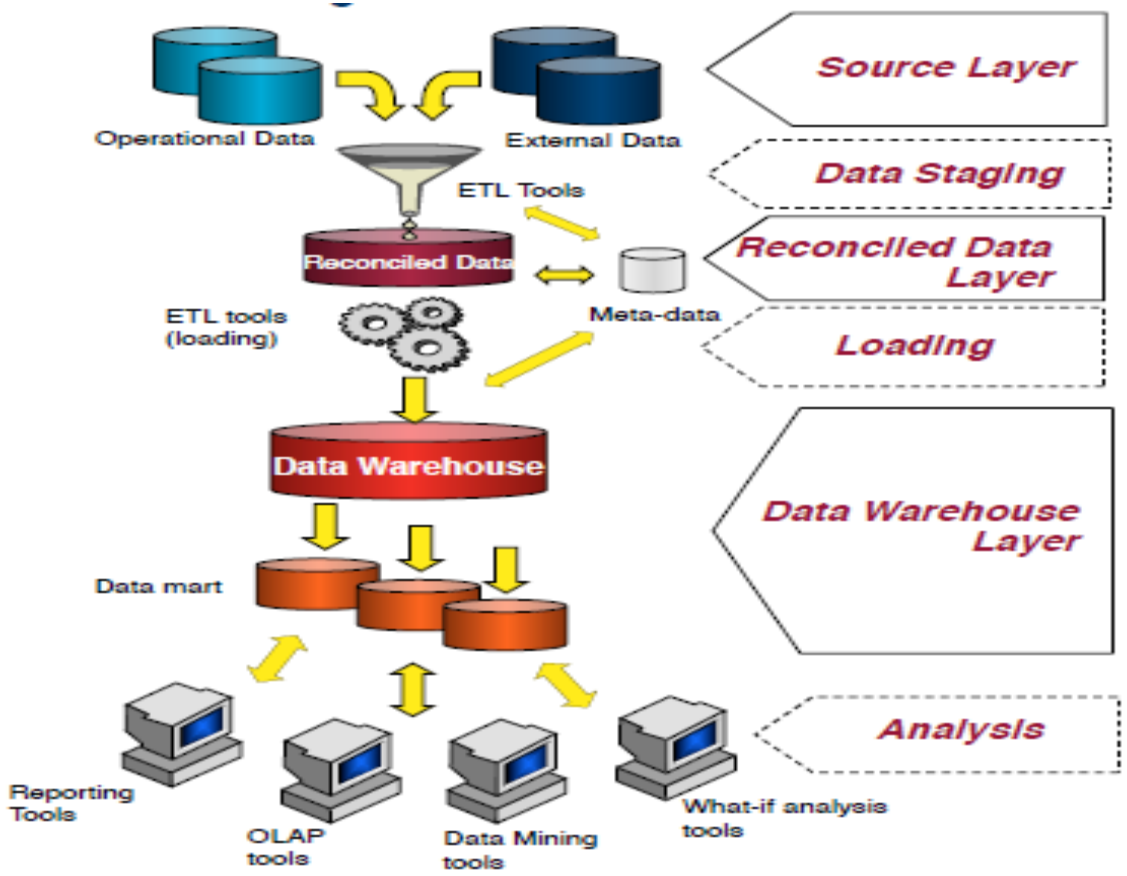
- ✓ **Vantaggi:**
- ✓ **Nei sistemi di data warehouse,** l'informazione di buona qualità è sempre disponibile, anche quando l'accesso alle sorgenti è temporaneamente negato per ragioni tecniche o organizzative.
- ✓ **Le query di analisi dei Data Warehouse** non influenzano la gestione delle transazioni, l'affidabilità della quale è vitale per le imprese per lavorare propriamente a livello operazionale.
- ✓ **I DW** sono logicamente strutturati secondo il modello multidimensionale, mentre le sorgenti operazionali sono generalmente basate su modelli relazionali o semi-strutturate.

Architettura Two-Layer

- ✓ **Vantaggi:**
- ✓ **Una mancata corrispondenza** in termini di tempo e granularità può verificarsi tra sistemi **OLTP** che gestiscono dati correnti ad un massimo livello di dettaglio, e i sistemi **OLAP,** che gestiscono i dati storici e sintetizzati (aggregati).
- ✓ **I data warehouse possono usare specifiche soluzioni di progetto** che hanno lo scopo dell'ottimizzazione di performance per l'analisi e le applicazioni di reportistica

Architettura Three-Layer

- ✓ **Dati Riconciliati:**
- ✓ **Questo strato materializza** i dati operazionali ottenuti dopo l'integrazione e la pulizia dei dati provenienti dalle sorgenti. Come risultato questi dati sono integrati, consistenti, corretti e dettagliati.



Architettura Three-Layer

- ✓ **I principali vantaggi** di questo layer aggiuntivo per i **Dati Riconciliati** sono quelli di creare un dato di riferimento comune per l'intera organizzazione
- ✓ Allo stesso tempo, questo layer nettamente separa i problemi dell'estrazione dati dalle sorgenti e l'integrazione da quelle della popolazione di data warehouse.
- ✓ Tuttavia, i dati riconciliati portano ad avere **più ridondanza** delle sorgenti dati operazionali.

Bibliografia