Data warehousing

Data warehousing e OLAP

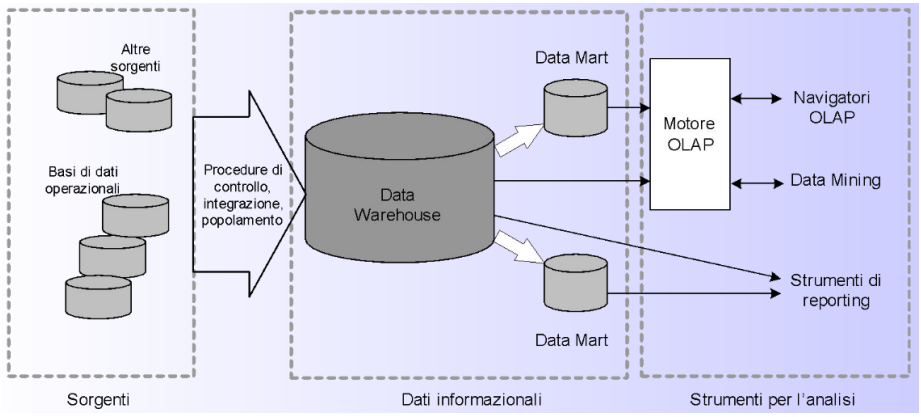
OLAP (On Line Analytic Processing) identifica l'insieme degli strumenti atti ad aiutare il processo decisionale all'interno di un'azienda.
Esistono alcune regole per la definizione dei sistemi OLAP, la cosiddetta *FASMI*:

- 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 si.
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 informazionali quindi dal DWH in poi.

Modelli concettuali per il data warehousing

i Sistemi informazionali 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.

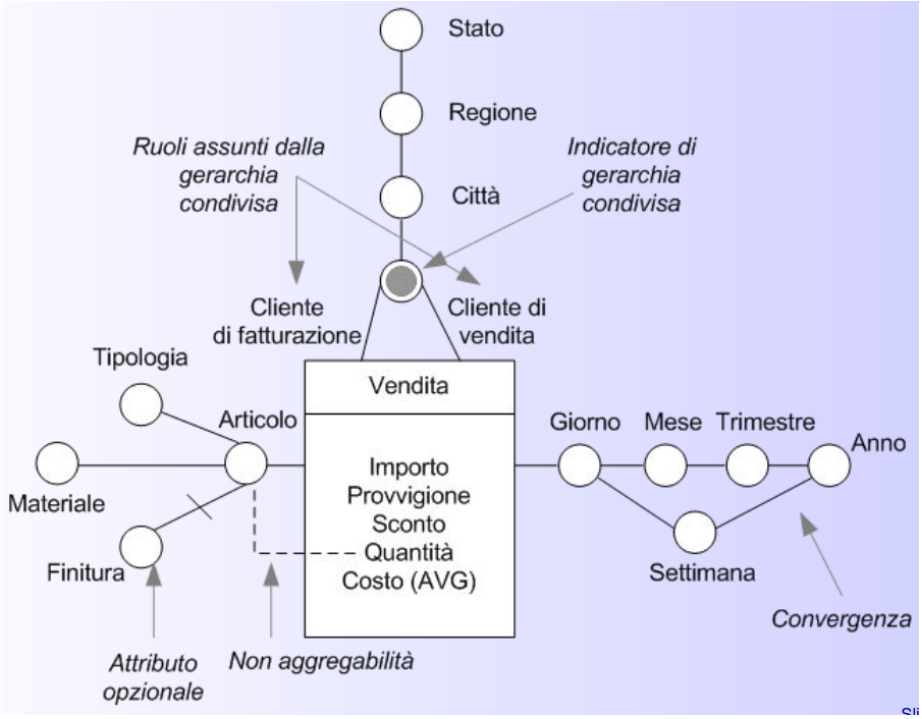
DFM (Dimensional Fact Model)

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 dimansioni 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 agreggabilità



Modelli logici per il data warehouse

Nel momento in cui bisogna realizare 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 quetso 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.

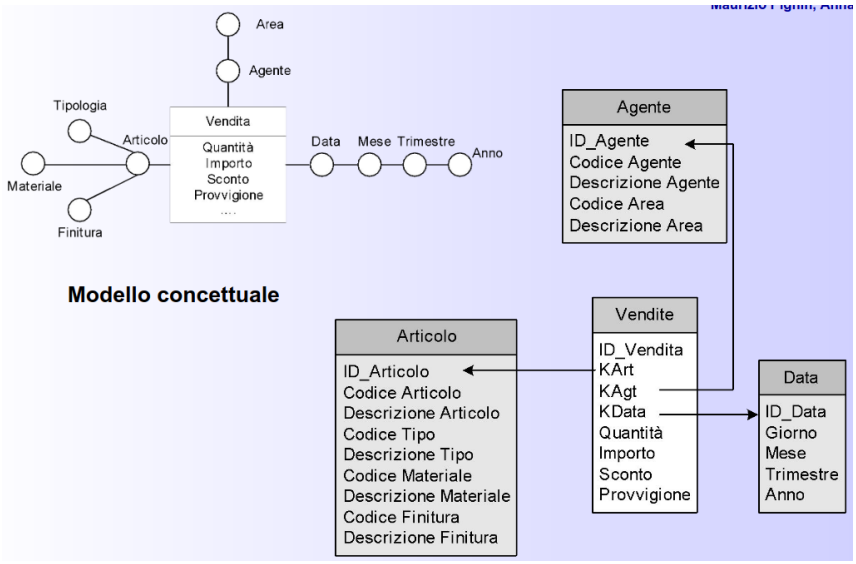
HOLAP

Soluzione intermedia che combina i bantaggi 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 efficenza nelle query e con un overhead dimensionale minore.

Schemi multidimensionali su db relazionali

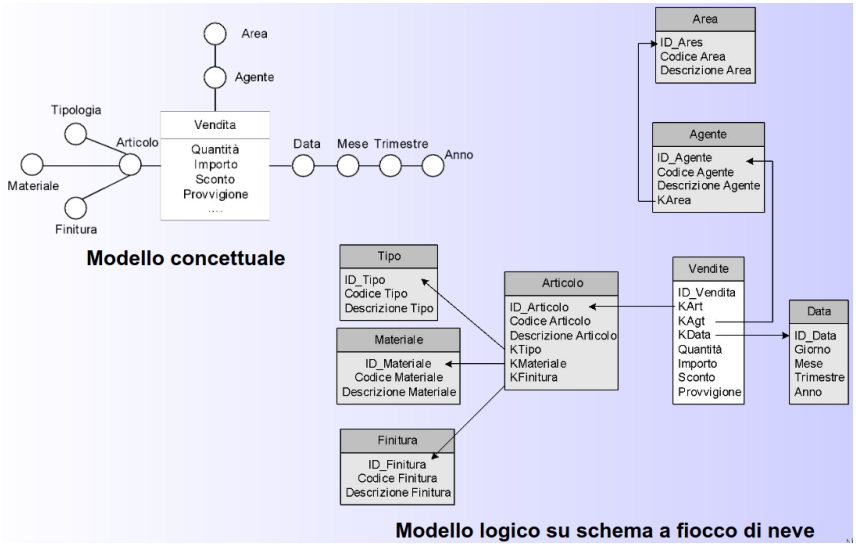
1. Schema a stella

Nelle soluzuioni ROLAP e HOLAP la modellazione logica segue lo schema a stella e le sue varianti.
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 ognni dimensione di base, queste tabelle sono soggetete 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.



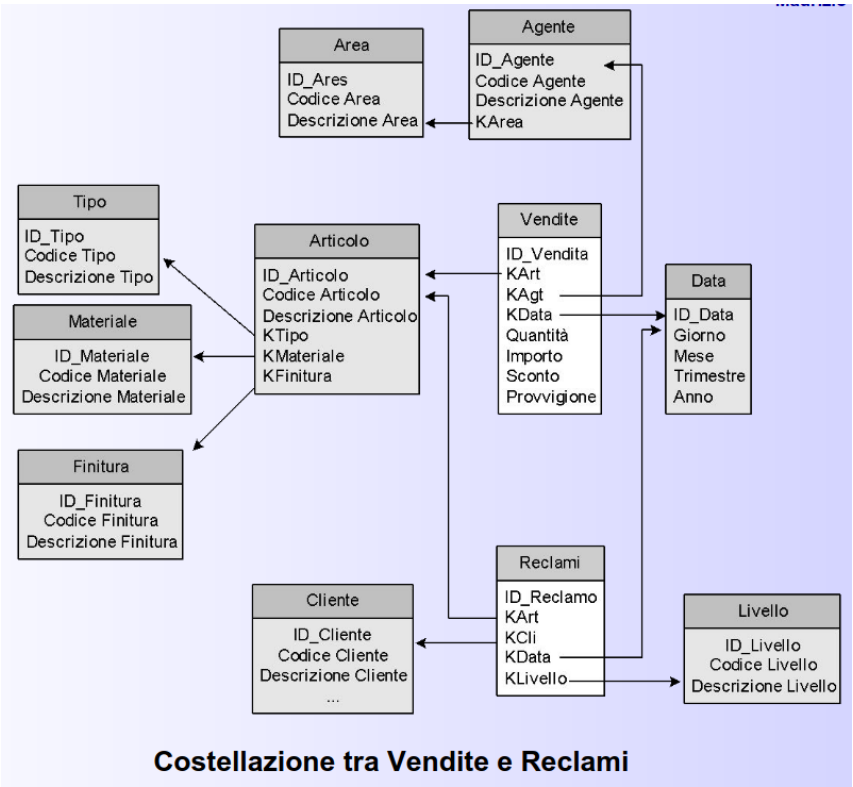
2. Schema a fiocco di neve

   Questo schema riduce la denormalizzazione esplicitando delle dipendenze funzionali.
   Questo permette di chiarificare la separazione tra i soggetti, migliora le prestazioni e riduce la sensibilità alle variazioni logiche.
   Ne risente però la velocita 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.



Costellazione tra Vendite e Reclami

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
- Invetsimenti obbligatoriamente diluiti
- Sviluppare il modello in base alle necessità

Costruzione dei data mart

E' costituita dai seguenti passaggi:

- **Analisi delle dorgenti:** 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.

Popolamento del data warehouse

Fasi di popolamento

1. Estrazione dei dati

   I dati vengono estratti dalle diverse fonti, viene definito quali dati devono essere acquisiti (tabelle e campi) e anche come devono essere trattati gli eventi di origine (aggregaizone o massimo dettaglio).
   Esistono due tipi di estrazione, la prima è tipicamente usata per la popolazione iniziale:

   (a) **Statica:** vengono prelevati tutti i dati presenti nella sorgente.
   (b) **Dinamica:** vengono estratti solo i dati modificati o prodotti alla fonte.
      - L'estrazione avviene in modo automatico, guidata da apposite funzioni trigger.
      - L'estrazione avviene in modo periodico, tipica delle PMI.
      - Viene fatto un confronto diretto con la sorgente.

   Spesso i dati estratti non vengono subito scritti nel DWH ma messi nella staging area per essere modificati.

2. Integrazione e trasformazione

   Prima di essere scritti i dati devono essere resi omogenei.

   - **Riconciliazione:** i dati che riguardano lo stesso soggetto ma provengono da fonti diverse sono messi in relazione.
   - **Riconoscimento dei duplicati**.
   - **Trasformazione dei valori continui:** vengono parametrizzati in valori discreti.
   - **Standardizzazione dei formati**.

3. Pulizia

   Questa fase può essere fatta anche prima dell'integrazione ma anche in modo parallelo ad essa.
   Devono essere eseguite analisi per rilevare, possibilmente anche riparare, le incorrette presenti nei dati.

   - **Dati incompleti:** si può inserire un codice per indicare la mancanza del dato, oppure nei sistemi più complessi si fanno inferenze sugli altri dati per ricavare quello mancante.
   - **Dati errati o incomprensibili:** vengono confrontati con dizionari di dati o valori limite, per stabilire la loro ammissibilità.

- **Dati inconsistenti:** vengono applicate delle regole per inferire quale risultato sia corretto.

4. Caricamento dei dati

   Questa è la fase in cui i dati vengono effettivamente caricati nel DWH, l'aggiornamento avviene dalle dimensioni verso i fatti con l'applicazione delle politiche di aggiornamento agli elementi già esistenti.

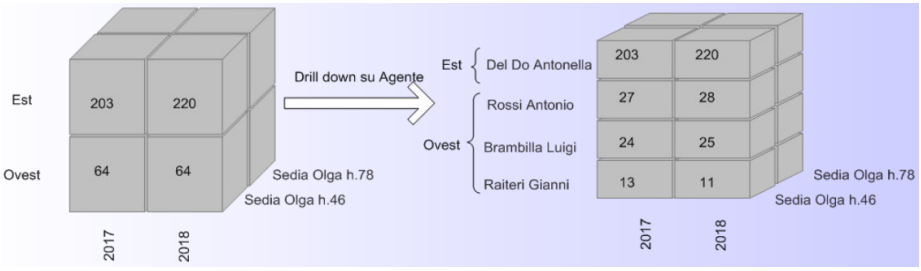
   - Non fare nulla, ogni fatto usa gli attributi dimensionali validi all'inserimento della dimansione.
   - Aggiornare l'elemento sovrascrivendolo.
   - Creare una nuova istanza che verrà associata ai nuovi fatti che verranno inseriti.
   - Creare una nuova istanza con marcatori temporali.

Tecniche di analisi dei dati

L'analisi OLAP è la principale modalità di interrogazione interattiva del warehouse.
Il paradigma che si usa è quello dell'esplorazione guidata da ipotesi, l'utente formula un'ipotesi e inoltra la richiesta per verificarla.
Durante le interrogazioni viene costruita una sessione di analisi, ciascun passo diventa conseguenza dei risultati precedentemente ottenuti.
Ogni passaggio di navigazione è costituito da un operatore OLAP, se non diversamente detto viene applicato all'ultimo risultato ottenuto.

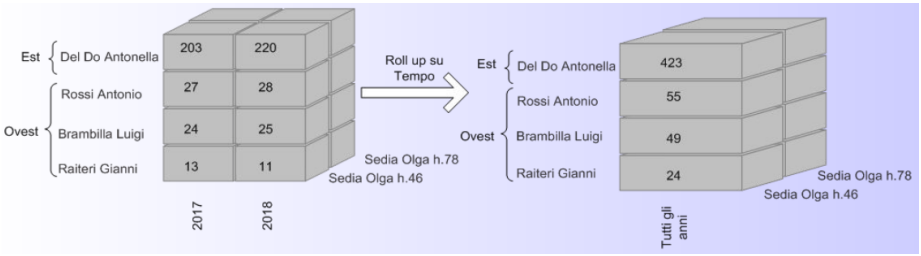
Drill down

Permette di partire da un livello generale e approfondire i dettagli passo passo, si scende lungo una gerarchia o aggiungendo una dimensione di analisi.



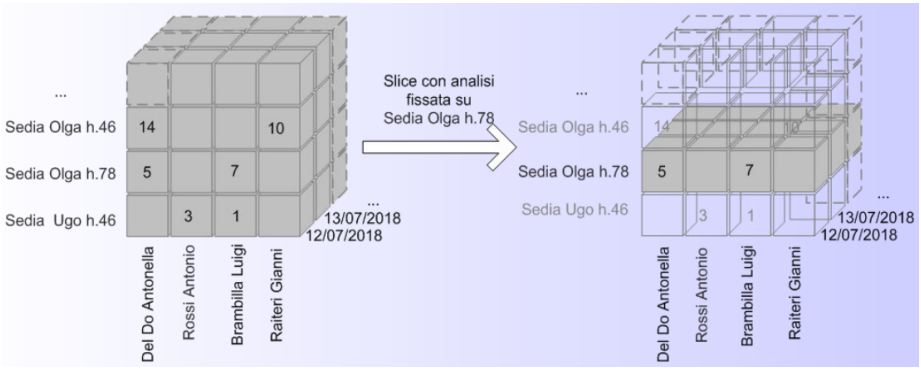
Roll up

E' il reciproco del drill down, si procede risalendo una gerarchia oppure eliminando una dimensione di analisi.



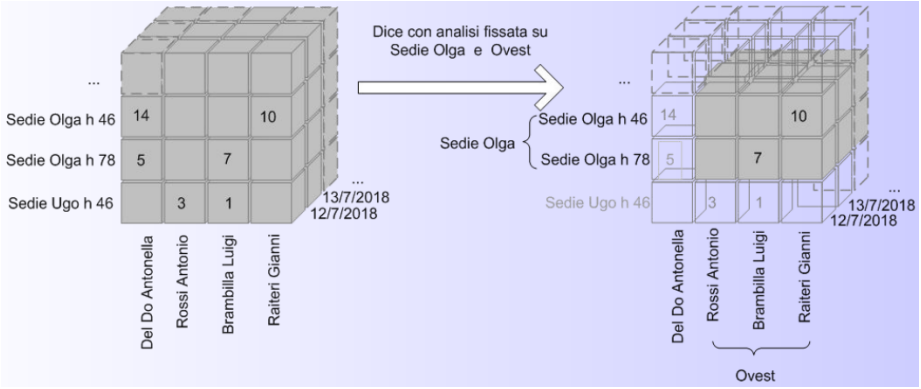
Slice

Viene fissato il valore di una dimensione e vengono analizzati i dati così ottenuti.



Dice

Simile allo slice ma può operare su più dimensioni fissandone il valore, si può applicare a dimensioni di qualsiasi livello.



Pivot

(Trasposizione di matrice nel caso 2D) Le dimensioni della matrice vengono invertite.

Prodotto	Area/Anno	2017	2018
Sedia Olga h.46	Est	203	220
	Ovest	64	64

Prodotto	Anno/Area	Est	Ovest
Sedia Olga h.46	2017	203	64
	2018	220	64