

Progettazione dimensionale

Luca Cabibbo
marzo 2010

Progettazione dimensionale

La progettazione dimensionale è la progettazione logica dei dati del data warehouse, basata sull'architettura a bus

- progettazione dei data mart
- progettazione di un insieme di dimensioni conformi
- progettazione degli schemi dimensionali
- analisi delle sorgenti informative
 - comprensione delle sorgenti informative disponibili e delle loro qualità
 - progettazione preliminare del mapping dei dati dalle sorgenti informative al data warehouse
- piano preliminare delle aggregazioni

Lo schema logico del data warehouse è la pietra angolare della progettazione dell'intero data warehouse

Progettazione dei data mart

Un data warehouse dimensionale viene progettato come un insieme coerente di data mart

- un **data mart** è
 - un sottoinsieme logico dell'intero data warehouse
 - è la restrizione del data warehouse a un singolo processo dell'organizzazione, o a un insieme di attività correlate
 - una collezione di fatti correlati che devono essere analizzati insieme
 - un insieme di schemi dimensionali correlati

Un insieme di data mart può essere considerato “coerente” se è organizzato secondo una architettura a bus basata su dimensioni conformi e fatti conformi

- ovvero, su dimensioni e fatti con significato uniforme in tutto il data warehouse

3

Progettazione dimensionale

Luca Cabibbo

Come iniziare?

E' possibile iniziare la progettazione di un data warehouse dimensionale adottando un approccio top-down

- identifica i data mart candidati
- identifica le dimensioni implicate da tali data mart
- prosegui con la progettazione specifica dei data mart solo dopo un accordo preliminare su questi due aspetti

4

Progettazione dimensionale

Luca Cabibbo

Identifica i data mart candidati

Identificazione dei data mart candidati

- il criterio principale è
 - un data mart deve rappresentare una collezione di fatti correlati che devono essere analizzati insieme
- ad esempio, in una banca i fatti di interesse relativi alle attività di un conto corrente potrebbero essere
 - importi depositati – importi prelevati
 - canoni/imposte – interessi
 - numero di transazioni
 - numero di persone in attesa allo sportello
- dunque, in prima approssimazione, un data mart è un gruppo di fatti che possono essere usati insieme a fini decisionali

Identifica i data mart candidati

Un data mart può essere relativo ad uno oppure più attività/processi aziendali

- alcuni data mart – **data mart di primo livello** – hanno origine in un singolo processo dell'organizzazione ed in una singola sorgente informativa
- altri data mart – **data mart consolidati** – sono invece relativi a più processi e/o con dati derivanti da più sorgenti informative
- i data mart possono essere (parzialmente) sovrapposti
- in una grande organizzazione dovrebbe essere possibile identificare da 10 a 30 data mart

Identifica i data mart candidati

Lavoro da fare per sostenere la candidatura di un data mart

- quali sono le necessità di analisi dell'organizzazione? quali corrispondenze tra le necessità di analisi (report) ed i data mart candidati?
- quali sono le sorgenti informative disponibili all'organizzazione? quali corrispondenze tra le sorgenti informative disponibili ed i data mart candidati?

Esempio — una grande compagnia telefonica

Data mart a sorgente singola

- fatturazione clienti (residenziali e commerciali)
- gestione ordini
- gestione dei malfunzionamenti
- pubblicità sulle pagine gialle
- servizio clienti e informazioni sulle fatture
- offerte promozionali e comunicazioni ai clienti
- dettaglio delle chiamate dal punto di vista della fatturazione
- dettaglio delle chiamate dal punto di vista del carico della rete telefonica
- inventario clienti
- inventario della rete telefonica
- ...

Selezione dei data mart

La realizzazione di un data warehouse deve iniziare da un data mart con le seguenti caratteristiche

- essere significativo
 - ovvero, permettere analisi significative
- essere semplice da realizzare
 - in particolare, essere a sorgente singola

Successivamente, possono essere realizzati altri data mart, più complessi

- ad esempio, a sorgente multipla
 - come il data mart della profittabilità dei clienti

Identifica le dimensioni candidate

Scelti i data mart di interesse, si procede identificando ed elencando le dimensioni (candidate) di interesse

- bisogna progettare un insieme di dimensioni da usare in modo conforme (o conformato) in tutti i data mart del data warehouse
- si può iniziare identificando le dimensioni di interesse per ciascun data mart
- utile mostrarle in una matrice data mart/dimensioni

Esempio — una grande compagnia telefonica

Dimensioni per il data mart della fatturazione clienti

- tempo (data di fatturazione)
- cliente (residenziale o commerciale)
- servizio
- tariffa (compresa promozione)
- fornitore di servizi locali

Dimensioni per il data mart del dettaglio delle chiamate dal punto di vista della fatturazione

- chiamante
- chiamato
- fornitore di servizi non locali

La matrice dell'architettura a bus

I data mart e le dimensioni possono essere utilmente correlati in una matrice che descrive l'architettura a bus del data warehouse

- ciascuna riga della matrice rappresenta un data mart
- ciascuna colonna della matrice rappresenta una dimensione
- ciascun elemento della matrice, all'intersezione di un data mart e di una dimensione, viene marcato se la dimensione è di interesse per il data mart

La definizione della matrice che descrive l'architettura a bus del data warehouse è una “pietra miliare” fondamentale nella progettazione dell'intero data warehouse

- è il luogo dove viene fissato l'insieme delle dimensioni conformi del data warehouse

Esempio — una grande compagnia telefonica

	Time	Customer	Service	Rate Category	Local Service Provider	Calling Party	Called Party	Long-Distance Provider	International Organization	Employee	Location	Equipment Type	Supplier	Item Supplied	Weather	Account Status
Customer billing	X	X	X	X	X			X			X					X
Service orders	X	X	X		X			X	X	X	X	X			X	X
Trouble reports	X	X	X		X	X		X	X	X	X	X	X	X	X	X
Yellow Page Ads	X	X		X		X			X	X	X					X
Customer Inquiries	X	X	X	X	X	X		X	X	X	X				X	X
Promotions & Comm'n	X	X	X	X	X	X		X	X	X	X	X	X	X		X
Billing Call Detail	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
Network Call Detail	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
Customer Inventory	X	X	X	X	X			X	X		X	X	X	X		X
Network Inventory	X		X						X	X	X	X	X	X		
Real Estate	X								X	X	X	X				
Labor & Payroll	X								X	X	X					
Computer Charges	X	X	X		X			X	X	X	X	X	X	X		
Purchase Orders	X								X	X	X	X	X	X		
Supplier Deliveries	X								X	X	X	X	X	X		
Combined Fields Ops.	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X
Customer Reln. Mgmnt.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Customer Profit	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

13

Progettazione dimensionale

Luca Cabibbo

La matrice dell'architettura a bus

La creazione di dimensioni conformi è frutto di una decisione tanto politica quanto tecnica – e deve essere sostenuta dai livelli esecutivi più alti

- la matrice dell'architettura a bus è una specie di mappa del processo politico che è stato intrapreso per far in modo che tutti i gruppi coinvolti siano d'accordo sulle definizioni comuni delle dimensioni

14

Progettazione dimensionale

Luca Cabibbo

Progettazione delle tabelle fatti

È poi possibile procedere, nell'ambito di ciascun data mart, con la progettazione delle tabelle fatti

- la progettazione di ciascuna tabella fatti avviene in quattro fasi
 - scegli il processo
 - scegli la grana
 - che cosa rappresenta un record della tabella fatti?
 - scegli le dimensioni
 - dimensioni primarie (fissate dalla grana) e dimensioni supplementari
 - scegli i fatti
 - numerici o fittizi

Progettazione degli schemi dimensionali

Successivamente, va completato il progetto degli schemi dimensionali

- selezione degli attributi delle dimensioni
- scelta della strategia di gestione dei cambiamenti lenti, per ciascuna dimensione
- altre scelte di rappresentazione
 - minidimensioni, dimensioni e fatti eterogenei, aggregazioni
- durata storica del data warehouse
 - quanti dati storici devono essere rappresentati nel data warehouse? con quale grana?
- pianificazione del caricamento incrementale
 - con che periodicità deve essere aggiornato il data warehouse? con che urgenza?

Convenzioni nella progettazione

Alcune indicazioni stilistiche (e non) da adottare nella progettazione

- i nomi (etichette) per data mart, dimensioni, attributi e fatti devono essere scelti attentamente nel dominio applicativo del data warehouse
 - devono essere nomi accettabili per gli utenti finali
- ogni attributo vive in una sola dimensione, un fatto può essere ripetuto in più tabelle fatti
- se una dimensione deve essere ripetuta, probabilmente indica ruoli diversi della stessa dimensione e, quindi, dimensioni diverse
 - ad esempio, data del servizio e data di scadenza della fattura

Convenzioni nella progettazione

- i campi significativi delle sorgenti informative corrispondono a uno o più campi del data warehouse
 - ad esempio, un campo prodotto può essere rappresentato dal codice del prodotto, descrizione sintetica, descrizione completa
- ogni fatto dovrebbe essere associato a una modalità di aggregazione di default
 - ad esempio, somma, minimo, massimo, ultimo valore, semi additivo, algoritmo speciale, non additivo, ...
- è opportuno evidenziare nelle dimensioni le eventuali gerarchie di aggregazione significative per l'utente

Analisi delle sorgenti informative

Progettato lo schema logico del data warehouse, bisogna

- descrivere le sorgenti informative a disposizione
 - ovvero, le sorgenti informative individuate nella fase di raccolta e analisi dei requisiti
- progettare la trasformazione dei dati dalle sorgenti informative al data warehouse

Descrizione delle sorgenti informative

Alcune caratteristiche delle sorgenti informative

- nome della sorgente informativa
- responsabile
- responsabile tecnico
- piattaforma hardware e software
- ubicazione
- breve descrizione

Queste caratteristiche non sono sufficienti per la selezione delle sorgenti informative

Selezione delle sorgenti informative

Il criterio principale per la selezione delle sorgenti informative da cui estrarre i dati per il data warehouse è relativo all'accuratezza dei dati

- uno stesso dato può attraversare più sistemi, per essere elaborato in più modi
 - il transito dei dati da un sistema all'altro avviene insieme a delle trasformazioni, che arricchiscono o sintetizzano i dati originari
- in generale, la qualità di un dato può diminuire allontanandosi dal sistema in cui è stato immesso o generato
 - è quindi opportuno catturare i dati quando vengono generati (possibilmente, dopo che sono stati puliti)

Reverse engineering delle sorgenti informative

Le sorgenti informative selezionate per l'estrazione devono essere comprese in dettaglio

- è opportuno l'uso di modelli formali per descrivere la struttura dei dati
 - schema logico dei dati
 - schema concettuale dei dati
 - glossario dei dati
- gli schemi concettuali, se non sono disponibili, possono essere ottenuti mediante una attività di reverse engineering dei dati
 - orientata appunto alla comprensione e descrizione concettuale delle sorgenti informative

Progettazione della trasformazione dei dati

Per ciascun elemento (record e campo) del data warehouse bisogna progettare la trasformazione necessaria a calcolare l'elemento dalle sorgenti informative

- descrivendo per ciascun dato
 - il ruolo nel data warehouse
 - la sorgente (o le sorgenti) da cui viene estratto
 - le trasformazioni necessarie

Piano delle aggregazioni

Ogni data warehouse contiene dati pre-aggregati

- la disponibilità di dati pre-aggregati è lo strumento singolo più efficace nel controllo delle prestazioni delle attività di interrogazione del data warehouse
- i dati aggregati devono essere rappresentati in tabelle fatti apposite, separate dalle tabelle fatti da cui sono calcolati
 - usando anche tabelle dimensione aggregate, contratte e conformate
- i dati effettivamente aggregati possono cambiare nel corso del tempo
 - le aggregazioni deve essere gestite mediante un "navigatore" e metadati opportuni
- un piano preliminare delle aggregazioni è comunque utile, ad esempio nella stima dell'occupazione di memoria

Discussione

Come in tutte le attività di progettazione, le fasi di una metodologia non vengono mai eseguite in una sequenza perfetta

- spesso, lo svolgimento di una fase richiede la correzione di scelte fatte nei passi precedenti
 - ad esempio, se la scelta delle dimensioni portasse a una grana diversa per uno schema dimensionale, o se fosse impossibile estrarre dei dati dalle sorgenti informative
- in alcuni casi, può essere opportuno avviare una fase anche se la fase immediatamente precedente non è stata conclusa
 - ad esempio, iniziare la progettazione di alcuni data mart anche quando la selezione dei data mart non è stata completata