

Capitolo 8

Esercizio 8.1

Progettare un cubo multidimensionale relativo all'analisi dei sinistri per una compagnia assicurativa, basandosi sulle specifiche accennate nel paragrafo 8.2.1.

Soluzione:

Per la progettazione di un cubo dimensionale si devono individuare i tre concetti di base: il fatto, la misura e la dimensione.

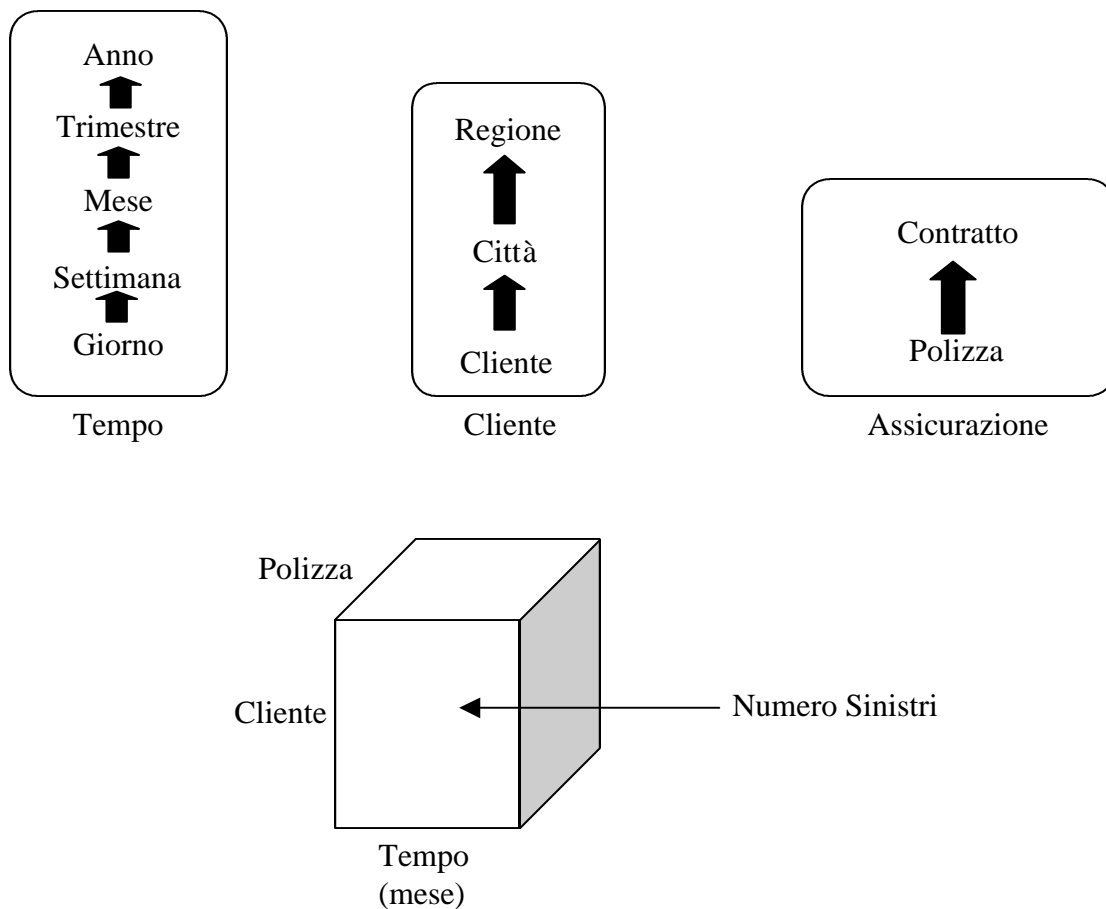
Nel nostro caso, la compagnia di assicurazioni avrà:

Fatto: SINISTRO

Misura: NUMERO DI SINISTRI – COSTO DEI SINISTRI

Dimensione: CLIENTE – TIPOLOGIA SINISTRO – POLIZZA – PERIODO DI TEMPO

Il passo successivo consiste nell'organizzare le dimensioni in gerarchie.



Esercizio 8.2

Descrivere alcune operazioni slice-and-dice, roll-up e drill down per il cubo multidimensionale definito nell'esercizio precedente.

Soluzione:

Sul cubo definito nell'esercizio precedente possiamo effettuare le seguenti operazioni:

Slice-and-dice:

Consiste nel definire un sottoinsieme del un cubo dimensionale.

Possiamo quindi selezionare i sinistri delle sole polizze RCA divisi per periodo e categoria di cliente.

Roll-up:

L'operazione di roll-up può essere di due tipi.

Il primo tipo consiste nell'applicare una funzione aggregata. Nel nostro caso possiamo aggregare i dati sulla dimensione tempo, facendoli diventare trimestrali.

Il secondo tipo consiste nell'eliminazione completa di una dimensione. Se togliamo la dimensione Cliente otteniamo un cubo bidimensionale che indica i sinistri che sono avvenuti in un determinato periodo di tempo per ogni categoria di polizza.

Drill down:

L'operazione di drill down è l'operazione inversa a quella di roll-up. Quindi anche questa operazione può essere di due tipi.

Il primo tipo consiste nell'aggiungere un livello di dettaglio eliminando una funzione aggregata. Ad esempio commutando il livello dalla dimensione Tempo da mesi in giorni.

Il secondo tipo consiste nell'aggiungere una dimensione, come potrebbe essere il Luogo in cui si è verificato il sinistro.

Esercizio 8.3

Indicare cosa si ottiene applicando un roll-up che elimina la condizione ARTICOLO dal cubo in figura 8.7.

Soluzione:

L'eliminazione della dimensione ARTICOLO dal cubo in figura 8.7 porta ad avere le vendite complessive di tutti i negozi organizzate per trimestre.

Esercizio 8.4

Scrivere una interrogazione SQL in grado di eseguire un roll-up che, a partire dallo schema stella in figura 8.10, calcola l'incasso totale per marca di prodotto e città.

Soluzione:

L'interrogazione SQL è la seguente:

```
SELECT Articolo.Marca, Luogo.Città, Sum(Vendite.Incasso) AS Totale
FROM Tempo JOIN (Luogo JOIN (Cliente JOIN (Articolo JOIN
      Vendite ON Articolo.CodiceArticolo = Vendite.CodiceArticolo)
      ON Cliente.CodiceCliente = Vendite.CodiceCliente)
      ON Luogo.CodiceLuogo = Vendite.CodiceLuogo)
      ON Tempo.CodiceTempo = Vendite.CodiceTempo)
GROUP BY Articolo.Marca, Luogo.Città
```

Esercizio 8.5

Scrivere una interrogazione SQL in grado di eseguire un roll-up che, a partire dallo schema a fiocco di neve in figura 8.12, calcola il numero di articoli venduti per categoria di prodotto, mese ed età del cliente.

Soluzione:

```
SELECT Categoria.Categoria, Tempo.Mese, Cliente.Età, Sum(Vendite.Quantità) AS
      Quantità
FROM Tempo JOIN (Luogo JOIN (Cliente JOIN (
      (Articolo JOIN Categoria ON Articolo.CodiceCategoria =
      Categoria.CodiceCategoria) JOIN Vendite ON
      Articolo.CodiceArticolo = Vendite.CodiceArticolo) ON
      Cliente.CodiceCliente = Vendite.CodiceCliente) ON Luogo.CodiceLuogo
      = Vendite.CodiceLuogo) ON Tempo.CodiceTempo = Vendite.CodiceTempo)
GROUP BY Categoria.Categoria, Tempo.Mese, Cliente.Età
```

Esercizio 8.6

Scrivere una interrogazione SQL che mediante la clausola CUBE calcola le vendite complessive per trimestre e marca di prodotto a partire dallo schema stella in figura 8.10 e indicare un possibile risultato.

Soluzione:

```
SELECT Tempo.Trimestre, Articolo.Marca, Sum(Vendite.Incasso) AS Vendite
FROM Tempo JOIN (Luogo JOIN (Cliente JOIN (Articolo JOIN
      Vendite ON Articolo.CodiceArticolo =
      Vendite.CodiceArticolo) ON Cliente.CodiceCliente =
      Vendite.CodiceCliente) ON Luogo.CodiceLuogo =
      Vendite.CodiceLuogo) ON Tempo.CodiceTempo =
      Vendite.CodiceTempo)
GROUP BY CUBE (Tempo.Trimestre, Articolo.Marca)
```

Un possibile risultato potrebbe essere il seguente:

Trimestre	Marca	Vendite
1 Trimestre	Agnesi	50000

1 Trimestre	Barilla	200000
1 Trimestre	Ferrero	100000
1 Trimestre	Nestlè	350000
2 Trimestre	Agnesi	80000
2 Trimestre	Barilla	250000
2 Trimestre	Ferrero	100000
2 Trimestre	Nestlè	300000
3 Trimestre	Agnesi	40000
3 Trimestre	Barilla	200000
3 Trimestre	Ferrero	60000
3 Trimestre	Nestlè	10000
4 Trimestre	Agnesi	80000
4 Trimestre	Barilla	300000
4 Trimestre	Ferrero	200000
4 Trimestre	Nestlè	300000
1 Trimestre	ALL	700000
2 Trimestre	ALL	730000
3 Trimestre	ALL	310000
4 Trimestre	ALL	880000
ALL	Agnesi	250000
ALL	Barilla	950000
ALL	Ferrero	460000
ALL	Nestlè	960000
ALL	ALL	2620000

Esercizio 8.7

Mostrare il risultato che si ottiene sostituendo la clausola CUBE con la clausola ROLLUP nell'interrogazione dell'esercizio precedente.

Soluzione:

```
SELECT Tempo.Trimestre, Articolo.Marca, Sum(Vendite.Incasso) AS Vendite
FROM Tempo JOIN (Luogo JOIN (Cliente JOIN (Articolo JOIN
      Vendite ON Articolo.CodiceArticolo =
        Vendite.CodiceArticolo) ON Cliente.CodiceCliente =
        Vendite.CodiceCliente) ON Luogo.CodiceLuogo =
        Vendite.CodiceLuogo) ON Tempo.CodiceTempo =
        Vendite.CodiceTempo
GROUP BY ROLLUP (Tempo.Trimestre, Articolo.Marca)
```

Il risultato riferito all'esempio dell'esercizio precedente è il seguente:

Trimestre	Marca	Vendite
1 Trimestre	Agnesi	50000
1 Trimestre	Barilla	200000
1 Trimestre	Ferrero	100000
1 Trimestre	Nestlè	350000
2 Trimestre	Agnesi	80000

Basi di dati - Architetture e linee di evoluzione 2/ed

Paolo Atzeni, Stefano Ceri, Piero Fraternali, Stefano Paraboschi, Riccardo Torlone

2 Trimestre	Barilla	250000
2 Trimestre	Ferrero	100000
2 Trimestre	Nestlè	300000
3 Trimestre	Agnesi	40000
3 Trimestre	Barilla	200000
3 Trimestre	Ferrero	60000
3 Trimestre	Nestlè	10000
4 Trimestre	Agnesi	80000
4 Trimestre	Barilla	300000
4 Trimestre	Ferrero	200000
4 Trimestre	Nestlè	300000
ALL	Agnesi	250000
ALL	Barilla	950000
ALL	Ferrero	460000
ALL	Nestlè	960000
ALL	ALL	2620000

Esercizio 8.8

Indicare una scelta motivata di indici bitmap, indici di join e viste materializzate per lo schema a stella in figura 8.10.

Soluzione:

Indici bitmap:

Nello schema a stella della figura 8.10 si possono creare indici bitmap sui seguenti campi:

Marca e Categorie della relazione Articolo. Una selezione che deve scegliere un determinato articolo di una determinata categoria e marca dovrebbe solamente eseguire l'and bit a bit sui vettori corrispondenti alla Marca e alla Categoria. Questi campi hanno solitamente valori predefiniti o che vengono aggiornati piuttosto raramente, quindi non appesantiscono troppo la gestione della base dati.

Ad esempio, immaginando una tabella con cardinalità 10, tre Categorie e due Marche di prodotto, si creerebbero i seguenti indici Bitmap:

Rating bitmaps (Marche)

```
1: < 1 0 0 0 1 0 0 0 1 1 >    (Sony)
2: < 0 1 1 1 0 1 1 1 0 0 >    (Philips)
```

Rating bitmaps (Categoria)

```
1: < 0 1 0 0 1 0 0 0 1 0 >    (Notebook)
2: < 1 0 0 0 0 1 1 0 0 0 >    (Televisori)
3: < 0 0 1 1 0 0 0 1 0 1 >    (Accessori)
```

Una query che estrae i Televisori Sony dovrà prendere il primo vettore di Marche e il secondo di Categoria ed eseguire l'and bit a bit nel seguente modo.

```
M: < 1 0 0 0 1 0 0 0 1 1 >    (Sony)
C: < 1 0 0 0 0 1 1 0 0 0 >    (Televisori)
=====
R: < 1 0 0 0 0 0 0 0 0 0 >    (Risultato)
```

I bit posti a uno del vettore Risultato sono quelli che soddisfano le richieste.

Indici di join:

Si possono definire indici di join, relativamente allo schema a stella in figura 8.10, sugli attributi CodiceArticolo e CodiceLuogo. Grazie ad essi, la tabella dei fatti Vendite avrà le due chiavi che puntano ad insiemi di tuple che soddisfano la loro condizione nelle tabelle LUOGO e ARTICOLO. Questi campi non dovrebbero essere troppo frazionati e quindi l'appesantimento generale della base dati non sarà eccessivo.

Viste materializzate:

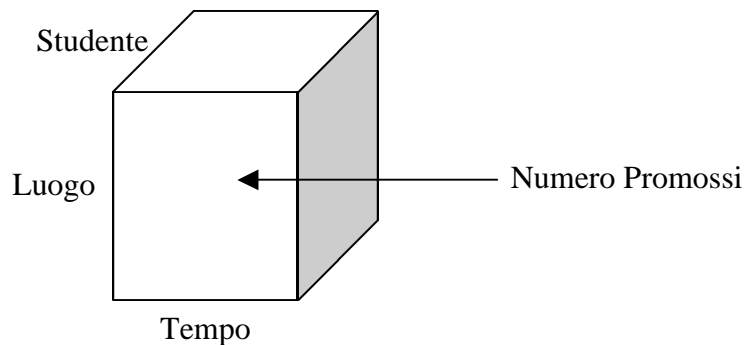
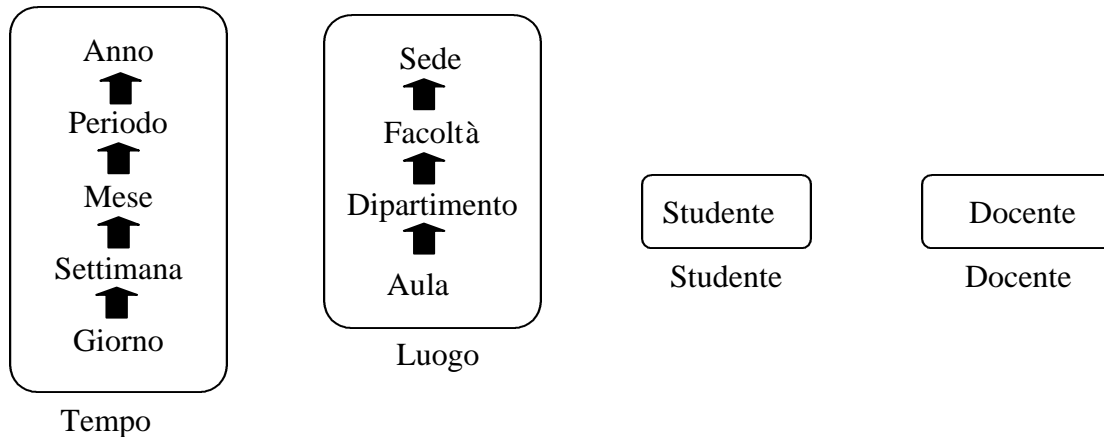
Per definire delle viste materializzate si dovrebbe essere a conoscenza delle tipiche interrogazioni e della loro frequenza di esecuzione.

Nella nostra base di dati sarebbe utile avere materializzati gli incassi mensili divisi per città, oppure le vendite mensili suddivise per sesso e fasce di età dei clienti. Questo genere di informazioni sono solitamente molto richieste in una grande azienda e possono servire come base per ulteriori interrogazioni, quindi è utile averle sempre a portata di mano.

Esercizio 8.9

Progettare un cubo dimensionale relativo alla gestione degli esami universitari, considerando come fatti gli esiti degli esami sostenuti dagli studenti e come dimensioni di analisi il tempo, la sede dell'esame (supponendo una facoltà distribuita su più sedi), il docente coinvolto, e le caratteristiche degli allievi (ad esempio, i dati relativi all'andamento scolastico pre-universitario, il punteggio nell'esame di ammissione e il corso di laurea prescelto).

Soluzione:

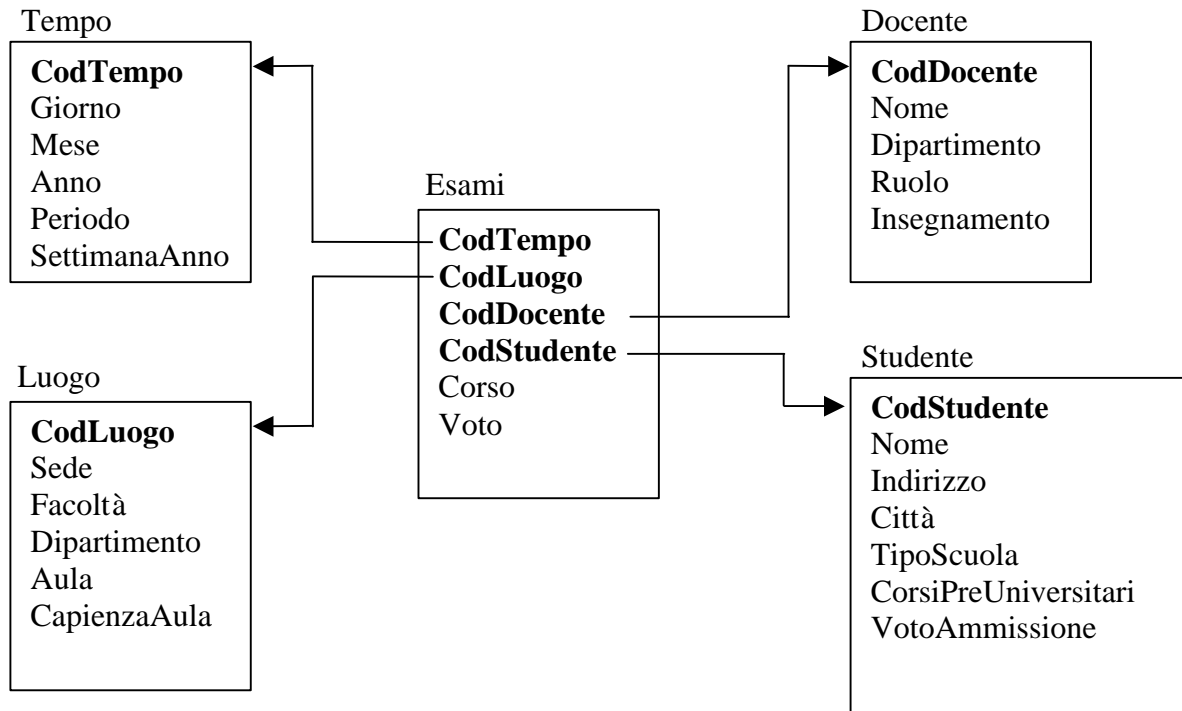


Esercizio 8.10

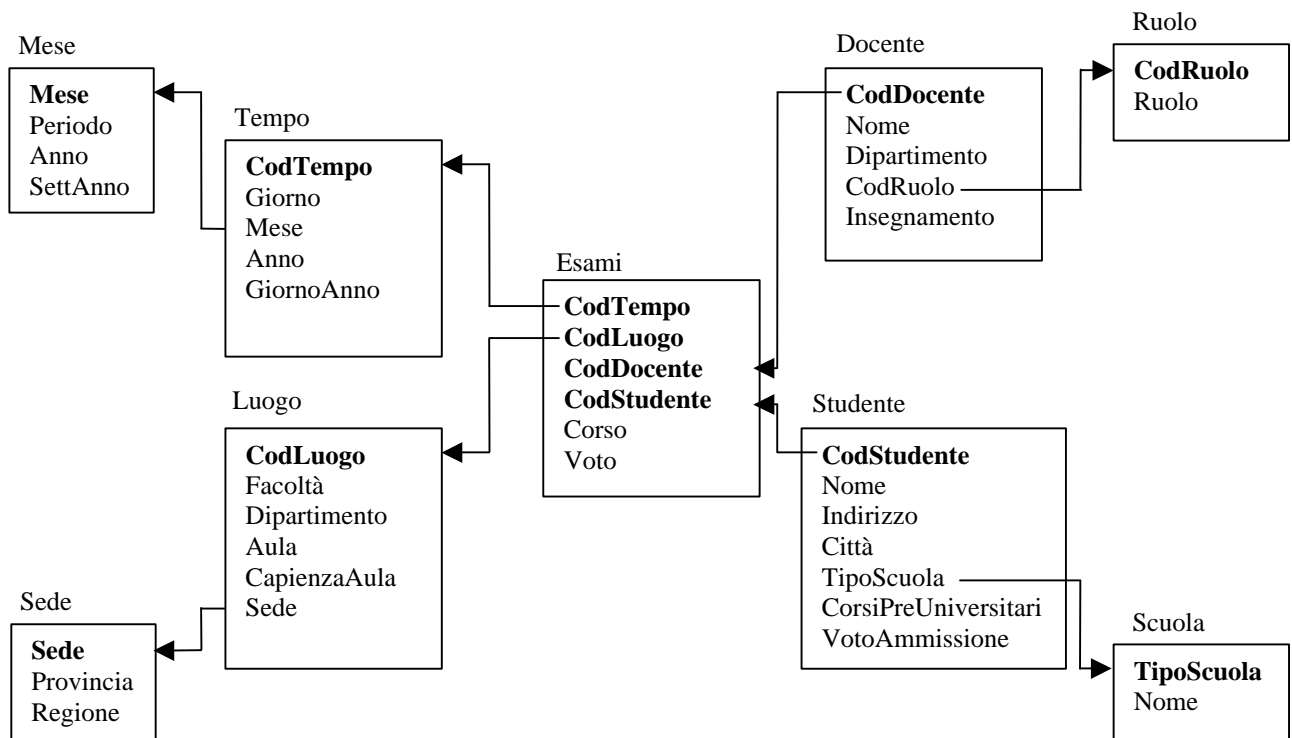
Realizzare in un sistema ROLAP il cubo definito nell'esercizio precedente definendo sia uno schema a stella sia uno schema a fiocco di neve. Indicare infine una scelta motivata di indici bitmap, indici di join e viste materializzate per gli schemi ottenuti.

Soluzione:

Schema a stella



Schema a fiocco di neve



Prima di realizzare indici o viste materializzate dobbiamo chiederci quali saranno le dimensioni delle tabelle e le operazioni svolte più di frequente sulla base dati. Alla prima domanda si può rispondere difficilmente in quanto non si possono conoscere a priori le dimensioni delle strutture che adotteranno la nostra base dati, anche se sicuramente le tabelle con più dati saranno la tabella dei fatti, Esami, quella del Tempo e quella degli Studenti.

Per quanto riguarda le operazioni svolte più di frequente, si intuisce che la tabella Esami verrà utilizzata prevalentemente per inserimenti e molto più raramente per modifiche o cancellazioni. La tabella studente verrà modificata prevalentemente nel periodo delle iscrizioni e più raramente per modifiche di tipo anagrafico.

La tabella Tempo conterrà comunque i soli dati temporali generali e quindi non vi sarà bisogno di manutenzione presumendola già impostata.

Le altre tabelle verranno invece utilizzate molto meno di frequente.

Le interrogazioni più frequenti potrebbero essere il numero di esami sostenuti dagli studenti che provengono da un determinato tipo di scuola o da un Voto di ammissione appartenente ad una determinata fascia. Il numero degli esami che ogni Docente deve correggere o il periodo dell'anno in cui vengono dati il maggior numero di esami, suddiviso per dipartimento, facoltà e sede.

La scelta degli indici bitmap in questa situazione dovrebbe ricadere su dati che vengono raramente modificati, o che vengono modificati solamente in determinate condizioni. Per la tabella Studente si potrebbero creare indici bitmap sugli attributi TipoScuola CorsiPreUniversitari e VotoAmmissione. Per la tabella Tempo si potrebbero utilizzare gli attributi Periodo, Mese e Anno.

Gli indici di join dovrebbero essere costruiti sulle chiavi delle tabelle dimensione, in modo da facilitare l'estrazione di gruppi di dati che corrispondono ad un determinato criterio anche grazie agli indici bitmap.

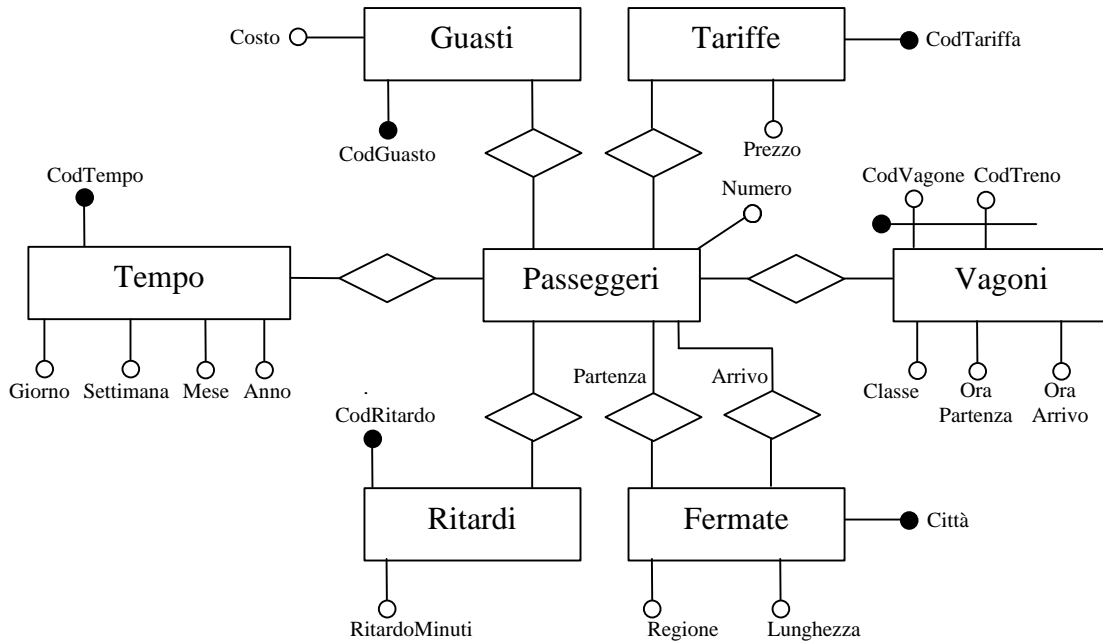
Delle viste materializzate potrebbero essere realizzate sulla tabella Esami in modo che ogni sede abbia la propria tabella dei fatti.

Esercizio 8.11

Progettare uno o più data mart relativi alla gestione delle linee ferroviarie, considerando come fatti il numero complessivo di passeggeri giornaliero per ciascuna tariffa su ciascun treno e su ciascuna tratta della rete e come dimensione le tariffe, la distribuzione geografica delle città attraversate, la composizione del treno, i guasti e i ritardi. Realizzare uno o più cubi dimensionali e darne la traduzione in forma relazionale.

Soluzione:

Ecco lo schema concettuale del data mart



Dal precedente schema concettuale si ottiene il seguente schema relazionale.

PASSEGGERI(CodTempo, CodVagone, CodTreno, CodTariffa, Città, CodRitardo, CodGuasto, Numero)

GUASTI(CodGuasto, Costo)

TARIFFE(CodTariffa, Prezzo)

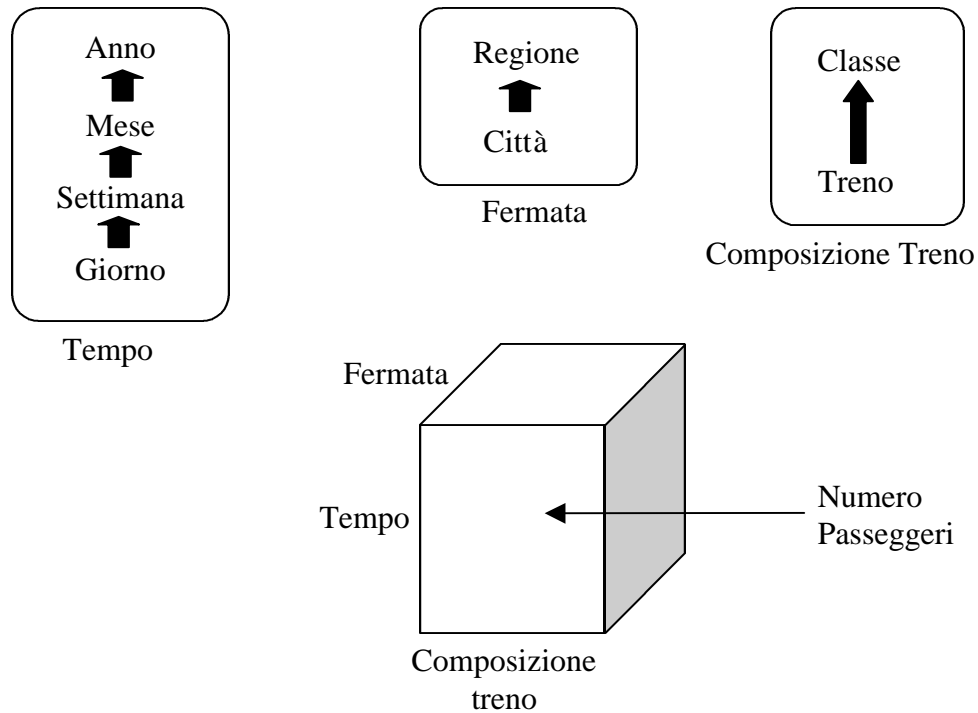
VAGONI(CodVagone, CodTreno, Classe, OraPartenza, OraArrivo)

FERMATE(Città, Regione, Lunghezza)

RITARDI(CodRitardo, RitardoMinuti)

TEMPO(CodTempo, Giorno, Settimana, Mese, Anno)

Ecco uno dei cubi dimensionali che possiamo realizzare:



Esercizio 8.12

Si consideri la tabella descritta in figura 8.21. Estrarre le regole di associazione con supporto e confidenza maggiori o uguali al 20%. Indicare poi quali regole sono estratte se si richiede invece un supporto superiore al 50%.

Soluzione:

CodTrans	Data	Oggetto	Qta	Prezzo
1	17/12/03	pantaloni-sci	1	140
1	17/12/03	scarponi	1	180
1	17/12/03	bastoncini	1	20
2	18/12/03	maglietta	1	25
2	18/12/03	Giacca	1	200
2	18/12/03	Stivali	1	70
3	18/12/03	Giacca	1	200
4	19/12/03	Giacca	1	200
4	19/12/03	maglietta	3	25
5	20/12/03	maglietta	1	25
5	20/12/03	Giacca	1	200
5	20/12/03	cravatta	1	25

Figura 8.21 Tabella che contiene dati di vendita

Soluzione:

Premessa	Conseguenza	N° Pre	N° Cons	Supporto	Confidenza
Pantaloni-sci	Scarponi	1	1	20%	100%
Scarponi	Pantaloni-sci	2	1	20%	50%
Maglietta	Giacca	3	3	60%	100%
Giacca	Maglietta	4	3	60%	75%
Maglietta	Bastoncini	3	1	20%	33%
Bastoncini	Maglietta	1	1	20%	100%
Maglietta	Scarponi	3	1	20%	33%
Scarponi	Maglietta	2	1	20%	50%
Giacca	Scarponi	4	1	20%	25%
Scarponi	Giacca	2	1	20%	50%
Giacca	Cravatta	4	1	20%	25%
Cravatta	Giacca	1	1	20%	100%
Maglietta	Cravatta	3	1	20%	33%
Cravatta	Maglietta	1	1	20%	100%
{Bastoncini, Maglietta}	{Giacca, Stivali}	1	1	20%	100%
{Giacca, Stivali}	{Bastoncini, Maglietta}	1	1	20%	100%
{Bastoncini, Giacca}	{Maglietta, Stivali}	1	1	20%	100%
{Maglietta, Stivali}	{Bastoncini, Giacca}	1	1	20%	100%
{Bastoncini, Stivali}	{Maglietta, Giacca}	1	1	20%	100%
{Maglietta, Giacca}	{Bastoncini, Stivali}	3	1	20%	33%
{Stivali, Bastoncini, Maglietta}	Giacca	1	1	20%	100%
Giacca	{Stivali, Bastoncini, Maglietta}	4	1	20%	25%
{Bastoncini, Maglietta, Giacca}	Stivali	1	1	20%	100%
Stivali	{Bastoncini, Maglietta, Giacca}	2	1	20%	50%
{Bastoncini, Giacca, Stivali}	Maglietta	1	1	20%	100%
Maglietta	{Bastoncini, Giacca, Stivali}	3	1	20%	33%
{Giacca, Stivali, Maglietta}	Bastoncini	1	1	20%	100%
Bastoncini	{Giacca, Stivali, Maglietta}	1	1	20%	100%
{Maglietta, Giacca}	Cravatta	3	1	20%	33%
Cravatta	{Maglietta, Giacca}	1	1	20%	100%
{Maglietta, Cravatta}	Giacca	1	1	20%	100%
Giacca	{Maglietta, Cravatta}	4	1	20%	25%
{Giacca, Cravatta}	Maglietta	1	1	20%	100%
Maglietta	{Giacca, Cravatta}	3	1	20%	33%

Le regole di associazione sopra estratte hanno tutte supporto e confidenza maggiori o uguali al 20%. Passando all'estrazione delle regole di associazione con supporto superiore al 50% si ottiene:

Premessa	Conseguenza	N° Pre	N° Cons	Supporto	Confidenza
Maglietta	Giacca	3	3	60%	100%
Giacca	Maglietta	4	3	60%	75%

Esercizio 8.13

Riferendosi alla tabella di figura 8.21, discretizzare i prezzi in tre valori (basso, medio e alto); trasformare i dati in modo che per ciascuna transazione si indichi con una sola tupla la presenza di una vendita di una particolare classe. Costruire poi regole di associazione che indicano la compresenza nella stessa transazione di oggetti appartenenti a differenti classi di prezzo. Infine, interpretare i risultati.

Soluzione:

Discretizziamo i prezzi in tre fasce:

Basso: prezzo ≤ 25

Medio: $25 < \text{prezzo} < 200$

Alto: prezzo ≥ 200

Nuova tabella discretizzata :

CodTrans	Data	Qta	Prezzo
1	17/12/03	2	Medio
1	17/12/03	1	Basso
2	18/12/03	1	Basso
2	18/12/03	1	Medio
2	18/12/03	1	Alto
3	18/12/03	1	Alto
4	19/12/03	1	Alto
4	19/12/03	3	Basso
5	20/12/03	2	Basso
5	20/12/03	1	Alto

Regole di associazione:

Premessa	Conseguenza	N° Pre	N° Cons	Supporto	Confidenza
Medio	Alto	2	1	20%	50%
Alto	Medio	4	1	20%	25%
Medio	Basso	2	2	40%	100%
Basso	Medio	4	2	40%	50%
Alto	Basso	4	3	60%	75%
Basso	Alto	4	3	60%	75%
{Basso, Medio}	Alto	2	1	20%	50%
Alto	{Basso, Medio}	4	1	20%	25%
{Alto, Basso}	Medio	3	1	20%	33%
Medio	{Alto, Basso}	2	1	20%	50%
{Alto, Medio}	Basso	1	1	20%	100%
Basso	{Alto, Medio}	4	1	20%	25%

Questi risultati mostrano che le relazioni più importanti sono

- Alto \rightarrow Basso
- Basso \rightarrow Alto

La maggior parte delle vendite si riferisce ad articoli di basso costo.

Significa che gli articoli di basso e alto costo sono spesso venduti assieme, mentre gli articoli con un costo medio non hanno una grande importanza per i clienti.

I risultati possono essere utili per il posizionamento degli articoli nei vari settori del supermarket.

Esercizio 8.14

Descrivere una base di dati delle vendite di automobili in cui compaiono descrizioni delle auto (auto sportive, compact, station wagon, ecc.), costo e cilindrata delle auto (discretizzati in classi) , età e reddito degli acquirenti (discretizzati in classi). Ipotizzare poi la struttura di un classificatore relativo alle propensioni all'acquisto di automobili da parte di categorie differenti di persone.

Soluzione:

Il database è composto dalle seguenti tabelle:

AUTO(CodAuto, Costruttore, Modello, Cilindrata, VelocitàMax, Categoria)

VENDITA(CodAuto, CodCliente, Colore, Optional, Prezzo, Data)

CLIENTE(CodCliente, Nome, Età, Salario)

Esempio di istanze della base dati:

Auto

<u>CodAuto</u>	Costruttore	Modello	Cilindrata	VelocitàMax	Categoria
A1	Audi	A4	2000	200	Berlina
F1	Ferrari	Enzo	3500	350	SuperSportiva
F2	Fiat	Punto	1000	160	Cittadina
S1	Smart	Smart	1000	150	Compact

Vendita

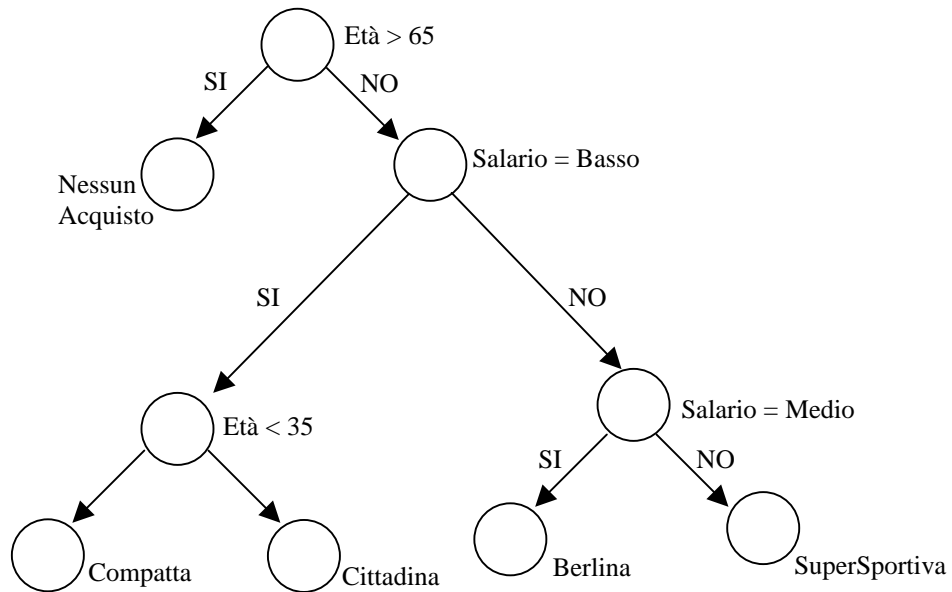
<u>CodAuto</u>	<u>CodCliente</u>	Colore	Optional	Prezzo	Data
A1	3	Grigio	Sedili in pelle	Medio	01/08/03
A1	2	Blu	Autoradio	Medio	05/08/03
S1	4	Nero		Basso	26/08/03
F1	3	Rosso	Sedili in pelle	Alto	01/09/03
F2	1	Nero		Basso	10/09/03

Cliente

<u>CodCliente</u>	Nome	Età	Salario
1	Rossi	60	Basso
2	Verdi	50	Medio
3	Bianchi	35	Alto
4	Neri	30	Basso

Classificatore:

Un ipotetico classificatore relativo alle propensioni all'acquisto di automobili da parte di categorie differenti di persone.



Esercizio 8.15

Si considerino i data mart dell'esercizio 8.11. Quali sono i problemi di data mining che vi vengono in mente.

Soluzione:

Come definito nel libro, per problemi di data mining si intende: trovare tutte le regole di associazione con supporto e confidenza superiori a valori prefissati.

In questo caso si può pensare di trovare:

- I treni con più guasti
- I treni con più ritardo
- Il ritardo dei treni in base al periodo
- Il numero di passeggeri per ogni tratta
- Il numero di passeggeri in base alla classe e al periodo

Ma dal data mart possiamo anche pensare di estrarre delle classificazioni, tipo:

- La classificazione dei ritardi in base al Tipo di treno, al periodo e alla tratta
- La classificazione dei guasti in base al Tipo di treno, al periodo e alla tratta