Data Warehouse
Architettura e Progettazione

Introduzione

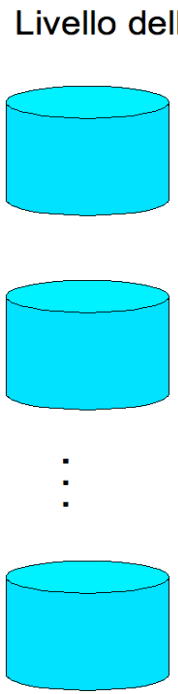
- Nei seguenti lucidi verrà fornita una panoramica del mondo dei Data Warehouse.
- Verranno riportate diverse definizioni per identificare i molteplici aspetti che caratterizzano uno strumento cosi complesso.
- Nel campo della BI infatti il termine Data Warehouse è spesso usato senza precisione, a volte ci si riferisce a tutto il sistema di analisi, a volte ad una singola base di dati implementata in maniera non relazionale.

Definizione di data warehouse (Ralph Kimball)

- Una collezione di dati di supporto al processo decisionale con le seguenti proprietà:
  - E' orientata ai soggetto (o argomenti di analisi).
  - E' integrata e consistente.
  - E' rappresentativa dell'evoluzione temporale.
  - E' non volatile.

- Di seguito analizziamo più in dettaglio questi quattro punti

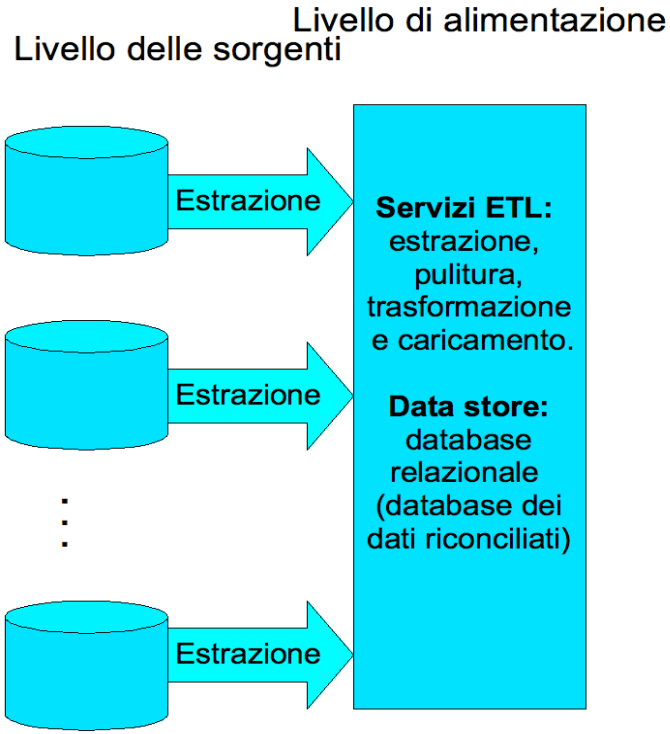
Componenti di un data warehouse (1)



Livello delle sorgenti (Operational source system)

- Il livello delle sorgenti incapsula i sistemi che memorizzano i dati di base.
- Può essere in realtà pensato come un livello esterno al DW, visto che non abbiamo nessun controllo sul formato dei dati e sulla loro qualità.
- Le interrogazioni sono di solito piuttosto standardizzate e coinvolgono pochi record alla volta.
- Spesso le sorgenti non sono integrate tra loro, cioè ogni applicazione ha il suo database, senza condivisione di dati comuni.

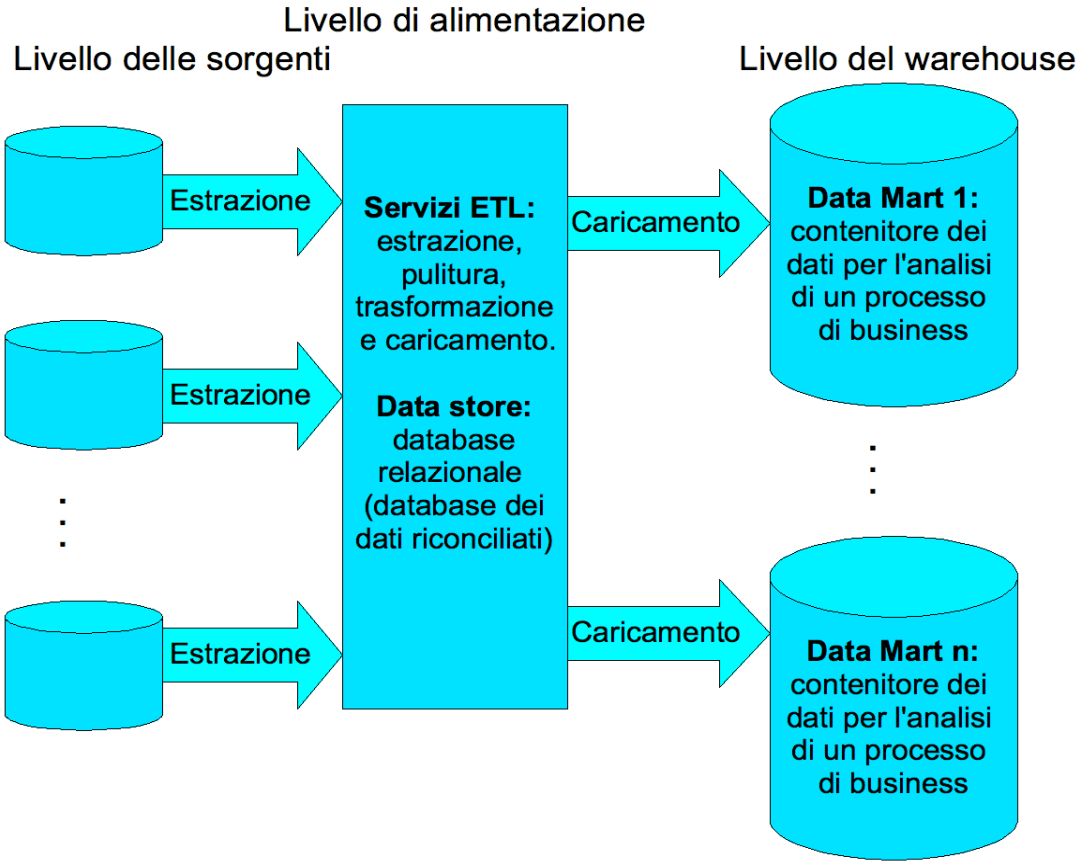
Componenti di un data warehouse (2)



Livello di alimentazione (Data staging area)

- Il livello di alimentazione fornisce i dati ottenuti al livello delle sorgenti al livello del warehouse.
- Contiene inoltre una serie di processi che servono per trasformare i dati, noti con il nome di *strumenti ETL* (*Extraction-Transformation-Load*).

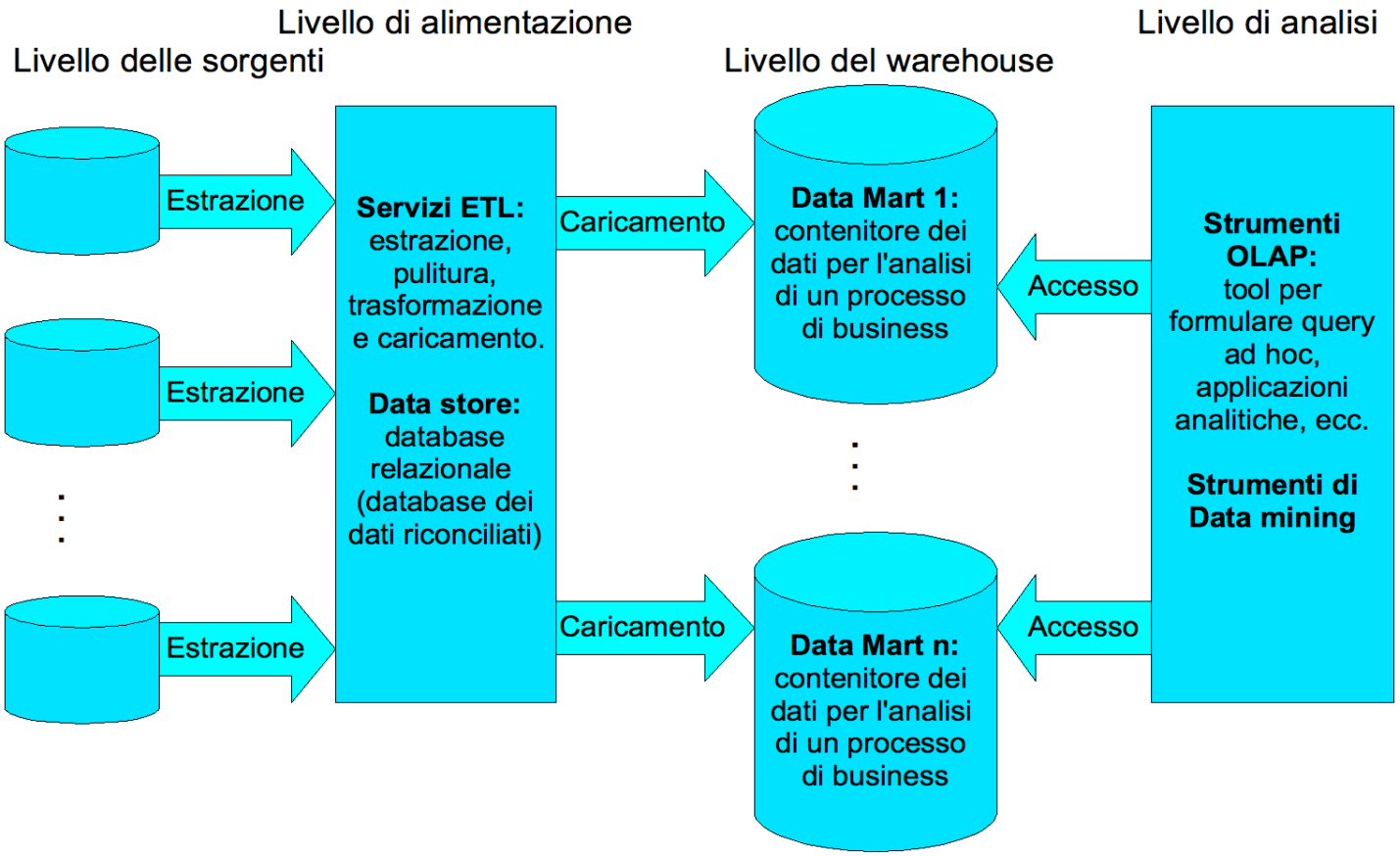
Componenti di un data warehouse (3)



Livello del warehouse (Data presentation area)

- Dal punto di vista dell'utente il livello del warehouse *è* il data warehouse, perché è l'unica area accessibile, in cui i dati sono disponibili per l'interrogazione.
- I dati sono presentati usando la metafora dell'ipercubo, come visto in precedenza.
- I dati sono memorizzati in *Data Mart*, ognuno dei quali contiene i dati relativi ad un processo aziendale; tipicamente i Data Mart condividono un insieme di dimensioni e fatti.
- Vengono memorizzati dati atomici; per ragioni prestazionali è possibile memorizzare anche (ma non solo) dati riassuntivi.

Componenti di un data warehouse (4)



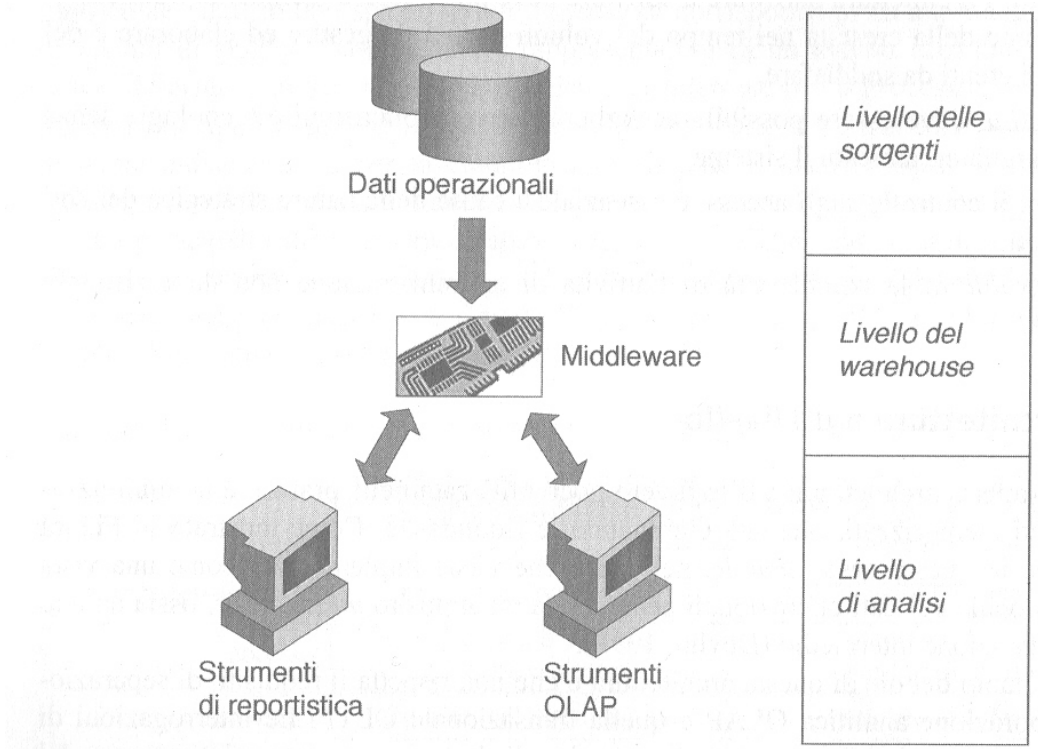
Livello di analisi (Data access tools)

- Il livello di analisi contiene:
  - per gli utenti più esperti, strumenti per la scrittura di interrogazioni ad hoc.
  - per gli utenti meno esperti, applicazioni di analisi predefinite, nelle quali è sufficiente definire una serie di parametri per ottenere il risultato.
- Esistono inoltre altri sofisticati strumenti di analisi (data mining), di cui parleremo in seguito.

Architetture del data warehouse

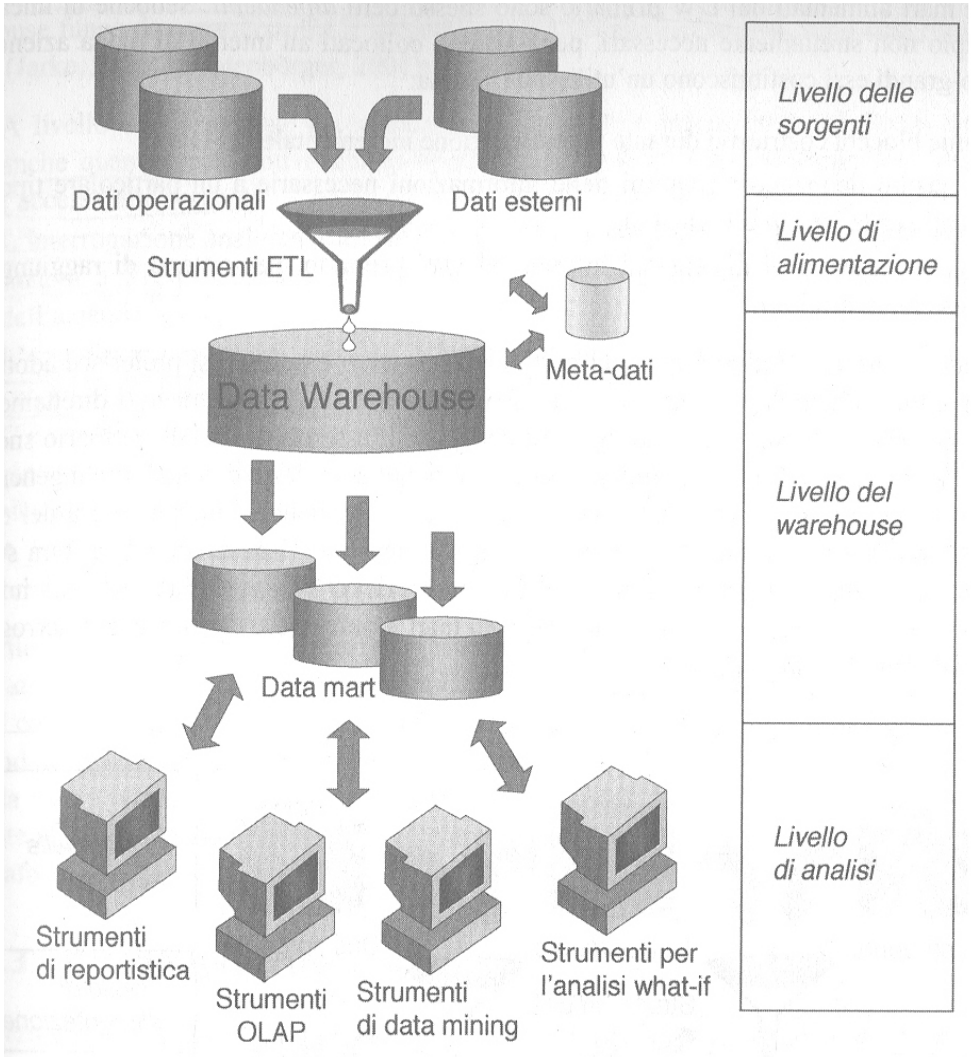
- Qualsiasi sistema di data warehousing deve contenere i quattro componenti analizzati.
- Sono però possibili varie scelte in ordine a come organizzare e coordinare tali componenti.
- Possiamo definire tre tipologie principali di architetture: a un livello, a due livelli, a tre livelli.

Architettura ad un livello



- Il DW è virtuale: implementato come una vista multidimensionale dei dati operativi, generata da un apposito middleware.
- Minimizza i dati memorizzati, eliminando le ridondanze, ma non rispetta il principio di separazione tra OLAP e OLTP.
- Elimina il livello di alimentazione.
- Poco utilizzato.

Architettura a due livelli (1)

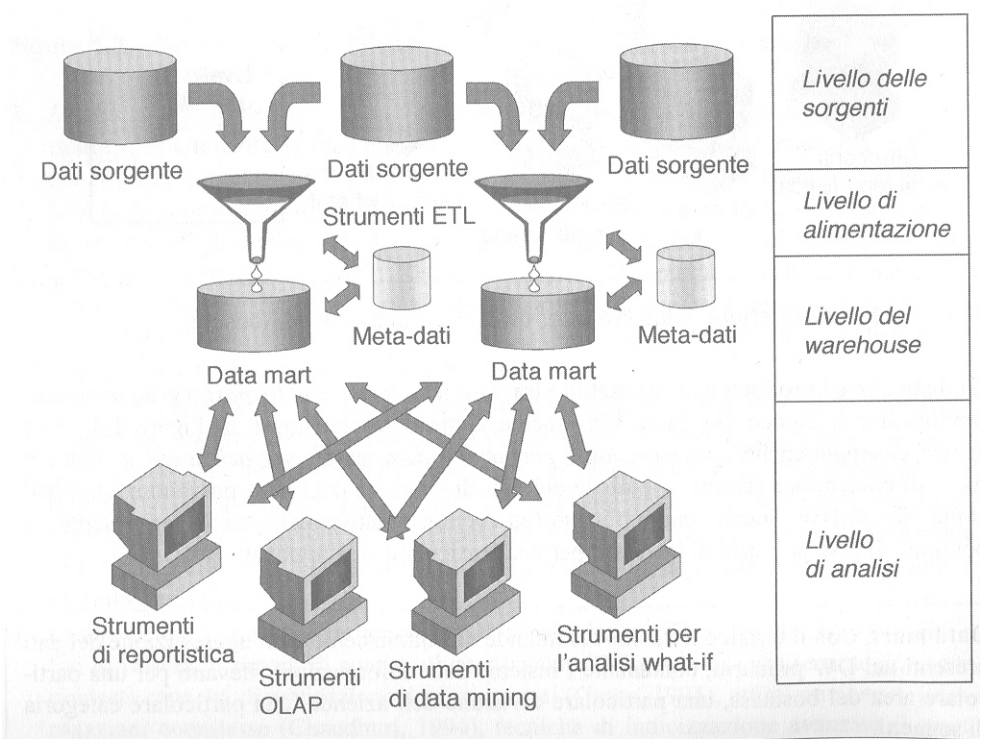


- La figura dovrebbe descrivere l'architettura di un Data Warehouse ma, allo stesso tempo, contiene un elemento chiamato Data Warehouse.
- Per quanto fuorviante manterremo questa notazione, il Data Warehouse è:
  - un sistema complesso;
  - e allo stesso tempo una base di dati facente parte di tale sistema complesso.

Architettura a due livelli (2)

- Implementa tutti i quattro componenti:
  - *Livello delle sorgenti*: dati interni (DB o sistemi legacy) e esterni all'azienda.
  - *Livello dell'alimentazione*: gli strumenti ETL (Extraction, Transformation and Loading) si occupano di ripulire i dati sorgente (eliminazione di informazione incompleta, incongruenze, ecc.), integrarli con dati provenienti da altre fonti, e caricarli nel DW.
  - *Livello del Warehouse*: può essere interrogato direttamente o utilizzato per costruire più Data Mart, realizzati come viste sul Data Warehouse.
  - *Livello di analisi*: consultazione dei dati per stesura di report, analisi, simulazioni.

Architettura a due livelli: Data Mart indipendenti

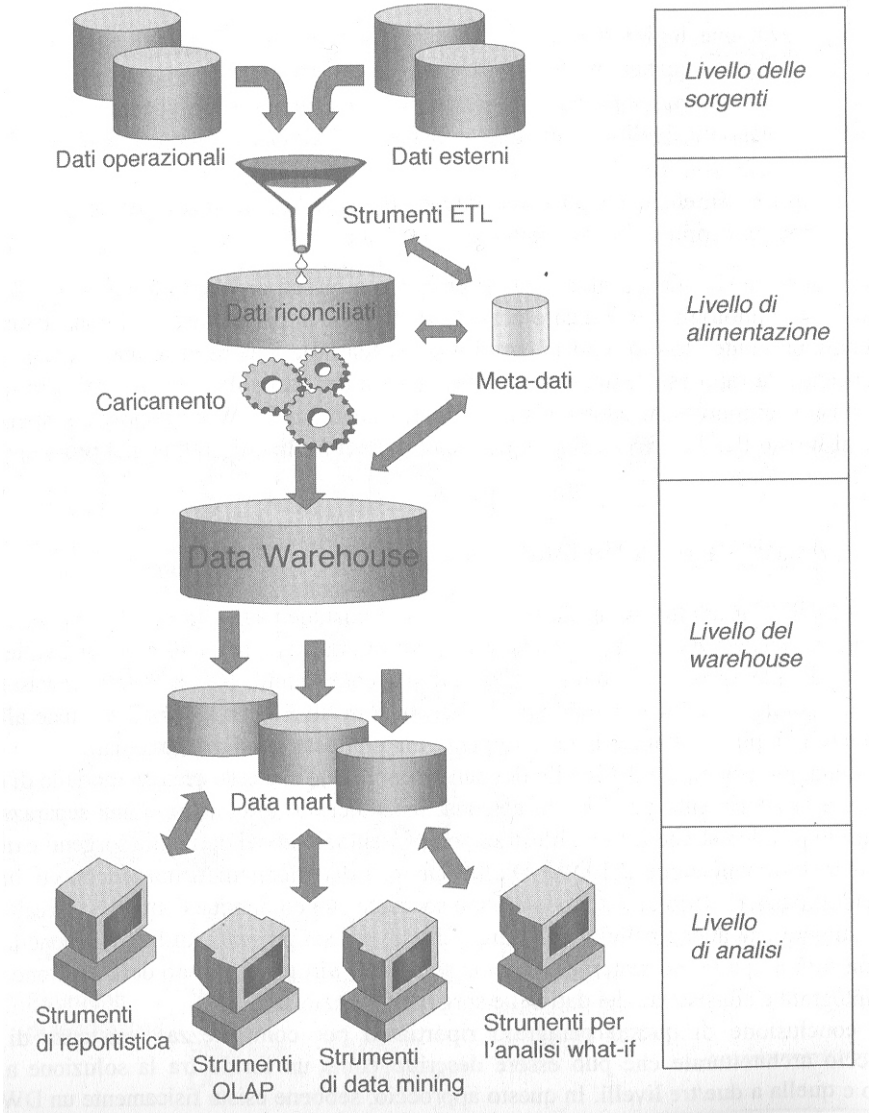


- I Data Mart vengono alimentati direttamente dalle sorgenti; vengono quindi detti *indipendenti*.
- L'assenza di un DW primario snellisce la fase di analisi, ma ingenera il rischio di incongruenze tra i Data Mart.
- A volte si preferisce quindi creare comunque un DW centrale, alimentato dai Data Mart.

Vantaggi dell'architettura a due livelli

- Mette sempre a disposizione dati di qualità, anche quando è temporaneamente precluso l'accesso alle sorgenti.
- Le interrogazioni sul DW non interferiscono con la gestione delle transazioni relative al normale funzionamento dell'azienda.
- Il DW è basato su un'organizzazione logica multidimensionale, mentre le sorgenti sono di solito relazionali o semi-strutturate.
- E' possibile applicare sul DW tecniche per l'ottimizzazione delle prestazioni, diverse da quelle utilizzate nell'ambito dei DBMS relazionali.

Architettura a tre livelli (1)



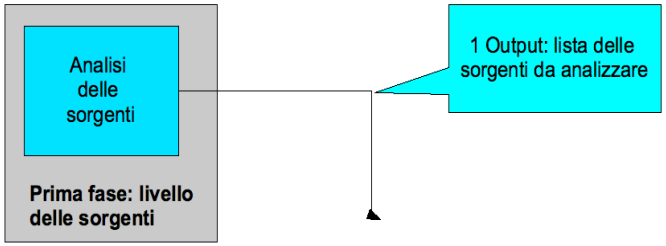
Architettura a tre livelli (2)

- Rispetto all'architettura a due livelli amplia il livello di alimentazione memorizzando i dati nel Database per i dati riconciliati.
- In pratica il risultato del processo di ripulitura e integrazione dei dati sorgenti viene materializzato; il DW viene dunque alimentato non più direttamente dalle sorgenti, ma dai dati riconciliati.
- Genera ulteriore ridondanza, ma consente una più netta separazione tra dati operazionali e DW.
- Soluzione architetturale preferibile alle altre due:
  - PRO: fault tolerance;
  - CON: costi dell'hardware per il database per i dati riconciliati.

Progettazione (1)

- Si individua una fase progettuale per la realizzazione di ogni componente del Data Warehouse:
  - Prima fase: si definiscono gli elementi del livello delle sorgenti;

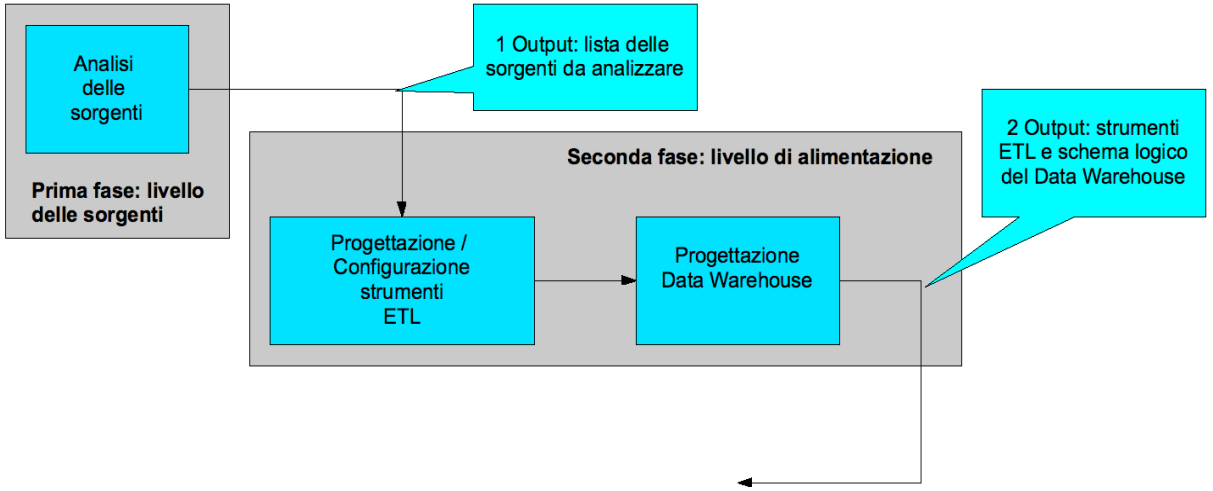
Diagramma di progettazione (1)



Progettazione (2)

- Si individua una fase progettuale per la realizzazione di ogni componente del Data Warehouse:
  - Prima fase: si definiscono gli elementi del livello delle sorgenti;
  - Seconda fase: viene definito il livello di alimentazione (strumenti ETL ed eventuale Data Warehouse);

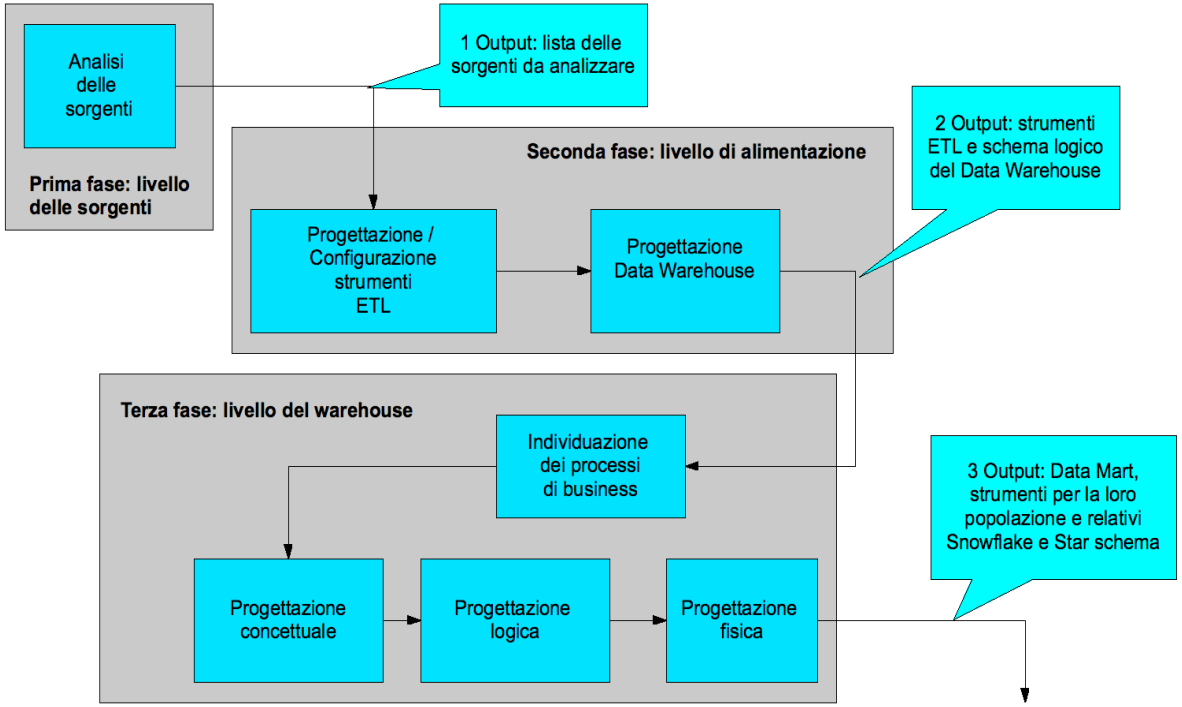
Diagramma di progettazione (2)



Progettazione (3)

- Si individua una fase progettuale per la realizzazione di ogni componente del Data Warehouse:
  - Prima fase: si definiscono gli elementi del livello delle sorgenti;
  - Seconda fase: viene definito il livello di alimentazione (strumenti ETL ed eventuale Data Warehouse);
  - Terza fase: viene definito il livello del warehouse (progettazione multidimensionale dei Data Mart);

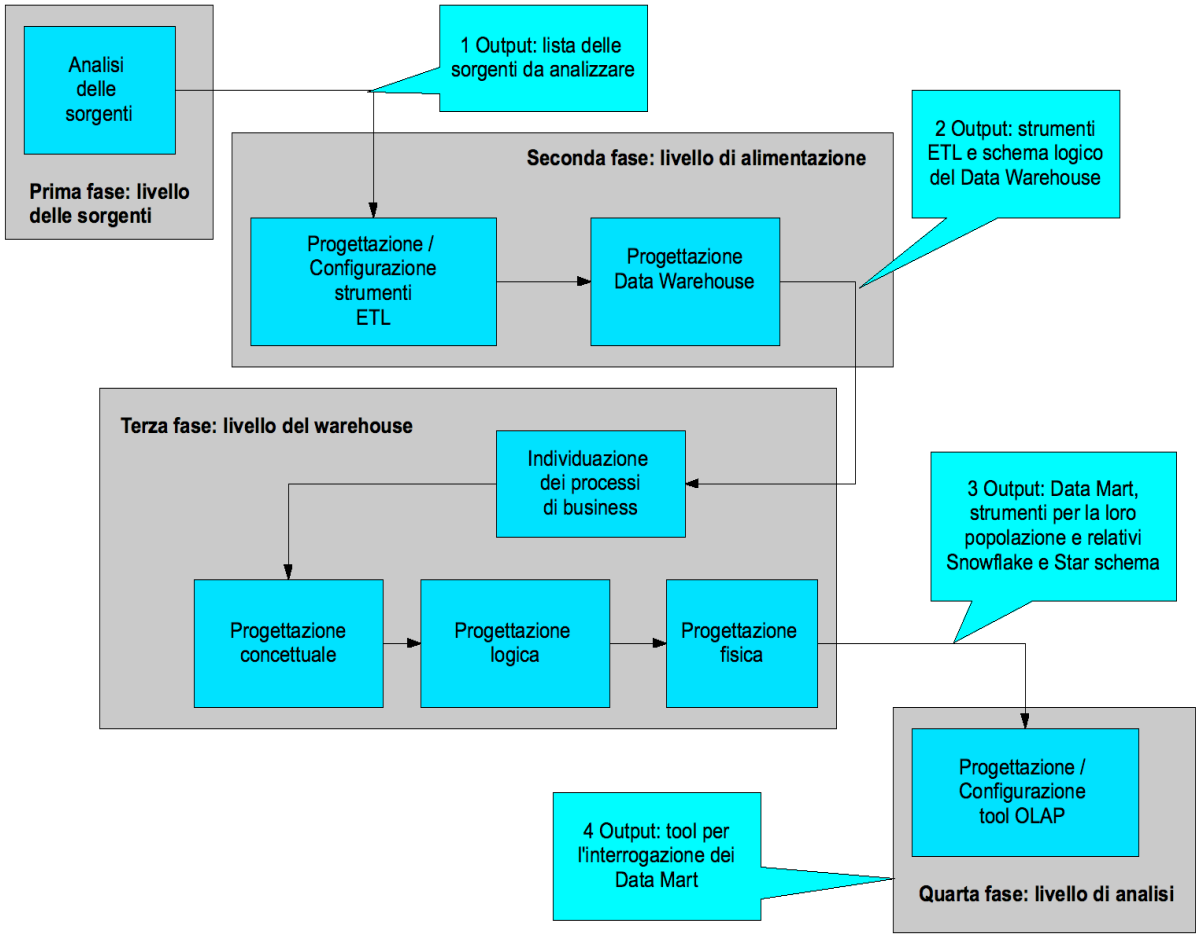
Diagramma di progettazione (3)



Progettazione (4)

- Si individua una fase progettuale per la realizzazione di ogni componente del Data Warehouse:
  - Prima fase: si definiscono gli elementi del livello delle sorgenti;
  - Seconda fase: viene definito il livello di alimentazione (strumenti ETL ed eventuale Data Warehouse);
  - Terza fase: viene definito il livello del warehouse (progettazione multidimensionale dei Data Mart);
  - Quarta fase: implementazione del livello di analisi (OLAP tools);

Diagramma di progettazione (4)



Analisi delle sorgenti

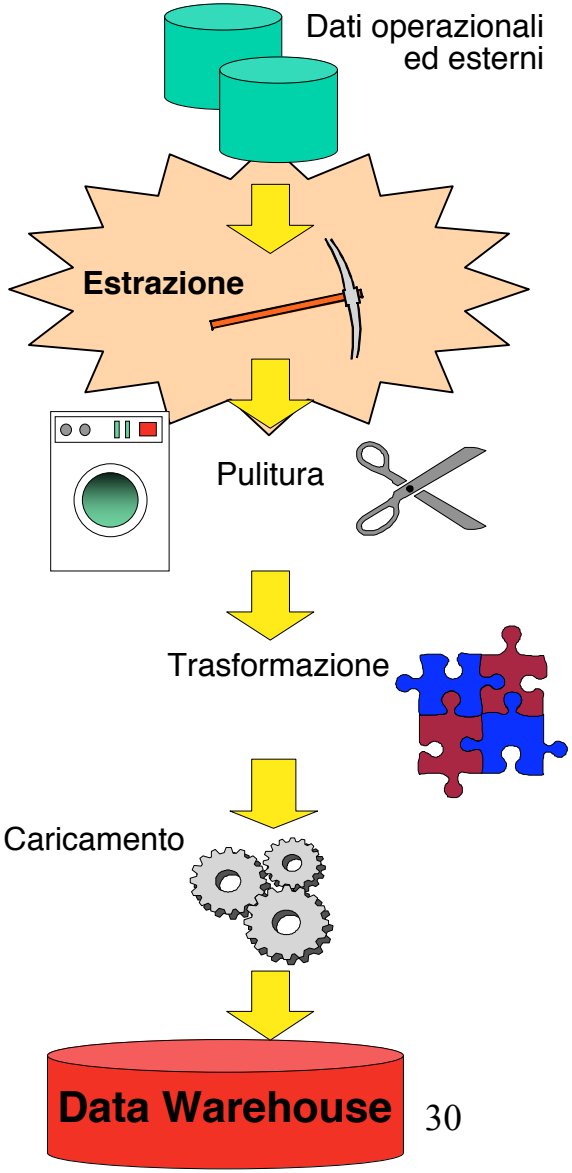
- La prima fase progettuale è l'analisi delle sorgenti.
- Si stila un elenco completo e dettagliato delle sorgenti che alimentano il Data warehouse.
  - Databases
  - Fogli elettronici
  - Documenti XML
- Si riporta la struttura (anche parziale) di tali documenti, cosi da sapere come maneggiarli, che informazioni contengono e come sono strutturate.
- Prepara alla seconda fase : la progettazione degli strumenti ETL.

Strumenti ETL

- Il ruolo degli strumenti di **Extraction, Transformation and Loading** è quello di alimentare una sorgente dati singola, dettagliata, esauriente e di alta qualità che possa a sua volta alimentare il DW (riconciliazione)
- Durante il processo di alimentazione del DW, la riconciliazione avviene in due occasioni: quando il DW viene popolato per la prima volta, e periodicamente quando il DW viene aggiornato.
  - estrazione
  - pulitura
  - trasformazione
  - caricamento
- Nei sistemi commerciali vengono forniti e sono da configurare per il lavoro che devono svolgere.

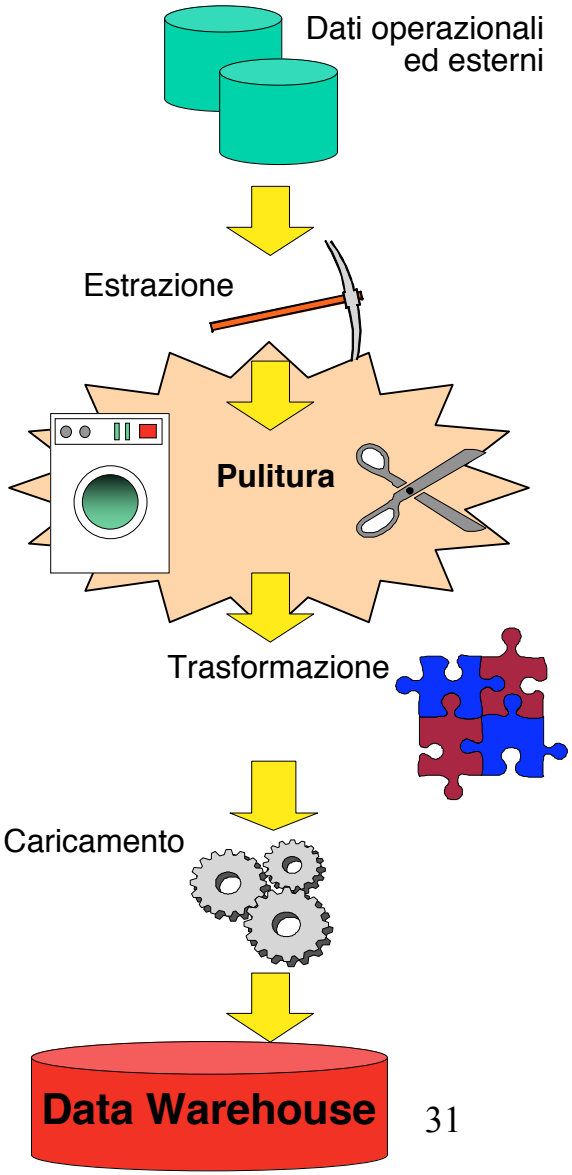
Estrazione

- I dati rilevanti vengono estratti dalle sorgenti.
  - L'estrazione **statica** viene effettuata quando il DW deve essere popolato per la prima volta e consiste concettualmente in una fotografia dei dati operazionali.
  - L'estrazione **incrementale** viene usata per l'aggiornamento periodico del DW, e cattura solamente i cambiamenti avvenuti nelle sorgenti dall'ultima estrazione
    - basata sul log mantenuto dal DBMS operazionale
    - basata su time-stamp
    - guidata dalle sorgenti
- La scelta dei dati da estrarre avviene principalmente in base alla loro qualità.



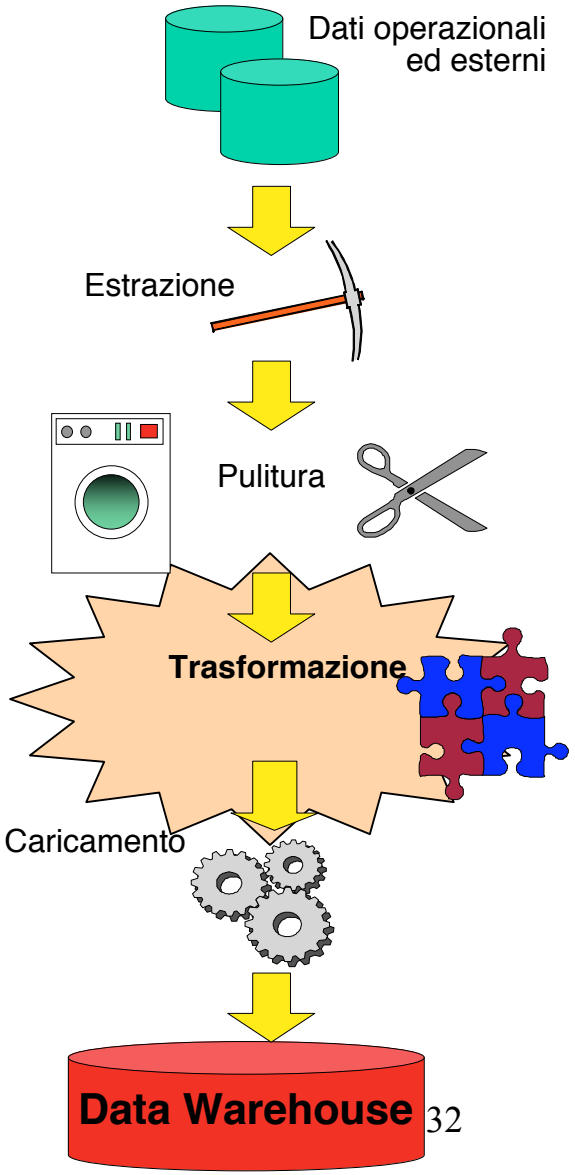
Pulitura

- Si incarica di migliorare la qualità dei dati delle sorgenti
  - dati duplicati
  - inconsistenza tra valori logicamente associati
  - dati mancanti
  - uso non previsto di un campo
  - valori impossibili o errati
  - valori inconsistenti per la stessa entità dovuti a differenti convenzioni
  - valori inconsistenti per la stessa entità dovuti a errori di battitura



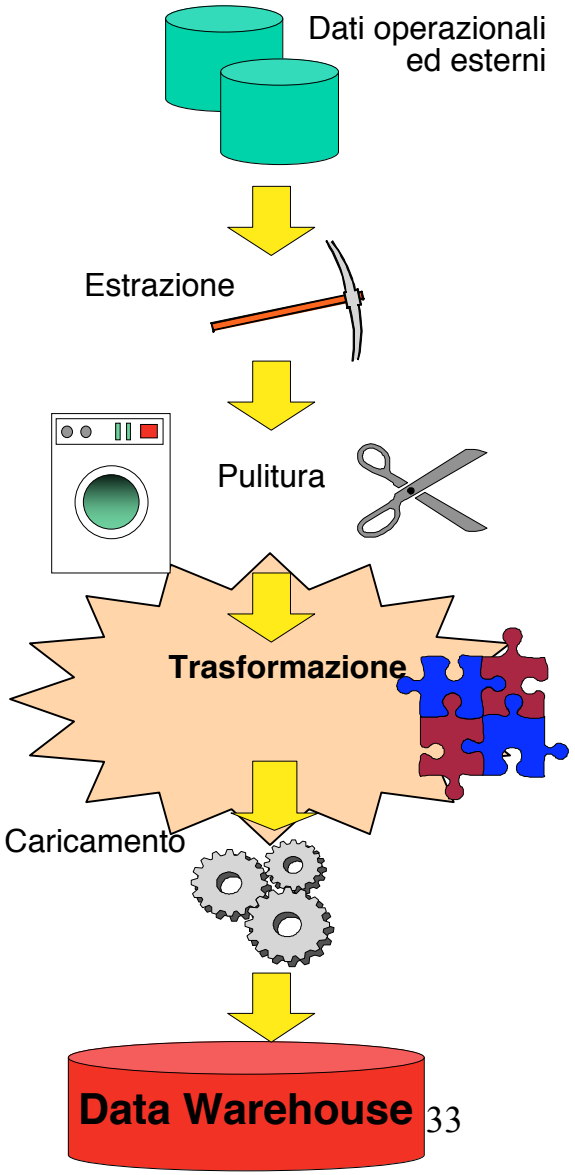
Trasformazione (1)

- Converte i dati dal formato operazionale sorgente a quello del DW. La corrispondenza con il livello sorgente è complicata dalla presenza di fonti distinte eterogenee, che richiede una complessa fase di integrazione
  - presenza di **testi liberi** che nascondono informazioni importanti
  - utilizzo di **formati differenti** per lo stesso dato

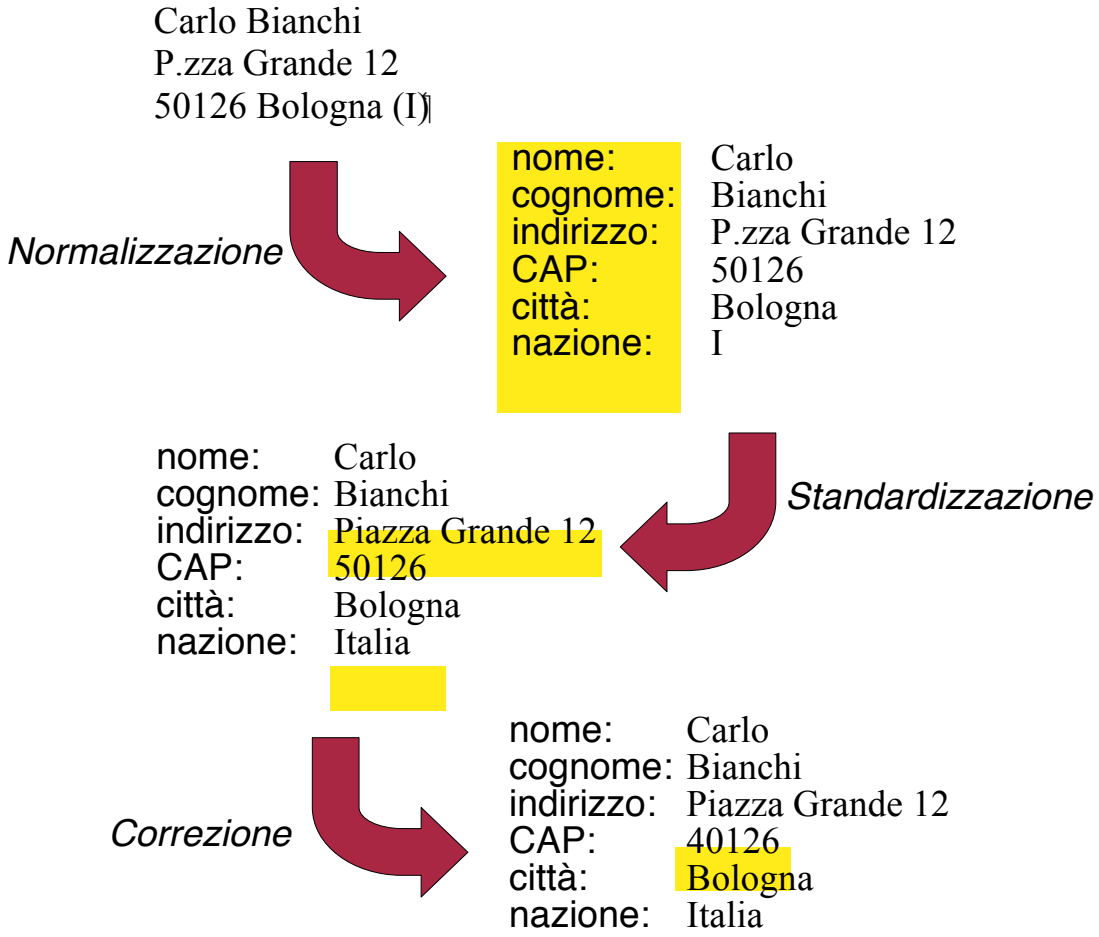


Trasformazione (2)

- Per l'alimentazione dei dati riconciliati:
  - **Conversione e normalizzazione.** Operano a livello di formato di memorizzazione e di unità di misura per uniformare i dati.
  - **Matching.** Stabilisce corrispondenze tra campi equivalenti in sorgenti diverse.
  - **Selezione.** Riduce il numero di campi e di record rispetto alle sorgenti.

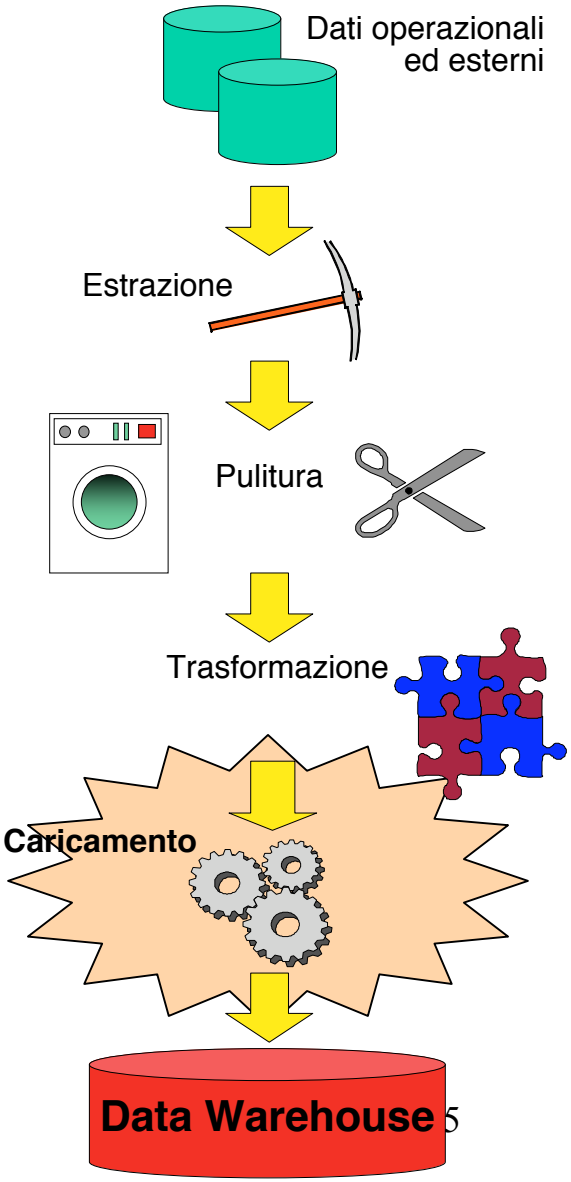


Esempio di pulitura e trasformazione



Caricamento

- Il caricamento dei dati nel DW
  - **Refresh.** I dati del DW vengono riscritti integralmente, sostituendo quelli precedenti (tecnica normalmente utilizzata solo per popolare inizialmente il DW)
  - **Update.** I soli cambiamenti occorsi nei dati sorgente vengono aggiunti nel DW (tecnica normalmente utilizzata per l'aggiornamento periodico del DW)



Software ETL

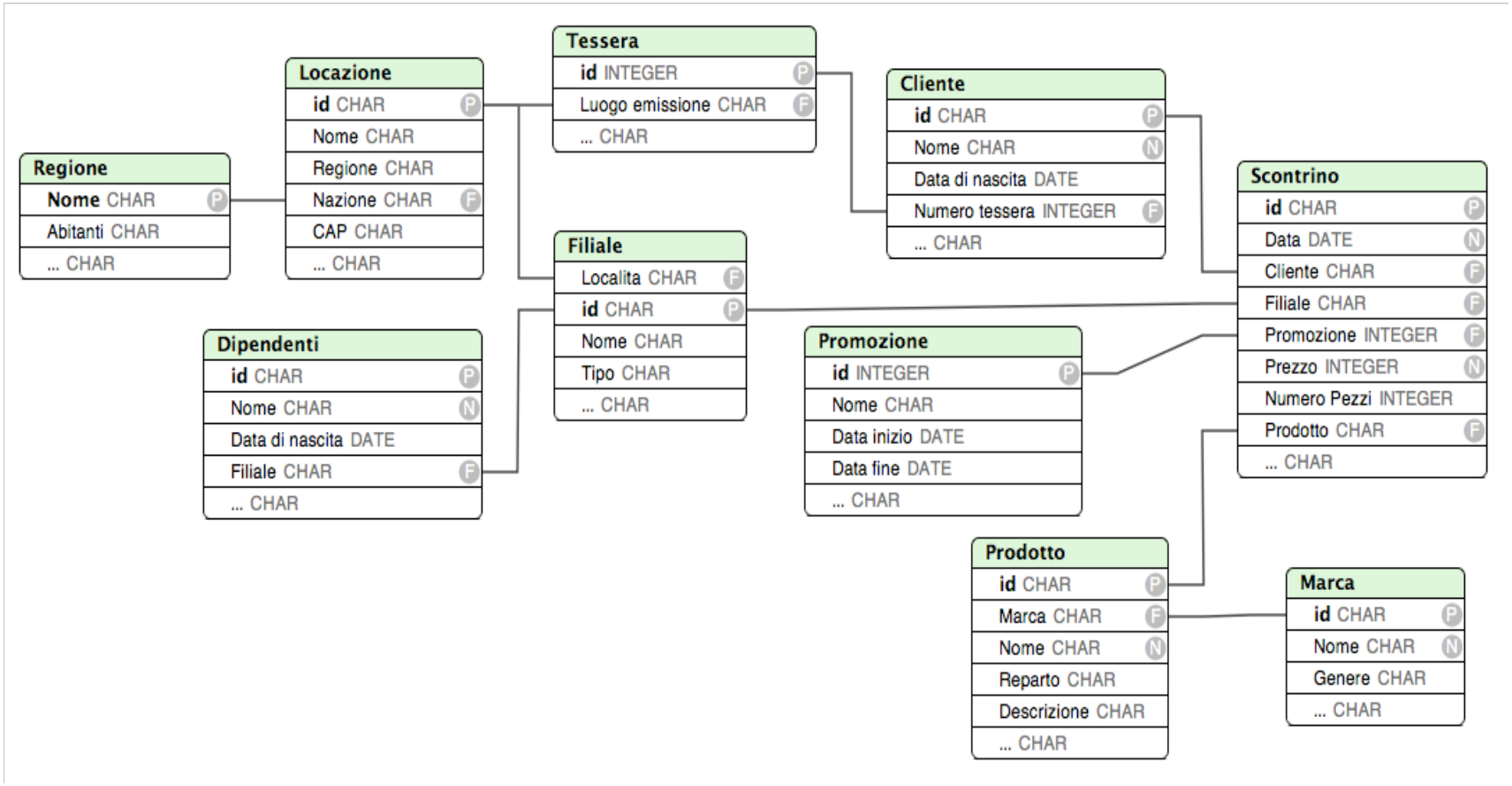
Alcuni software ETL commerciali:

- Microsoft DTS: Data Transformation Services, in SQL Server 2000 di *Microsoft.*
- Microsoft SSIS: SQL Server Integration Services, in SQL Server 2005 di *Microsoft.*
- PowerCenter, PowerExchange, Data Explorer, e Data Quality di *Informatica* (strumenti stand alone).
- Oracle Warehouse Builder di *Oracle.*
- Data Integrator di *BusinessObjects.*

Progettazione del Data Warehouse

- Il Data Warehouse di cui si parla qui, altro non è che un database che contiene tutti i dati che possono alimentare i Data Mart (Livello del warehouse).
- Nella architettura a tre livelli viene chiamato database dei dati riconciliati.
- Viene progettato in terza forma normale come i comuni database relazionali.
- Deve essere dotato di una tabella dove immagazzinare i dati temporali:
  - giorno, mese, anno,
  - ora, minuti, secondi,
  - trimestre, quadrimestre, semestre,
  - giorni festivi, ecc.

Esempio di un Data Warehouse



Bibliografia

[1] M. Golfarelli e S. Rizzi, *Data Warehouse – Teoria e Pratica della Progettazione,* Cap. 1. McGraw-Hill 2006.

[2] R. Kimball, *The Data Warehouse Toolkit: the Complete Guide to Dimensional Modeling (2nd Edition)*, Cap 1. Wiley, 2002.