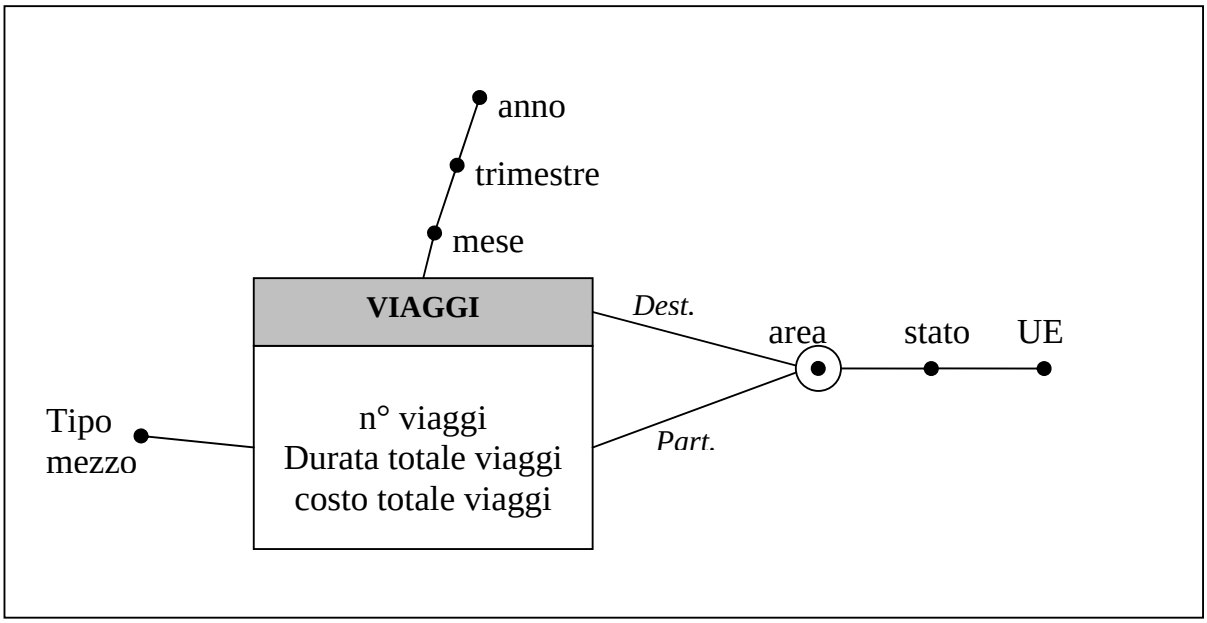
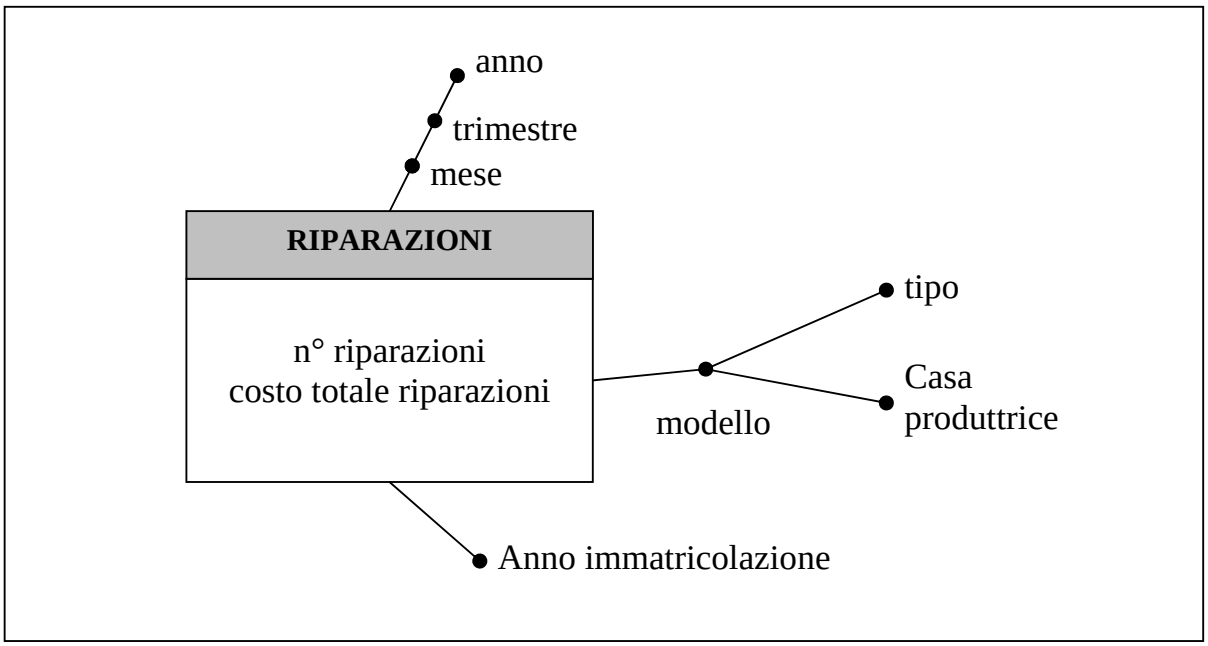
Data Base and Data Mining Group of Politecnico di Torino

Progetto di un data warehouse – BOZZA di Soluzione

Ditta trasporti

Progettazione concettuale




anno
trimestre
mese
RIPARAZIONI
n° riparazioni
costo totale riparazioni
modello
tipo
Casa
produttrice
Anno immatricolazione

Progettazione logica

Dimensioni:
AreeGeografiche(CodAreaGeo, AreaGeografica, Stato, AppartieneUE)

TipologieMezzi (CodTipologiaMezzo, TipologiaMezzo)
ModelliMezzi (CodModelloMezzo, ModelloMezzo, TipologiaMezzo, CasaProduttrice)

Tempo(CodTempo, Mese, Trimestre, Anno)

Fatti:
Viaggi(CodAreaGeoPart, CodAreaGeoDest, CodTipologiaMezzo, CodTempo, TotaleDurataViaggi, TotaleCostoViaggi, NumViaggi)

Riparazioni(AnnoImmatricolazione, CodModelloMezzo, CodTempo, NumeroRiparazioni, TotaleCostoRiparazioni)

Non serve un campo NumMedioRiparazioniMensili perché ogni riga corrisponde ad un mese e quindi NumMedioRiparazioniMensili=NumeroRiparazioni.

Una alternativa è aggiungere l'ID_mezzo alla gerarchia dei mezzi (per attaccarci direttamente la data di immatricolazione) ma questo aumenterebbe molto la cardinalità della tabella quindi è molto meglio non farlo e continuare a considerare l'"anno di immatricolazione" una dimensione a parte.

Gestione dinamicità delle dimensioni
L'unica dimensione sulla quale posso avere delle variazioni e quella relativa alle aree geografiche.

Queste possono passare da "non appartenente alla comunità europea" ad "appartenente alla comunità europea". Ciò succedere raramente, per poche tuple e può cambiare una volta sola nella vita.
Sovrascrivo semplicemente il valore del campo AppartieneUE. (Tipo 1 - "oggi per ieri")

Le aree geografiche le considero stabili. Se per caso ci sono dei cambiamenti ne definisco una o più nuove che vanno ad aggiungersi alle precedenti. Quelle precedenti (se non più valide) non vengono più usate. (Tipo 2)

Il Tipo 3 non è utilizzabile perchè un insieme di aree può essere sostituita da una sola (devo associare tutto a quella nuova). In questo caso non mi basta sovrascrivere dei valori. Devo fondere delle tuple che facevano riferimento ad aree diverse in una sola.

Posso dover anche dividere un'area preesistente in più sottoaree. In questo caso non ho modo di sapere quanto avevo venduto in passato in ognuna delle nuove aree (sono fuse nella macro area precedente).

Interrogazioni

a) SELECT Part.AreaGeografica, Dest.AreaGeografica, TM.TipologiaMezzo
SUM(TotaleDurataViaggi)/SUM(NumViaggi) as ValoreDurataMediaPerViaggio,
SUM(TotaleCostoViaggi)/SUM(NumViaggi) as ValoreCostoMedioPerViaggio,
FROM AreeGeografiche Part, AreeGeografiche Dest, Tempo T, TipologieMezzi TM, Viaggi V
WHERE V.CodAreaGeoPart=Part.CodAreaGeo
AND V.CodAreaGeoDest=Dest.CodAreaGeo
AND V.CodTempo=T.CodTempo
AND V.CodTipologiaMezzo=TM.CodTipologiaMezzo
AND T.Anno=2003
GROUP BY Part.AreaGeografica, Dest.AreaGeografica, TM.TipologiaMezzo;

b) SELECT M.ModelloMezzo,
SUM(TotaleCostoRiparazioni)/SUM(NumeroRiparazioni) as CostoMedioRiparazione,
RANK() over (ORDER BY SUM(TotaleCostoRiparazioni)/SUM(NumeroRiparazioni) DESC as RankCostoMedioRicaparazione)
FROM ModelliMezzi M, Tempo T, Riparazioni R
WHERE R.CodTempo=T.CodTempo
AND R.CodModelloMezzo=M.CodModelloMezzo
AND T.Trimestre="1-2005" **<- N.B.**
GROUP BY M.ModelloMezzo;

c) SELECT M.CasaProduttrice, R.AnnoImmatricolazione,
SUM(TotaleCostoRiparazione)/SUM(NumeroRiparazioni) as CostoMedioPerRiparazione,
SUM(NumeroRiparazioni)/COUNT(distinct mese) as NumMedioRiparazioniMensili
FROM ModelliMezzi M, Tempo T, Riparazioni R
WHERE R.CodTempo=T.CodTempo
AND R.CodModelloMezzo=M.CodModelloMezzo
AND T.Anno=2007
GROUP BY M.CasaProduttrice, R.AnnoImmatricolazione
ORDER BY NumMedioRiparazioniMensili

Per avere il numero medio di riparazioni mensili, occorre sapere quanti mesi ci sono nel gruppo che sto condensando. In questo caso i dati sono riferiti al 2007 e si possono fare alcune considerazioni:

- posso dividere direttamente per il numero dei mesi in un anno (12) se so che la ditta ha sempre lavorato tutto l'anno
- posso fare COUNT(distinct mese). Nel caso in cui in un mese non ci sia stata nessuna riparazione (e quindi il mese non compare in nessun record del db), devo ulteriormente valutare 2 casi:
    - in quel mese la ditta era chiusa e quindi è giusto che il COUNT non consideri quel mese
    - in quel mese la ditta era aperta ma per motivi casuali non sono state effettuate riparazioni; in questo caso il COUNT mi restituisce un valore diverso da quello che dovrei considerare. In questo caso posso anche

ipotizzare di aggiungere nel db dei dei record con n_riparazioni=0 relativi ai mesi in cui non ho effettuato riparazioni, in modo da non alterare la somma delle riparazioni ma di far comparire cmq esplicitamente il mese, in modo che sia considerato nel COUNT

VISTE

cardinalità VIAGGI
200 (aree partenza) *200 (aree arrivo) * 3x12 (anni x mesi) *10 (tipo dei mezzi di trasporto) = 14.4 M

Cardinalità RIPARAZIONI

36 (mesi x anni) *50 (modelli) *10 (anni di immatricolazione) = 18k

query	Group by	predicati
A	Area_part, area_arr, tipo_mezzo	Anno
B	Modello_mezzo	trimestre
C	Casa_prod, anno_immatr	Anno
D	Area_part, area_arr, tipo_mezzo	anno
e	Casa_prod	Anno

Query a) 200*200*10 = 400k (ben al di sotto del numero di tuple del fatto Viaggi quindi conviene generare una vista)

Area_partenza X Area_arrivo X tipo_mezzo X Anno

Query b) 50 (conviene creare una vista)

Query c) 10*10 = 100 (conviene creare una vista)

Casa_prod X Anno_imm X Anno

Query d) è molto simile alla query a). a questo punto genero una sola vista per la query a) e d) per rispondere a entrambe le richieste (800k record)

Query e) posso rispondere con la stessa vista della query c)