Esercizio: Progettazione di un data warehouse

Descrizione del problema

Una società di treni vuole analizzare i dati relativi ai movimenti dei treni passeggeri sul territorio italiano.
Ogni viaggio è caratterizzato da una stazione di partenza e una di arrivo, un'ora di partenza ed un'ora di arrivo "prevista" e dal treno che lo compie, Ogni treno ha un codice univoco ed è caratterizzato da un tipo (treno notte, treno ad alta frequentazione, treno alta velocità, etc.), da un modello e da una azienda produttrice (ogni modello è prodotto da una sola casa produttrice) e dalla presenza o meno di alcuni allestimenti (es. ristorante). Il numero dei posti disponibili sul treno varia invece a seconda del viaggio considerato. La stazione ferroviaria è caratterizzata, a seconda del traffico che vi circola, da un livello: internazionale, metropolitana, locale, etc. .

La dirigenza della società è interessata ad analizzare per ogni viaggio: il numero di persone, la percentuale di posti occupati sul treno sul totale dei posti disponibili, l'incasso effettuato dalla vendita dei biglietti, gli eventuali minuti di ritardo all'arrivo del treno e il numero di persone di servizio a bordo treno, in funzione:

- Per quanto riguarda sia la partenza che l'arrivo è necessario conoscere:
  - la data e il timestamp (espresso in ore più minuti) previsti
  - l'ora
  - il mese, trimestre, quadrimestre e anno
  - il giorno della settimana e se era festivo o meno
  - la stagione (primavera, estate, autunno, inverno)
  - il periodo del giorno (mattina dalle 6 alle 12, pomeriggio dalle 13 alle 18, …)
  - la stazione
  - la città, provincia e regione in cui è sita la stazione
  - il livello della stazione
- Per quanto riguarda il treno si è interessati a conoscere:
  - Il codice univoco del treno
  - il tipo di treno
  - il modello e l'azienda produttrice
  - la presenza o no di: vagone ristorante, cuccette, vagone letto e carrozze di prima classe.

Il data warehouse realizzato deve contenere le informazioni relative agli anni 1998-2008. Al fine di una corretta realizzazione del data warehouse sono state fornite le seguenti informazioni (le informazioni ritenute necessarie ma non presenti in questo elenco possono essere ipotizzate e stimate dal candidato):

- Numero di stazioni ~ 5000
- Numero di treni ~ 10000

I dirigenti vogliono poter disporre delle seguenti informazioni:

a) Trovare l'incasso medio mensile effettuato nel 2008.
b) Per ogni coppia tipo di treno e mese del 2008, visualizzare il mese, il tipo di treno e la percentuale di posti occupati sul totale dei posti disponibili considerando solo i treni in cui è presente il vagone ristorante. Inoltre, visualizzare la posizione in una graduatoria (rank) dalla coppia con la percentuale più alta a quella con la percentuale più bassa e ordinare i dati secondo la graduatoria.
c) Trovare il numero medio di minuti di ritardo accumulati dai treni prodotti dalla "Alstom Transport" che hanno viaggiato da Milano a Torino nel mese di maggio 2008
d) Per il mese di Ottobre 2005 calcolare il numero totale di minuti di ritardo accumulati dai treni di tipo "Treno ad alta Frequentazione"
e) Per ogni tipo di treno calcolare l'incasso effettuato in tutti i mesi del 2007
f) Considerando solo i viaggi in cui la data della partenza coincide con la data dell'arrivo e considerando solo i treni che a Settembre 2008 hanno incassato in totale (su tutte le corse effettuate) più di 10000 euro, trovare per ogni giorno festivo di Ottobre 2008 l'incasso effettuato

Progettazione

1) Progettare il data warehouse necessario per soddisfare le richieste descritte nelle specifiche del problema. Il data warehouse progettato deve inoltre permettere di rispondere in modo efficiente a **tutte** le interrogazioni frequenti proposte nelle specifiche del problema.

2) Esprimere le interrogazioni frequenti (b), (c), (f) delle specifiche del problema utilizzando il linguaggio SQL esteso.

3) Considerando le caratteristiche del data warehouse realizzato e la cardinalità dei dati memorizzati nel data warehouse, decidere se e quali viste materializzate potrebbe essere utile definire al fine di ottimizzare i tempi di risposta delle interrogazioni proposte nelle specifiche del problema (considerare **tutte** le interrogazioni proposte e non solo quelle risolte in SQL al punto 2). Motivare le scelte fatte.

4) Decidere come gestire la dinamicità (variazione) dei dati all'interno delle dimensioni.