

Data warehousing

Data warehousing e OLAP

OLAP (On Line Analytic Processing) identifica l'insieme degli strumenti atti ad aiutare il processo decisionale all'interno di un'azienda.

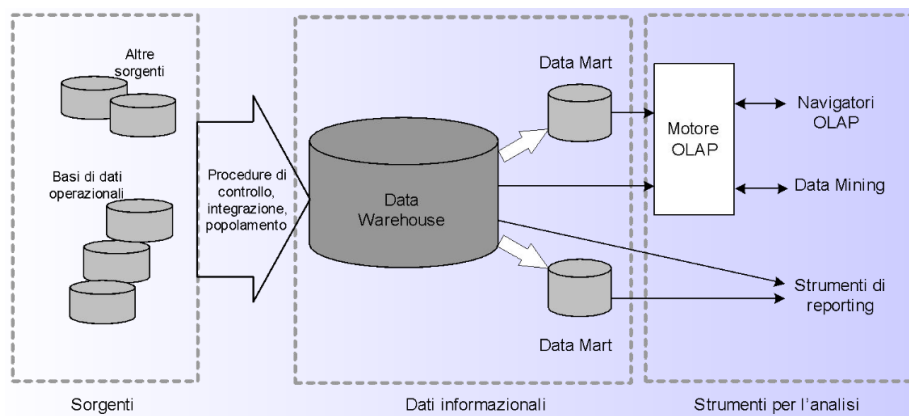
Esistono alcune regole per la definizione dei sistemi OLAP, la cosiddetta *FASMI*:

- Fast
- Analytical
- Shared
- Multidimensional
- Informatic

Architettura dei sistemi di data warehousing

I sistemi di DWH sono costituiti da DB posti a diversi livelli.

1. **Sorgenti:** db di origine dei dati, possono essere esterni o operazionali
 1. **Staging Area (opzionale):** area intermedia usata per la trasformazione dei dati.
 2. **Data warehouse:** db centrale, contiene tutti i dati necessari per le analisi.
 3. **Data mart:** db multidimensionali su cui si appoggia l'analisi.



Esistono architetture con un numero variabile di livelli, quelle a 2 livelli non comprendono la staging area mentre le architetture a 3 livelli sì.

Le soluzioni a 3 livelli sono spesso usate aziende più complesse, i sistemi a 2 livelli presentano elementi come:

- Un primo livello costituito dalle sorgenti dei dati.
- Il secondo livello contenente i dati informativi quindi dal DWH in poi.

Modelli concettuali per il data warehousing

i Sistemi informativi sono soggetti a molte evoluzioni nel corso della loro vita. Solitamente un'azienda costruisce un nucleo contenente i dati di maggior interesse, i quali verranno poi ampliati.

Non esiste un metodo evolutivo standard, vedremo solo DFM.

DFM (Dimensional Fact Model)

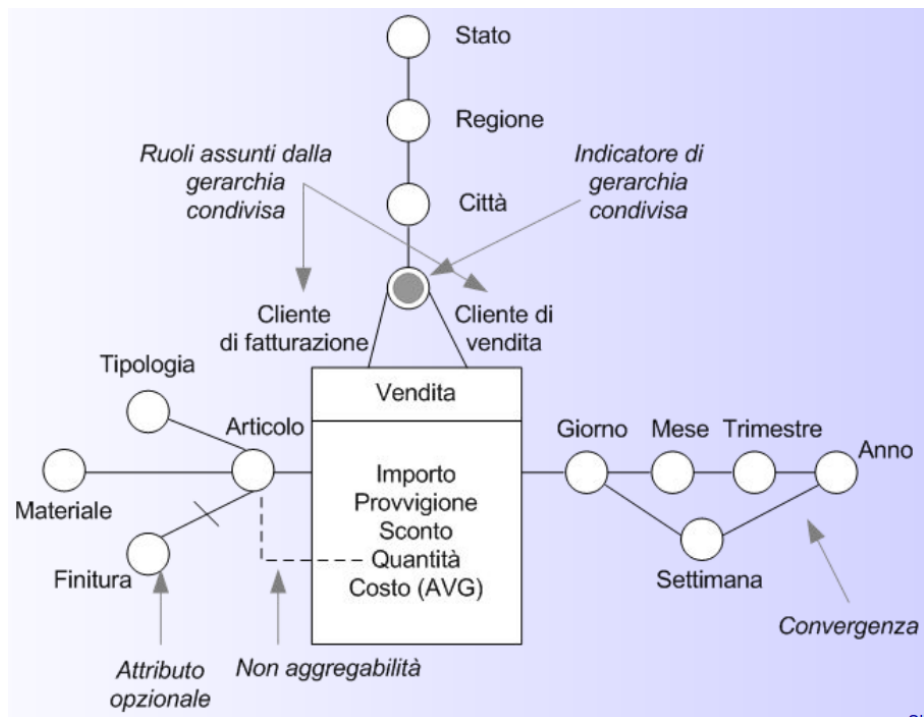
Fornisce una visione ad alto livello descrivendo graficamente i fatti attorno a cui si struttura un warehouse.

Ogni fatto è rappresentato tramite uno schema di fatto:

- **Fatto:** rettangolo contenente il nome del fatto e le sue misure
- **Dimensioni:** circoletti etichettati, vengono ollegati ai vari fatti

Le gerarchie dimensionali sono alberi con radice nelle dimensioni di base, mentre i nodi sono gli attributi su cui la gerarchia è costruita. DFM permette di rappresentare alcune caratteristiche dei fatti:

- L'opzionalità di una o più dimensioni
- La presenza di gerarchie
- La convergenza
- Non aggregabilità



Modelli logici per il data warehouse

Nel momento in cui bisogna realizzare un warehouse si deve scegliere quale DBMS usare.

I dati possono essere memorizzati in db relazionali oppure in db multidimensionali come gli ipercubi.

Dobbiamo scegliere anche il tipo di interrogazione da fare:

- Motori di db relazionali come SQL.
- Motori multidimensionali tramite linguaggi come MDX di Microsoft.
- Elaborazione delegata ai client tramite linguaggi proprietari.

Dalla combinazione delle caratteristiche sopra citate nascono tre tipo di modelli:

1. Relational OLAP (ROLAP)
2. Multidimensional OLAP (MOLAP)
3. Hybrid OLAP (HOLAP)

ROLAP

Si basa su una struttura a db puramente relazionali interrogati tramite query SQL.

Risultano quindi molto compatte e con un diffuso know-how, bisogna però anche considerare la ridotta velocità per query con molte dimensioni e che le soluzioni (denormalizzazione e materializzazione) fanno aumentare la complessità di gestione e le dimensioni.

MOLAP

Con questo approccio i dati sono memorizzati come strutture multidimensionali, basta pensare a dei vettori.

Questa soluzione non ha molta popolarità dato il tasso di spazio occupato in cui solo il 20% è spazio utile, la mancanza di standard e il grande successo dei db relazionali.

HOLAP

Soluzione intermedia che combina i vantaggi delle presenti.

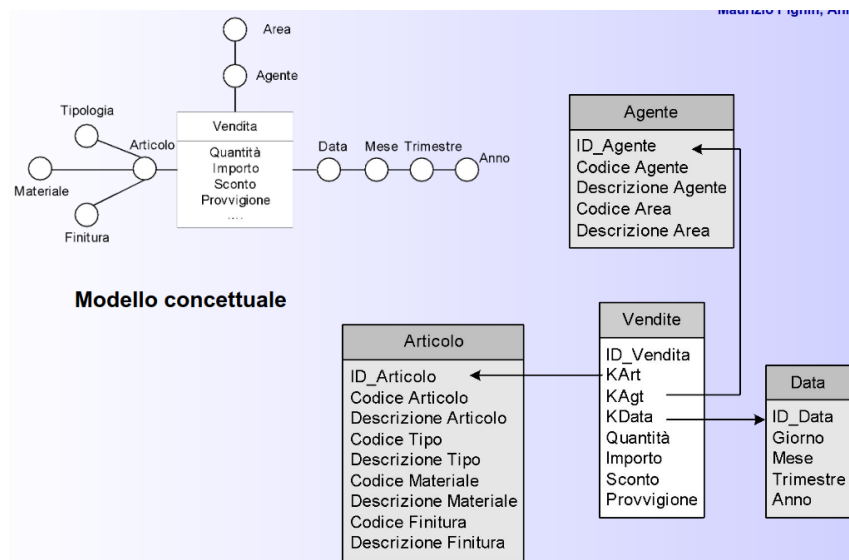
Il warehouse viene realizzato con un db relazionale così da essere mantenibile e scalabile facilmente.

Viene poi fatta una distinzione nei data mart, in cui i dati sono realizzati con db multidimensionali per avere una maggior efficienza nelle query e con un overhead dimensionale minore.

Schemi multidimensionali su db relazionali

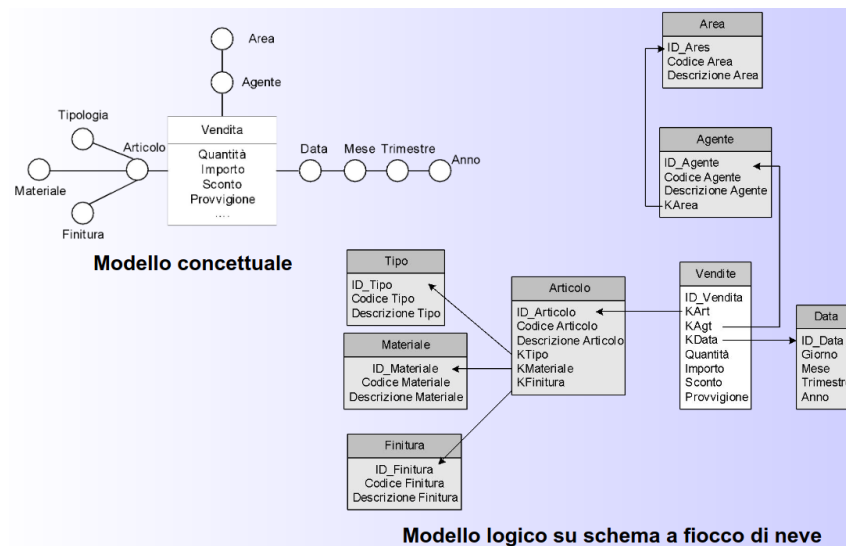
1. Schema a stella

Viene usata una tabella dei fatti in cui ogni elemento è un fatto elementare, per ogni misura propria del fatto viene inserito un campo di tipo numerico. Vengono anche definite le tabelle delle dimensioni per ogni dimensione di base, queste tabelle sono soggette ad una denormalizzazione completa. L'elevata denormalizzazione permette di fare un unico join per avere tutti i dati relativi ad un'unica dimensione, questo massimizza la velocità. La denormalizzazione porta anche molti svantaggi come la scarsa intuitività e lo spazio occupato da gerarchie profonde.



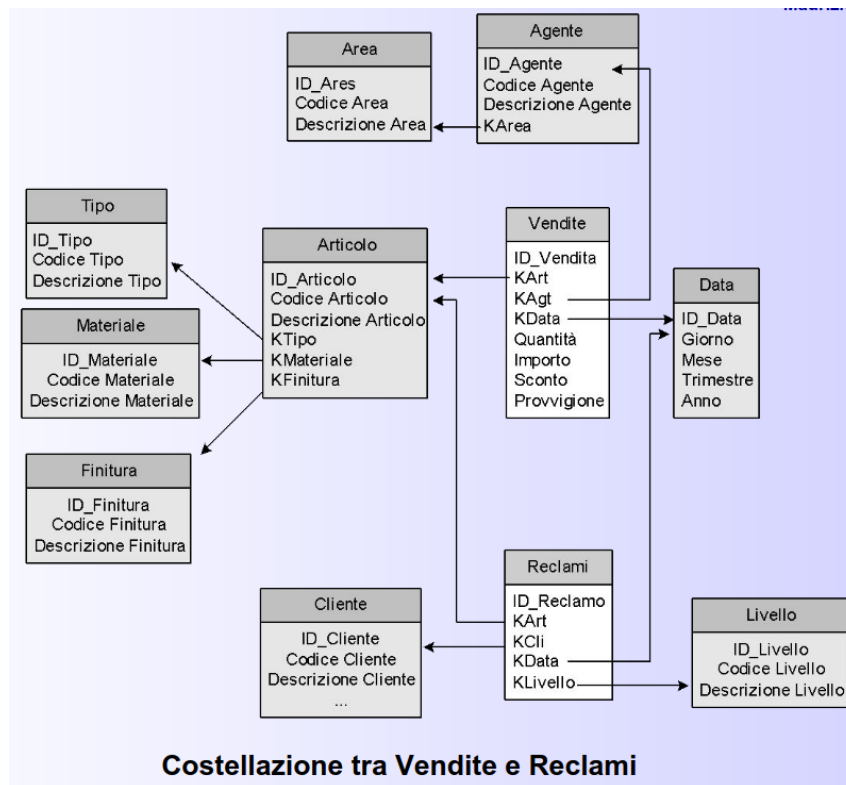
Questo schema riduce la denormalizzazione esplicitando delle dipendenze funzionali.

Ne risente però la velocità di risposta alle richieste.



3. Schema a costellazione

Se diverse tabelle dei fatti condividono delle tabelle dimensionali, risulta essere il miglior approccio da seguire quando più fatti coinvolgono gli stessi soggetti.



Ciclo di vita del DWH

La costruzione di un data warehouse è un processo che avviene, solitamente, in modalità iterativa.

Viene prima definito e popolato un ipercubo principale e man mano vengono aggiunti gli altri fatti, una volta rilasciati tutti i fatti di uno specifico interesse aziendale è possibile rilasciare il corrispettivo data mart.

Vantaggi:

- Premi risultati disponibili subito
- Investimenti obbligatoriamente diluiti
- Sviluppare il modello in base alle necessità

Costruzione dei data mart

E' costituita dai seguenti passaggi:

- **Analisi delle sorgenti:** capire quali dati sono disponibili e verificare che siano compatibili con i requisiti lato utente.
- **Progettazione concettuale degli schemi:** identificare misure, dimensioni ed eventuali limiti di aggregabilità.

- **Progettazione logica:** decisione su schemi a stella/fiocco di neve e la necessità di costruire viste materializzate o ipercubi con molta aggregazione.
- **Alimentazione:** le procedure che straggono i dati dalle sorgenti e li processano per prepararli al DWH.