Introduzione al data warehousing

Luca Cabibbo, Riccardo Torlone

aprile 2012

Motivazioni

I sistemi informatici permettono di aumentare la produttività delle organizzazioni automatizzandone la gestione quotidiana dei **processi operativi**

- vendite nelle catene di supermercati
- instradamento e la contabilizzazione delle telefonate

Questi dati – se opportunamente **accumulati** e **analizzati** – possono essere utilizzati per supportare i **processi gestionali** e **direzionali**, ovvero per il controllo, la pianificazione e il supporto alle decisioni

- promozioni dei prodotti
- offerta di contratti diversificati

Perché?

- una corretta gestione dei dati storici può essere occasione di un grande vantaggio competitivo

Processi aziendali, dati e decisioni

I processi informativi svolti da un'organizzazione possono essere classificati in tre grandi categorie

- processi operativi
- processi gestionali
- processi direzionali

Banca

Processi operativi

- gestione di un movimento su un conto corrente bancario, presso sportello tradizionale o automatico

Processi gestionali

- concessione di un fido
- revisione delle condizioni su un conto corrente

Processi direzionali

- verifica dell'andamento dei servizi di carta di credito
- lancio di una campagna promozionale
- stipula di accordi commerciali

Compagnia telefonica

Processi operativi

- stipula di contratti ordinari
- instradamento delle telefonate
- memorizzazione di dati contabili sulle telefonate (chiamante, chiamato, giorno, ora, durata, instradamento,..)

Processi gestionali

- stipula di contratti speciali
- installazione di infrastrutture

Processi direzionali

- scelta dei parametri che fissano il costo delle telefonate
- definizione di contratti diversificati
- pianificazione del potenziamento delle infrastrutture

Caratteristiche dei processi

Processi operativi

- operano sui dati dipartimentali e dettagliati
- le operazioni sono strutturate, basate su regole perfettamente definite

Processi gestionali

- operano su dati settoriali e parzialmente aggregati
- le operazioni sono semi-strutturate, basate su regole note, ma in cui è spesso necessario un intervento umano “creativo”

Processi direzionali

- operano su dati integrati e fortemente aggregati
- le operazioni sono non strutturate, non esistono criteri precisi, e la capacità personale è essenziale

Informatizzazione dei processi

L'informatizzazione di un processo è funzione del suo grado di strutturazione delle sue operazioni

- un processo altamente strutturato può essere facilmente informatizzato
- un processo non strutturato può essere al più supportato da applicazioni informatiche

Tipologie di sistemi informatici

Transaction Processing Systems

- dipartimentali, per i processi operativi

Management Information Systems

- settoriali, anche per processi gestionali

Decision Support Systems – Business Intelligence

- fortemente integrati, di supporto ai processi direzionali

Sistemi di supporto alle decisioni

In particolare, i sistemi di supporto alle decisioni (DSS) costituiscono la tecnologia che supporta la dirigenza aziendale nel prendere decisioni tattico-strategiche in modo efficace e veloce, mediante particolari tipologie di elaborazione (per esempio OLAP)

Ma su quali dati?

- quelli accumulati per i processi operativi e gestionali!

Operazioni supportate dai DSS

Esempi di operazioni supportate dai DSS

- quali sono stati i volumi di vendita dello scorso anno per regione e categoria di prodotto?
- quali ordini dovremmo soddisfare per massimizzare le entrate?
- uno sconto tra il 7% e il 10% potrebbe incrementare le vendite di un certo prodotto in modo sufficiente?
- in che modo i dividendi di aziende di hardware sono correlati ai loro profitti trimestrali negli ultimi 10 anni?

Business Intelligence

Business Intelligence (**BI**) è, genericamente, il processo di trasformazione di dati e informazioni in conoscenza

- nello specifico, le tecnologie BI hanno lo scopo di supportare un'organizzazione nello sfruttare al meglio il suo patrimonio informativo (interno ed esterno) nei processi decisionali (decision making)
- "Business Intelligence is a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making"
- le tecnologie BI forniscono delle viste storiche, correnti e predittive sui processi di business – alcune funzioni comuni
  - reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining e predictive analytics

Tipologie di elaborazione

Nei Transaction Processing Systems

- On-Line Transaction Processing

Nei Management Information Systems

- On-Line Transaction Processing + applicazioni evolute ("intelligenti")

Nei Decision Support Systems

- On-Line Analytical Processing

Più in generale, nei sistemi BI

- On-Line Analytical Processing
- querying
- reporting
- business analytics

OLTP

I sistemi di gestione di basi di dati relazionali sono normalmente ottimizzati per supportare le operazioni transazionali (OLTP, **On Line Transaction Processing**)

- tradizionale elaborazione di transazioni, che realizzano i processi operativi dell'organizzazione
- le transazioni sono predefinite e di breve durata
- i dati di interesse sono dettagliati, aggiornati e recenti
- i dati risiedono su una unica base di dati
- leggono e/o modificano pochi record
- le proprietà "acide" (atomicità, correttezza, isolamento, durabilità) delle transazioni sono essenziali
- architettura (principalmente) centralizzata

OLAP

I DSS devono invece supportare l'elaborazione analitica (OLAP, **On-Line Analytical Processing**), che ha le seguenti caratteristiche

- le interrogazioni sono complesse e casuali
- leggono un numero enorme di record –
- i dati di interesse sono tipicamente storici e aggregati
- i dati possono provenire da più basi di dati — possibilmente non omogenee
- le risposte alle interrogazioni sono attese in linea
- la visualizzazione dei dati è fondamentale
- non scrivono mai
- le proprietà "acide" non sono rilevanti, perché le operazioni sono di sola lettura
- architettura client-server

OLTP e OLAP

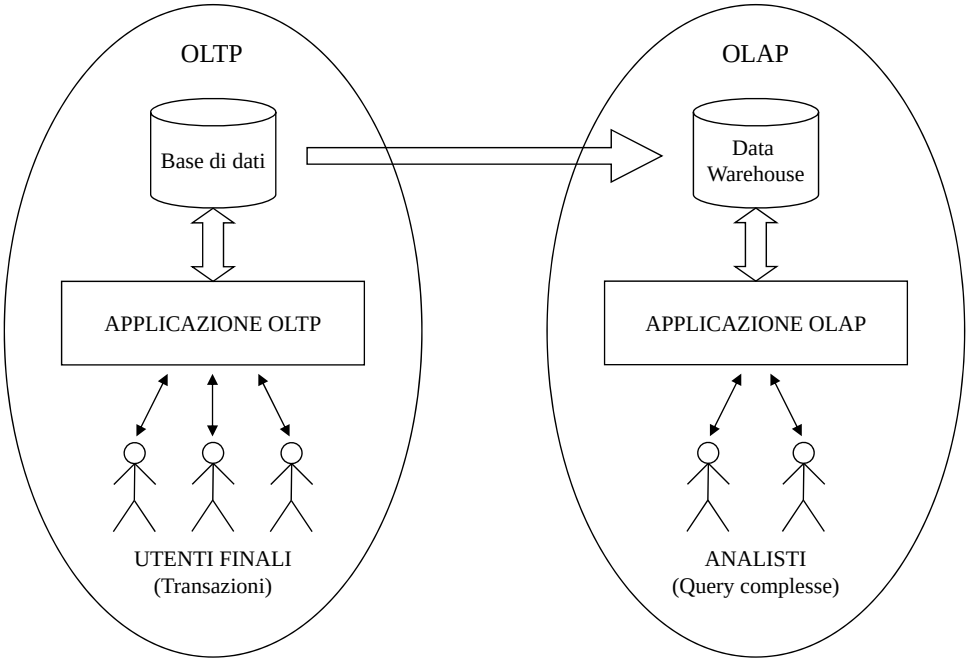
	OLTP	OLAP
Utente	impiegato	dirigente
Funzione	operazioni giornaliere	supporto alle decisioni
Progettazione	orientata all'applicazione	orientata ai dati
Dati	correnti, aggiornati, dettagliati, relazionali, omogenei	storici, aggregati, multidimensionali, eterogenei
Uso	ripetitivo	casuale
Accesso	read-write, indicizzato	read, sequenziale
Unità di lavoro	transazione breve	interrogazione complessa
Record acc.	decine	milioni
N. utenti	migliaia	centinaia
Dimensione	100MB - 10GB	100GB - 10TB
Metrica	throughput	tempo di risposta

OLTP e OLAP

I requisiti sono quindi contrastanti

Le applicazioni dei due tipi possono danneggiarsi a vicenda

Separazione degli ambienti



Una possibile definizione di data warehouse

Un **data warehouse** è

- una base di dati
- utilizzata principalmente per il supporto alle decisioni direzionali
- integrata – aziendale e non dipartimentale
- con dati storici – con un ampio orizzonte temporale, e indicazione di almeno un elemento di tempo
- con dati usualmente aggregati – per effettuare stime e valutazioni
- fuori linea – i dati sono aggiornati periodicamente
- mantenuta separata dalle basi di dati operazionali

... integrata ...

I dati di interesse provengono da tutte le sorgenti informative – ciascun dato proviene da una o più di esse

- il data warehouse rappresenta i dati in modo univoco – riconciliando le eterogeneità dalle diverse rappresentazioni
  - nomi
  - struttura
  - codifica
  - rappresentazione multipla

... dati storici ...

Le basi di dati operazionali mantengono il valore corrente delle informazioni

- l'orizzonte temporale di interesse è dell'ordine dei pochi mesi

Nel data warehouse è di interesse l'evoluzione storica delle informazioni

- l'orizzonte temporale di interesse è dell'ordine degli anni

... dati aggregati ...

Nelle attività di analisi dei dati per il supporto alle decisioni

- non interessa “chi” ma “quanti”
- non interessa un dato ma
  - la somma, la media, il minimo e il massimo, ...
- di un insieme di dati

Le operazioni di aggregazione sono quindi fondamentali nel warehousing e nella costruzione/mantenimento di un data warehouse

... fuori linea ...

In una base di dati operazionale, i dati vengono

- acceduti
- inseriti
- modificati
- cancellati
- pochi record alla volta

Nel data warehouse, abbiamo

- operazioni di accesso e interrogazione – “diurne”
- operazioni di caricamento e aggiornamento dei dati – “notturne”

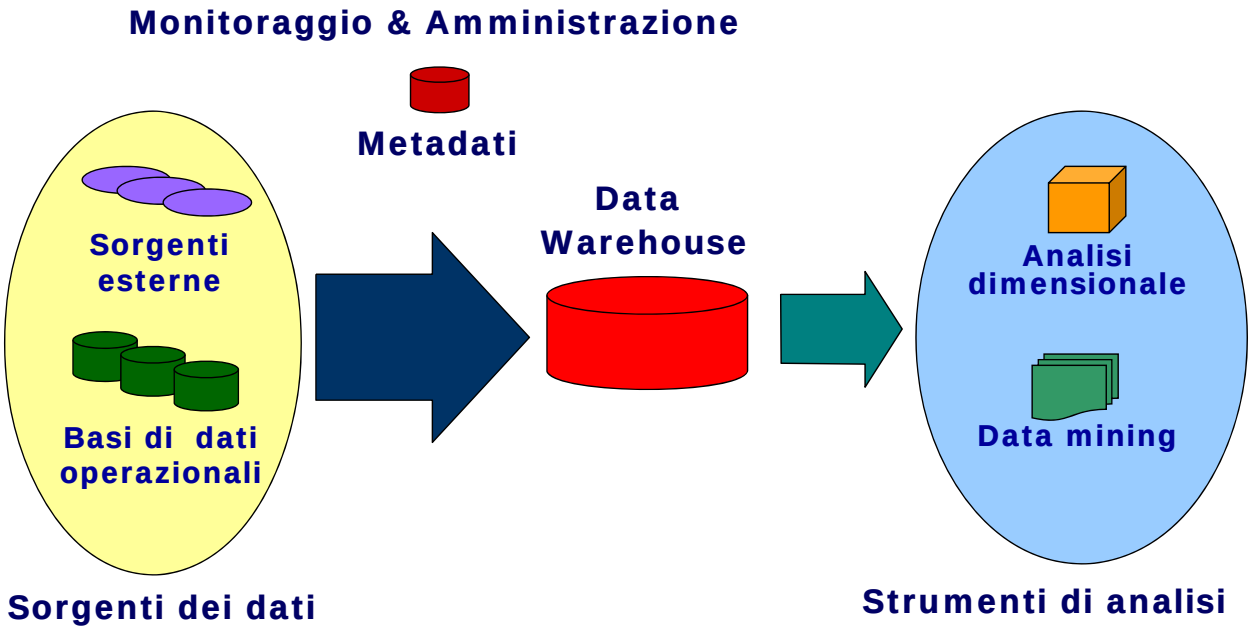
che riguardano milioni di record

... una base di dati separata ...

Diversi motivi

- non esiste un'unica base di dati operazionale che contiene tutti i dati di interesse
- la base di dati deve essere integrata
- non è tecnicamente possibile fare l'integrazione in linea
- i dati di interesse sarebbero comunque diversi
  - devono essere mantenuti dati storici e aggregati
- l'analisi dei dati richiede per i dati organizzazioni speciali e metodi di accesso specifici
- degrado generale delle prestazioni senza la separazione

Architettura generale per il data warehousing



Metadati

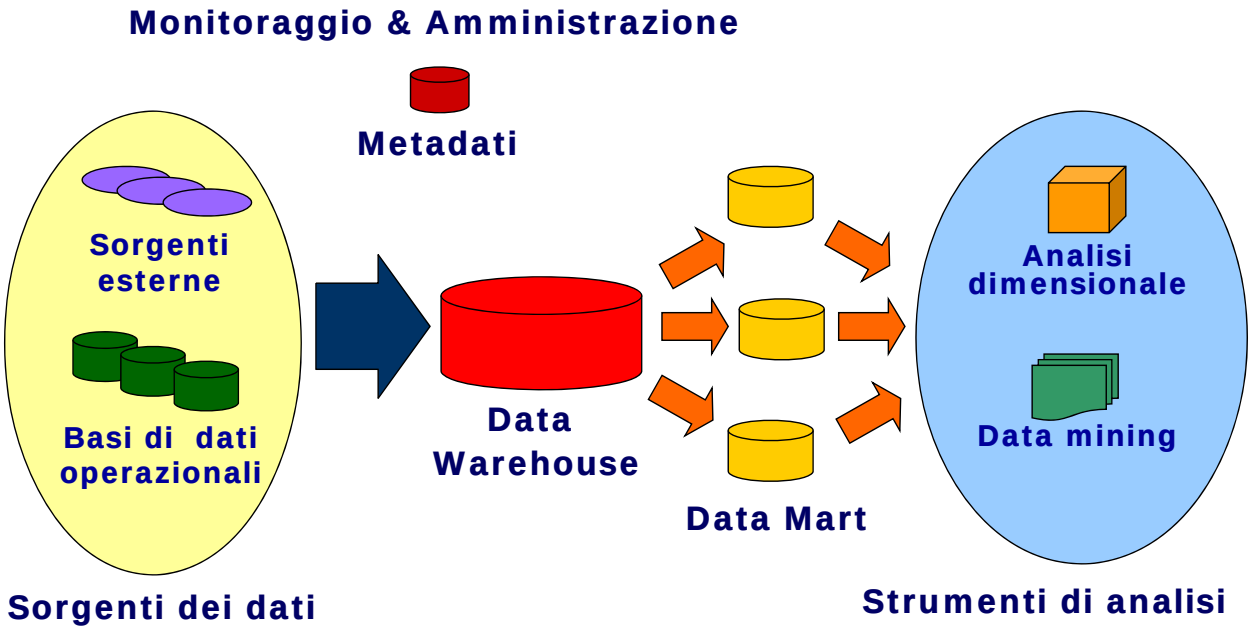
I metadati sono “dati sui dati”

- descrizioni logiche e fisiche dei dati (nelle sorgenti e nel DW)
- corrispondenze e trasformazioni
- dati quantitativi

I metadati – pur importanti nelle basi di dati – sono fondamentali in un DW

- in particolare, nell’analisi
  - a che periodo si riferiscono i dati in quella tabella?
  - che cosa rappresenta, con precisione, quel campo?
  - ...

Architettura per il data warehousing (Inmon)

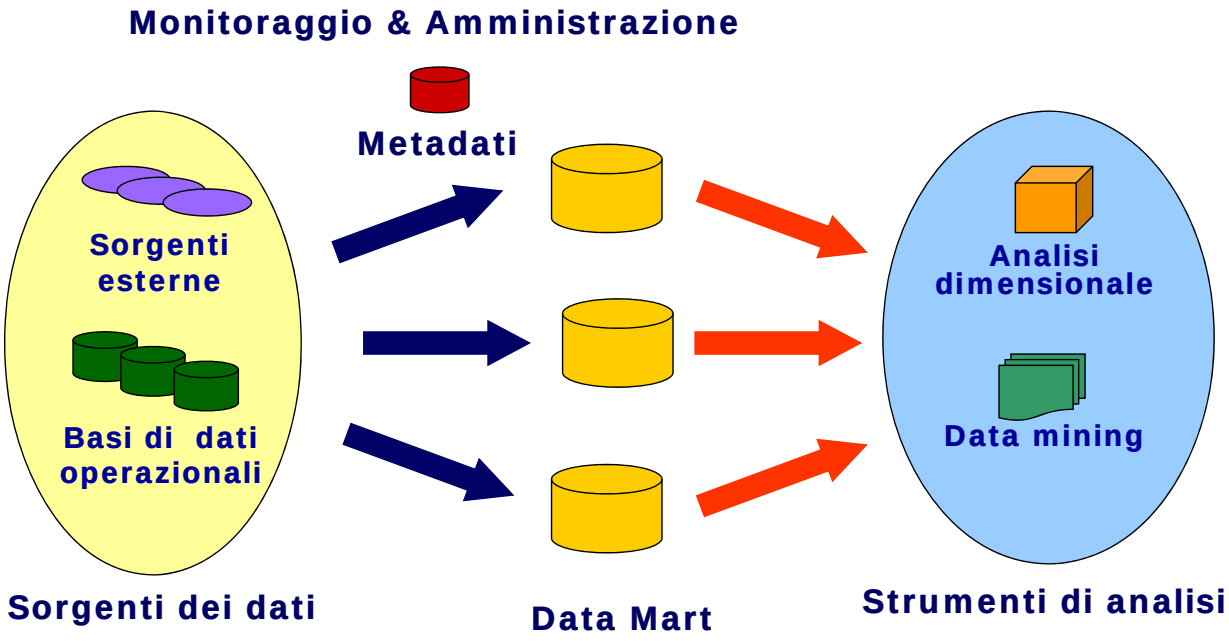


Data mart (Inmon)

Secondo Inmon, un **data mart**

- è una struttura dipartimentale di dati estratti dal data warehouse
- in un data mart, i dati sono rappresentati sulla base delle necessità di analisi del dipartimento – per questo, sono di solito aggregati
- dunque, un data mart è sostanzialmente la restrizione del data warehouse a un singolo problema di analisi

Architettura per il data warehousing (Kimball)



Data mart (Kimball)

Secondo Kimball, invece, un **data mart**

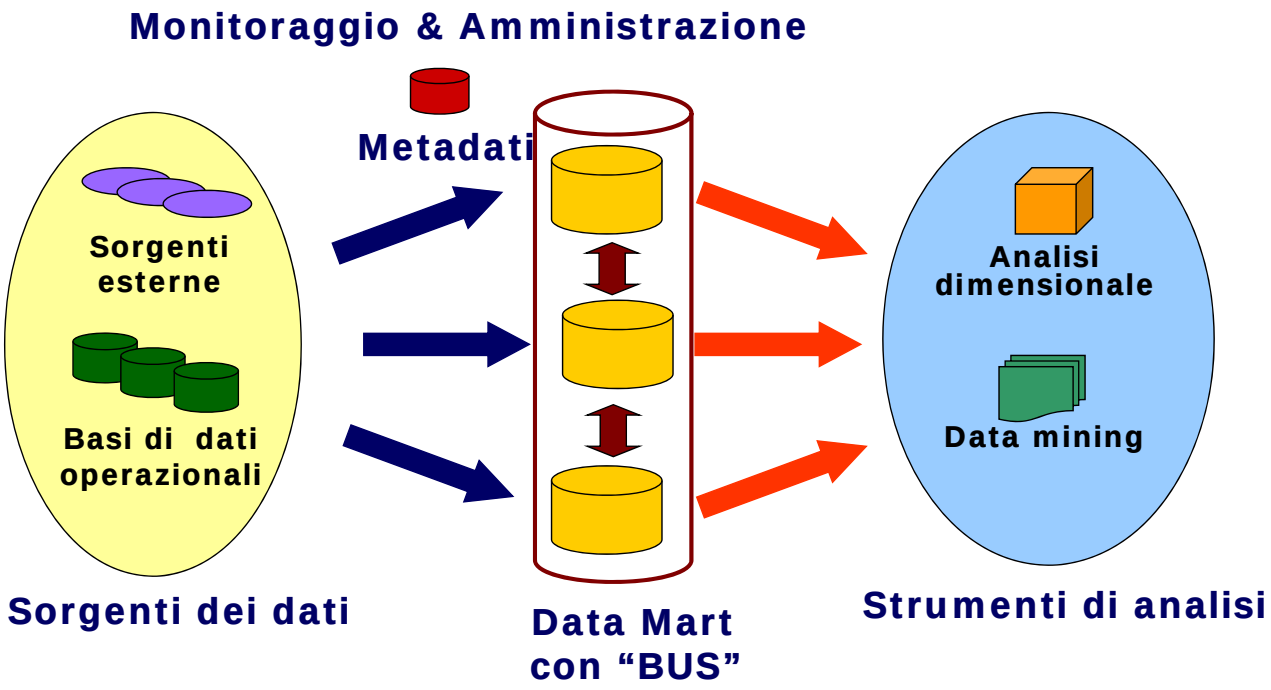
- è un sottoinsieme, logico e fisico, di un data warehouse
  - nella sua forma più semplice, un data mart rappresenta i dati da un singolo processo di business
- un data warehouse è l'unione di tutti i suoi data mart
- dunque, un data mart è un insieme di dati atomici, estratte e integrati dalle sorgenti operative, e organizzati in modo tale da soddisfare direttamente le necessità di interrogazione da parte degli strumenti di analisi
- inoltre, i data mart di un data warehouse possono essere interrogati in modo combinato, "trasversale"

Data warehouse – Inmon vs. Kimball

Due visioni diverse dai padri del data warehousing

- entrambi sostengono uno sviluppo iterativo del DW
- Inmon sostiene lo sviluppo di un singolo, grande DW per l'intera organizzazione
  - ma i dati nel DW non sono organizzati per essere interrogati direttamente dagli strumenti di analisi
  - piuttosto, devono prima essere estratti in un data mart (nel suo senso)
- Kimball sostiene invece che il DW va costruito come un insieme di data mart (nel suo senso)
  - i dati nel DW (un insieme di data mart) possono essere interrogati direttamente dagli strumenti di analisi
  - è necessaria un'architettura che garantisca la coerenza dei dati mart – affinché la loro unione formi effettivamente un DW – la cosiddetta architettura "a bus"

Architettura per il data warehousing (Kimball)



Dati multidimensionali

I dati in un DW sono di solito organizzati in forma *multidimensionale* – una forma adeguata all'analisi dei dati – ovvero organizzati mediante i seguenti concetti

- **fatto** (o processo)
  - un concetto sul quale centrare l'analisi
- **misura**
  - una proprietà atomica o misura di un fatto da analizzare
  - le misure sono solitamente valori numerici e additivi su un dominio continuo
- **dimensione**
  - una prospettiva rispetto alla quale effettuare l'analisi
  - le dimensioni descrivono domini discreti, solitamente organizzati in livelli di aggregazione

Dati multidimensionali – esempi

Data mart delle vendite

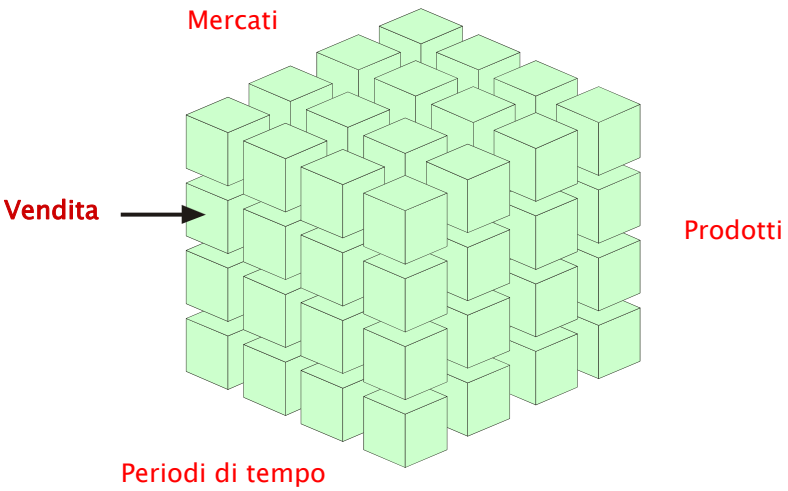
- fatto: vendite dei prodotti, giornaliere, per negozio
- dimensioni: prodotto, tempo (giorno), negozio, promozione
- misure: quantità venduta, incasso, costo, conteggio dei clienti

Data mart delle telefonate

- fatto: telefonata
- dimensioni: chiamante, chiamato, tariffa, tempo (giorno), tempo (ora del giorno)
- misure: durata, costo

Rappresentazione multidimensionale dei dati

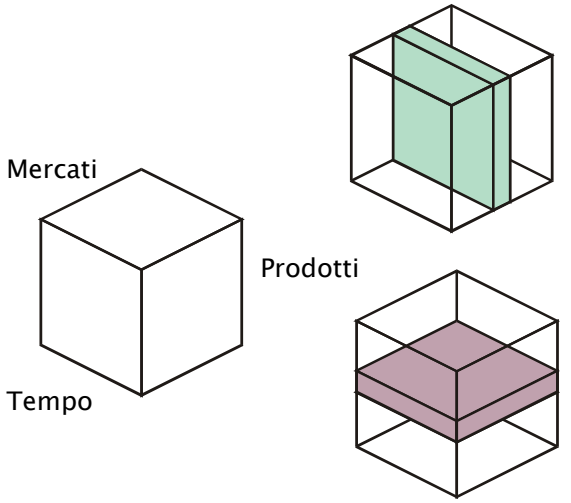
Gli analisti sono abituati a ragionare in termini di dimensioni e misure – non di schemi, tabelle e record



Viste su dati multidimensionali

Il manager regionale esamina la vendita dei prodotti in tutti i periodi relativamente ai propri mercati

Il manager finanziario esamina la vendita dei prodotti in tutti i mercati relativamente al periodo corrente e quello precedente



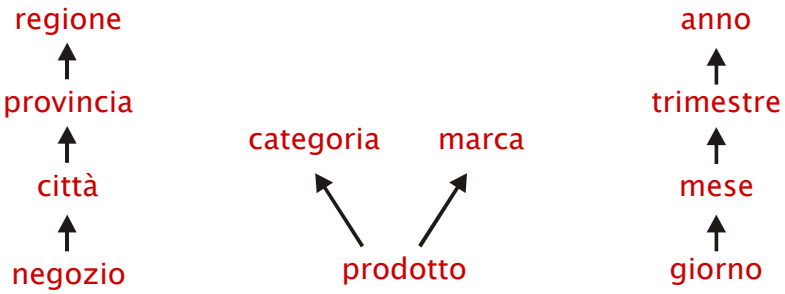
Il manager di prodotto esamina la vendita di un prodotto in tutti i periodi e in tutti i mercati

Il manager strategico si concentra su una categoria di prodotti, una area regionale e un orizzonte temporale medio

Dimensioni e gerarchie di livelli

Ciascuna dimensione è organizzata in una gerarchia che rappresenta i possibili **livelli** di aggregazione per i dati

- negozio, città, provincia, regione
- prodotto, categoria, marca
- giorno, mese, trimestre, anno



Operazioni su dati multidimensionali

Slice & dice – seleziona e proietta – solitamente su un piano bidimensionale

Roll up – aggrega i dati (rispetto all'interrogazione corrente), ovvero mostra dati a un maggior livello di aggregazione

Drill down – disaggrega i dati (rispetto all'interrogazione corrente), ovvero mostra dati a un minor livello di aggregazione

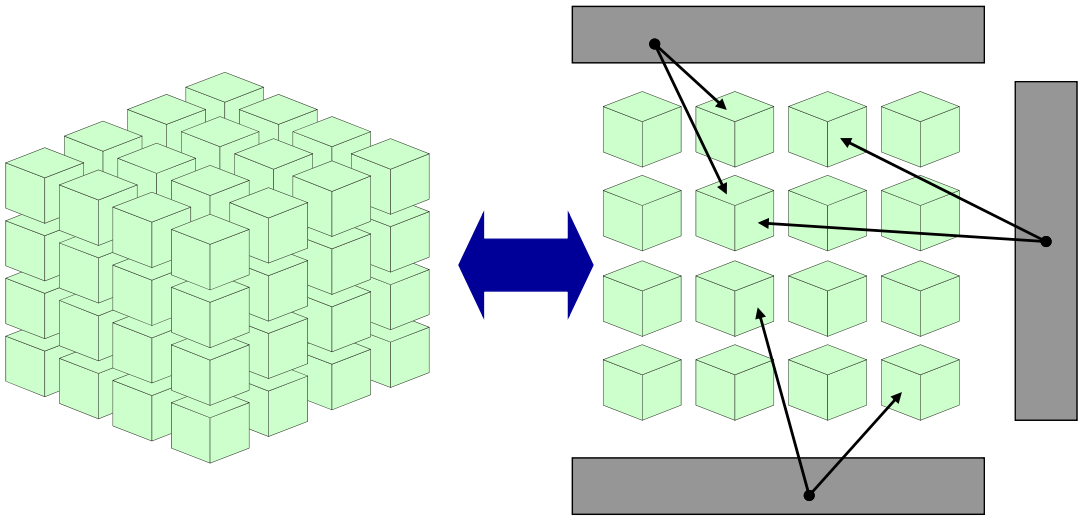
Drill across – combina i dati associati a più fatti

Pivot – re-orienta il cubo

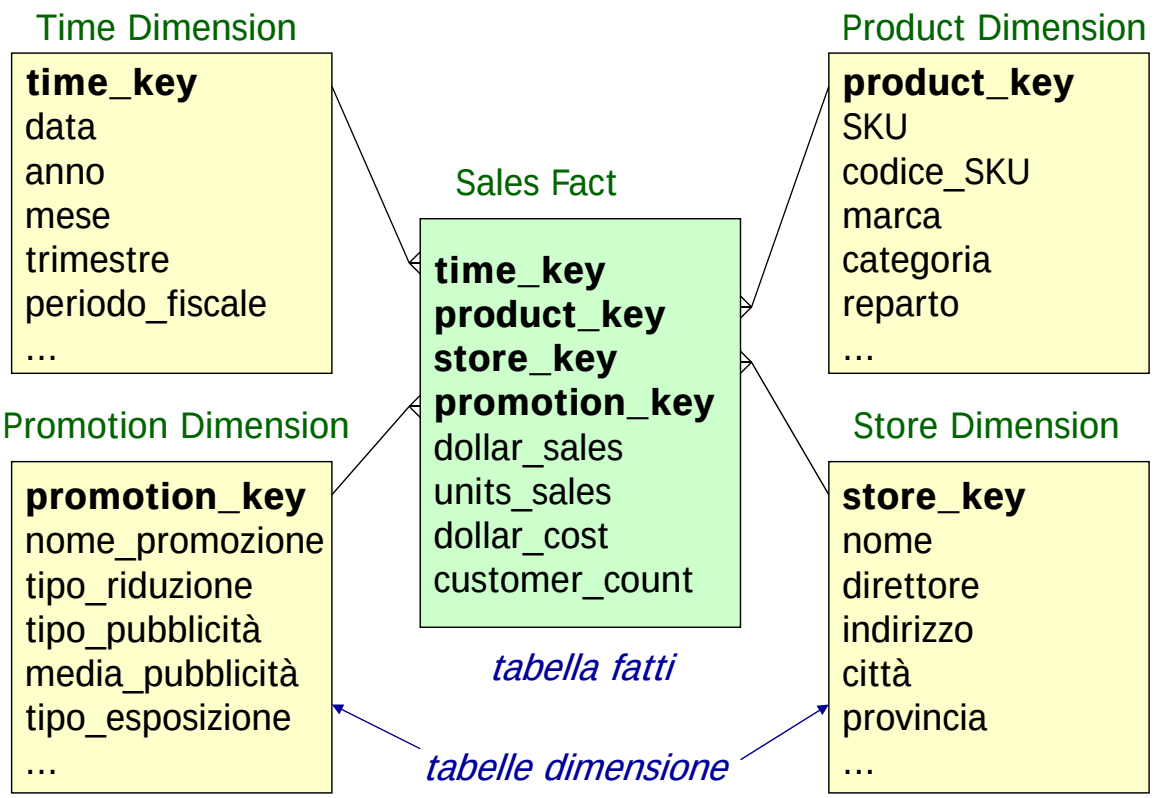
Rappresentazione MOLAP

I dati sono memorizzati direttamente in un formato dimensionale (proprietario)

- le gerarchie sui livelli sono codificate in indici di accesso alle matrici



Rappresentazione ROLAP



Strumenti di analisi

Gli strumenti di analisi devono supportare in modo efficace il lavoro degli analisti di business

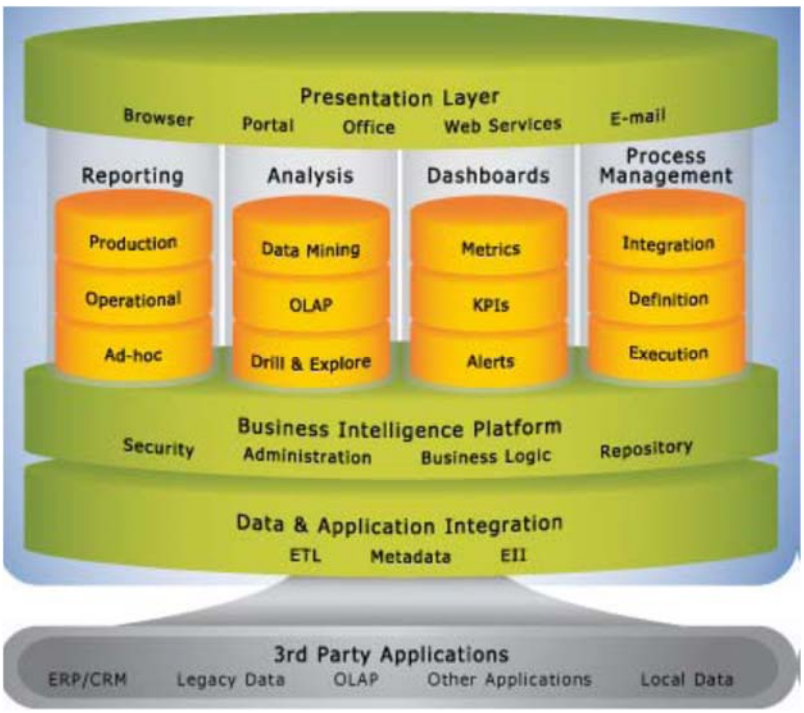
- utenti diversi hanno necessità di analisi differenti
  - da report standard a report personalizzabili a interrogazioni ad hoc
- devono consentire agli utenti di business di dominare la complessità dei dati
  - aiuto nella comprensione dei dati disponibili
  - visualizzazioni multiple degli stessi dati – tabelle, grafici, dashboard
- devono fornire un'ampia famiglia di algoritmi di analisi

Pentaho BI Open Source

A titolo di esempio, vengono brevemente presentati alcuni strumenti di Pentaho – una famiglia di prodotti open source per la Business Intelligence

- Pentaho BI Platform
  - fornisce un punto di accesso in rete alle capacità di BI – in modo sicuro e integrato
- Pentaho Reporting
  - consente l'organizzazione e la formattazione di report
- Mondrian – analisi di dati dimensionali (OLAP)
- Pentaho Dashboard
  - fornisce agli utenti di business l'accesso immediato e visuale a informazioni critiche e indicatori di prestazioni
- Weka – data mining
- Kettle – ETL

Pentaho BI Platform



Pentaho Reporting

Region: Central

Department	Position	Actual	Budget	Variance
Executive Management				
	SVP Partnerships	$367,415	$392,100	$24,685
	SVP WW Operations	$476,000	$725,887	$249,887
	SVP Strategic Development	$383,242	$403,405	$20,163
	CEO	$549,625	$522,250	-$27,375
	Total	**$1,776,282**	**$2,043,642**	**$267,360**
Department	**Position**	**Actual**	**Budget**	**Variance**
Finance				
	Controller	$570,373	$577,070	$6,697
	Payroll	$367,415	$432,100	$64,685
	Administrative Assistant	$827,861	$760,990	-$66,871
	IS	$570,759	$577,346	$6,587
	CFO	$770,272	$719,855	-$50,417
	Total	**$3,106,680**	**$3,067,361**	**-$39,319**
Department	**Position**	**Actual**	**Budget**	**Variance**
Human Resource				
	Sexual Harassment	$530,473	$538,570	$8,097
	EOE	$530,207	$538,380	$8,173
	HR Generalists	$856,190	$771,225	-$84,965
	HR Training	$397,473	$443,570	$46,097
	Administration	$549,625	$552,250	$2,625
	SVP HR	$574,896	$570,300	-$4,596
	Total	**$3,438,863**	**$3,414,295**	**-$24,568**
Department	**Position**	**Actual**	**Budget**	**Variance**
Marketing & Communication				
	Graphics	$782,375	$728,500	-$53,875
	Writer	$405,985	$459,650	$53,665
	Analyst Relations	$383,375	$443,500	$60,125
	Press Relations	$497,296	$524,872	$27,576
	CMO	$827,861	$760,990	-$66,871
	Product Marketing Mgr	$693,531	$665,040	-$28,491

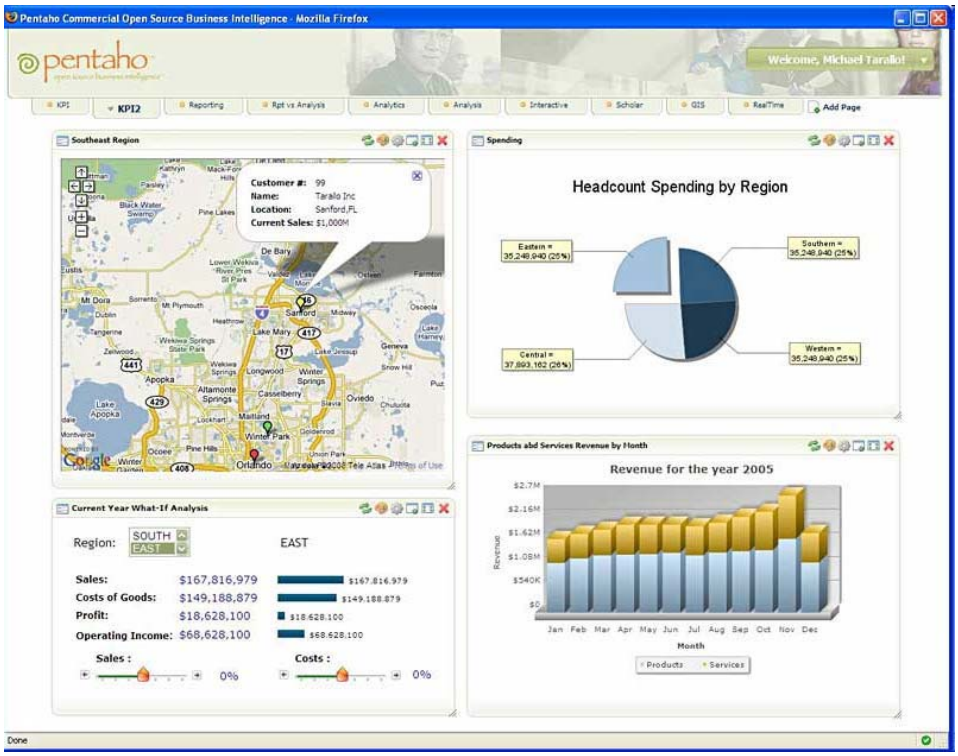
Mondrian

		Region			
Positions	**Measures**	**Central**	**Eastern**	**Southern**	**Western**
CEO	**Actual**	549,625.00	500,000.00	500,000.00	500,000.00
	Budget	522,250.00	488,750.00	498,750.00	478,750.00
	Variance Percent	-5.24%	-2.30%	-.25%	-4.44%
HR Generalists	**Actual**	856,190.00	961,000.00	961,000.00	961,000.00
	Budget	771,225.00	940,158.00	940,158.00	938,158.00
	Variance Percent	-11.02%	-2.22%	-2.22%	-2.43%
HR Training	**Actual**	397,473.00	271,200.00	271,200.00	271,200.00
	Budget	443,570.00	279,674.00	279,674.00	277,674.00
	Variance Percent	10.39%	3.03%	3.03%	2.33%

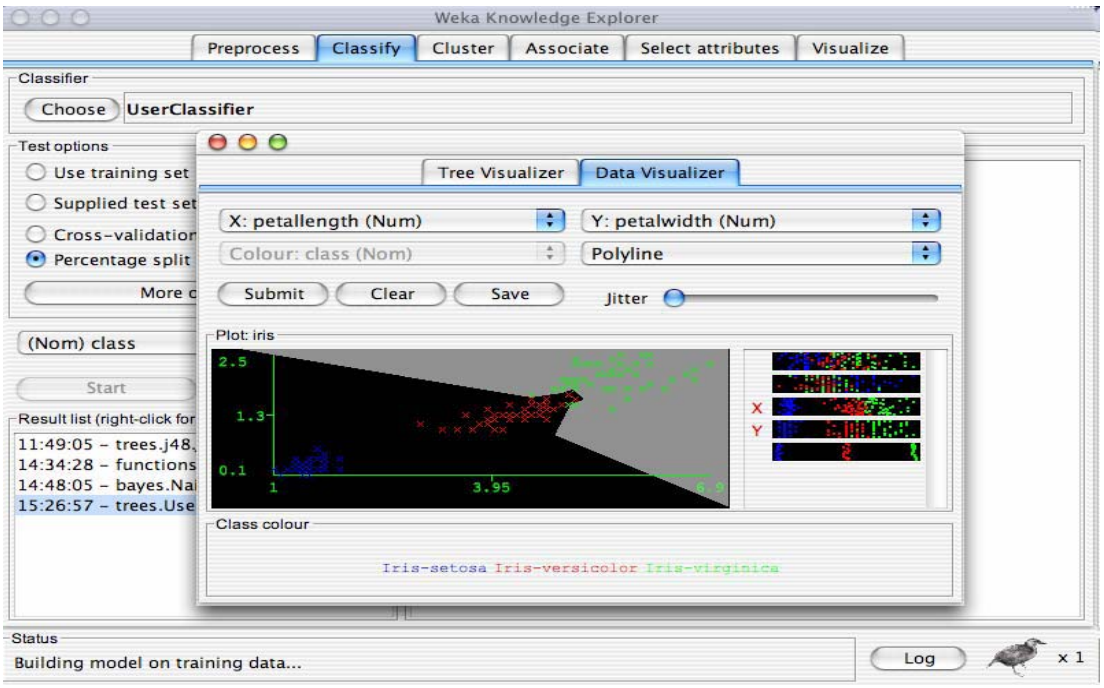
Slicer: [(All)=All Departments]

Pentaho Dashboard

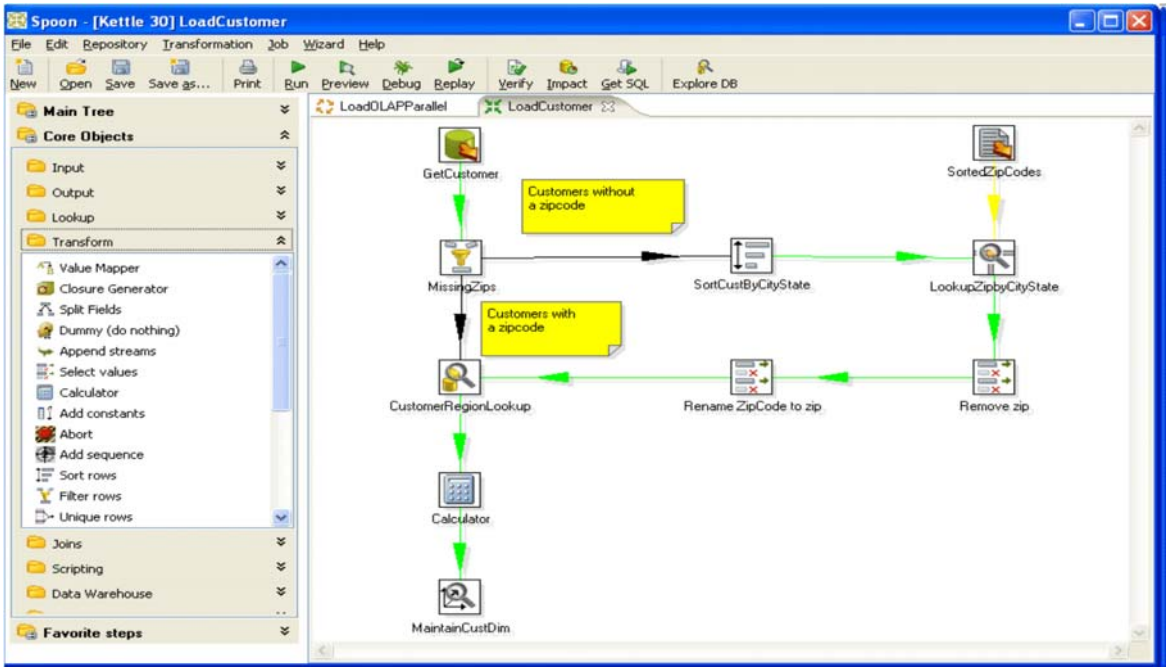
Pentaho Dashboard



Weka


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
UserClassifier
Test options
Use training set
Supplied test set
Cross-validation
Percentage split
More
(Nom) class
Start
Result list (right-click for
11:49:05 - trees.j48
14:34:28 - functions
14:48:05 - bayes.Na
15:26:57 - trees.Use
Tree Visualizer
Data Visualizer
X: petallength (Num)
Y: petalwidth (Num)
Colour: class (Nom)
Polyline
Submit
Clear
Save
Jitter
Plot: iris
2.5
1.3
0.1
1
3.95
X
Y
Class colour
Iris-setosa Iris-versicolor Iris-virginica
Status
Building model on training data...
Log
x 1

Kettle


Spoon - [Kettle 30] LoadCustomer
File Edit Repository Transformation Job Wizard Help
New Open Save Save as... Print Run Preview Debug Replay Verify Impact Get SQL Explore DB
Main Tree
Core Objects
Input
Output
Lookup
Transform
Value Mapper
Closure Generator
Split Fields
Dummy (do nothing)
Append streams
Select values
Calculator
Add constants
Abort
Add sequence
Sort rows
Filter rows
Unique rows
Joins
Scripting
Data Warehouse
Favorite steps
Load0LAPParallel
LoadCustomer
GetCustomer
Customers without a zipcode
SortedZipCodes
MissingZips
SortCustByCityState
LookupZipbyCityState
Customers with a zipcode
CustomerRegionLookup
Rename ZipCode to zip
Remove zip
Calculator
MaintainCustDim