IL DATA WAREHOUSE

Slide 2 – Sommario

Benvenuti!

In questa lezione analizzeremo innanzitutto il **cubo multidimensionale** che corrisponde al modello di riferimento per l'analisi dei dati e approfondiremo anche gli eventi secondari e l'aggregazione dei dati. Vedremo, poi, come vengono rappresentati i cubi multidimensionali nei sistemi relazionali ed esamineremo alcuni aspetti legati alla progettazione di un Data Warehouse.

Cominciamo…

Slide 3 – Il modello multidimensionale

Il modello multidimensionale è il modello alla base della rappresentazione e per l'interrogazione delle informazioni in un Data Warehouse. I fatti di interesse per l'analisi sono rappresentati in cubi dove ogni cella memorizza:

- le **misure numeriche** che quantificano il fatto da diversi punti di vista;
- inoltre, ogni **asse è una dimensione** per l'analisi dei valori delle misure;
- e, ancora, ogni dimensione può essere la radice di una **gerarchia di attributi** utilizzati per aggregare i valori delle misure, corrispondenti a diversi livelli di granularità.

Vediamo un esempio per discutere questi concetti un po' più in dettaglio.

Slide 4– Il cubo delle vendite

L'immagine mostra un esempio di cubo, relativo alle vendite. Ogni cella rappresenta le vendite di un prodotto in una determinata data e in un determinato punto vendita.

La cella blu corrisponde al fatto che 10 confezioni di shampoo BelliCapelli siano state vendute nel punto vendita MioSuper il 10 maggio 2022.

Le coordinate corrispondenti sono indicate sugli assi.

Slide 5 – Fatti

Un **fatto** è un **concetto rilevante** per i processi decisionali. Modella un insieme di eventi che si svolgono all'interno di un'organizzazione. Ha proprietà dinamiche o si evolve, in qualche modo, nel tempo.

Nel caso in esempio sono le vendite (di un prodotto in un punto vendita in una data).

Un fatto corrisponde a una cella del cubo.

Le celle trasparenti (tratteggiate) corrispondono a eventi che non si sono realizzati (quel prodotto non è stato venduto in quel punto vendita in quella data): questo corrisponde a dire che il cubo è **sparso**.

Notiamo che i fatti corrispondono tipicamente ad un'**aggregazione** rispetto alle transazioni di vendita (una cella aggrega le informazioni relative a più transazioni di vendita).

Slide 6 – Misure

Una **misura** è una **proprietà numerica** di un fatto e descrive una proprietà **quantitativa** rilevante per l'analisi. È un indicatore della rilevanza del fatto (dell'evento corrispondente).

Ad esempio, ogni vendita è quantificata dal numero di prodotti venduti e dal ricavo corrispondente.

La cella blu riporta il valore della misura quantità (10).

Slide 7 – Dimensioni

Una **dimensione** è una proprietà di un fatto con un **dominio finito** e descrive una **prospettiva**, un punto di vista, da cui analizzare il fatto.

Le dimensioni tipiche per il fatto di vendita sono:

- i prodotti (cosa è stato venduto?);
- i negozi (dove è stata effettuata la vendita?);
- le date (quando è stata effettuata?).

Le dimensioni **corrispondono agli assi dello spazio** e i valori delle proprietà corrispondenti sono le **coordinate delle celle**. Benché si parli di cubo, precisiamo che le coordinate non sono necessariamente tre come nell'esempio. Le dimensioni possono essere anche decisamente di più. Nel caso delle vendite altre dimensioni rilevanti potrebbero essere, ad esempio, chi ha effettuato l'acquisto (cliente) o come è stato effettuato il pagamento.

Slide 8 – Attributi dimensionali

Con il termine generico **attributi dimensionali** si intendono le **dimensioni** e gli altri possibili **attributi**, sempre con valori discreti, che le descrivono. Ad esempio, un prodotto è descritto dalla sua tipologia, dalla categoria di appartenenza, oltre che dal suo marchio.

Gli attributi dimensionali relativi a una stessa dimensione sono legati da associazioni **uno-a-molti** e da **dipendenze funzionali**.

Ogni prodotto ha un'unica tipologia e un unico marchio, mentre ci possono essere più prodotti della stessa tipologia e dello stesso marchio. Analogamente, una tipologia di prodotti appartiene a una sola categoria e una categoria include più tipologie.

In termini di **dipendenze funzionali** diciamo che il prodotto determina funzionalmente la sua tipologia: fissato un prodotto sappiamo a che categoria appartiene.

Slide 9 – Gerarchie

In questo modo possono venire a determinarsi, nelle diverse dimensioni, delle **gerarchie**.

Gli **attributi dimensionali** corrispondono ai diversi **livelli di dettaglio** rispetto a cui osservare i fatti dallo stesso punto di vista.

Spostarsi da un attributo dimensionale a un altro corrisponde a effettuare un «**cambio di scala**» nello spazio multidimensionale o a un **diverso livello di zoom** nell'osservare i fatti.

Possiamo cioè osservare le vendite più «da lontano», andando a unire in un'unica cella tutte le vendite di shampoo o tutte le vendite effettuate in Liguria.

Rispetto all'esempio precedente, ricordiamo che possono esserci diversi modi, indipendenti, di generalizzare, cioè guardare più da lontano un prodotto: la tipologia o il «marchio».

La gerarchia ha a un estremo (il più dettagliato) la dimensione e all'altro unisce in un unico punto tutto l'asse, cioè considera un unico valore «tutti i prodotti».

Slide 10 – Gerarchie e aggregazione

Le **gerarchie** definiscono il modo in cui è possibile **aggregare gli eventi** primari e selezionarli in modo efficace per i processi decisionali.

Possiamo infatti voler analizzare le vendite effettuate per mese, città e tipologia di prodotto o magari siamo solo interessati alle vendite di shampoo a Genova nel maggio 2022. Nelle nostre analisi, non tutte le dimensioni sono necessariamente di interesse: ad esempio possiamo non differenziare rispetto al negozio e così stiamo andando a ignorare la dimensione «dove» è stata effettuata la vendita e ad associare tutte le vendite al valore «tutti i negozi».

Mentre la **dimensione** alla radice di una **gerarchia** definisce la sua **granularità di aggregazione** più fine, gli altri attributi dimensionali corrispondono a una granularità gradualmente più grossolana.

Notiamo che — poiché la radice della gerarchia è la granularità più fine — è cruciale individuarla correttamente perché determina il livello di dettaglio delle analisi che saremo in grado di effettuare. Nel cubo con radice data per la dimensione temporale, ad esempio, non saremo in grado di analizzare le vendite rispetto all'orario.

Slide 11 – Eventi primari e secondari

Le celle, le cui coordinate si riferiscono alla **radice della gerarchia** (che ricordiamo essere l'attributo dimensionale dalla granularità più fine, cioè al massimo livello di dettaglio) corrispondono al livello massimo di disaggregazione a cui possiamo analizzare i nostri dati.

Sono dette **eventi primari**.

I cubi che si ottengono aggregando rispetto ad altri attributi dimensionali sono detti **eventi secondari** e possono essere derivati a partire dagli eventi primari.

Gli insiemi di attributi dimensionali rispetto alla relazione di ordine parziale «più dettagliato di» costituiscono un lattice.

Slide 12 – Misure e eventi secondari

I **valori** delle **misure** di un **evento secondario** sono calcolati a partire da quelli delle misure degli **eventi primari** corrispondenti.

Le vendite mensili per tipo di prodotto in una città sono calcolate a partire da quelle dei prodotti di quel tipo, venduti nei punti vendita di quella città, nelle date di quel mese.

Analogamente, le vendite per tipo di prodotto e per mese possono essere individuate aggregando tutte le vendite indipendentemente da dove sono state effettuate.

Slide 13 – Misure e aggregazione

L'**aggregazione** richiede la definizione di un **operatore idoneo** a comporre i valori di misura che contrassegnano gli eventi primari in valori da assegnare agli eventi secondari.

Da questo punto di vista, le misure possono essere classificate in tre categorie:

- le **misure di flusso** che si riferiscono ad un arco temporale, al termine del quale, vengono valutate cumulativamente (numero di prodotti venduti in un giorno, incassi mensili, numero di nascite annuali).
- Le **misure di livello** sono valutate in momenti particolari (il numero di prodotti in inventario, il numero di abitanti in una città).
- Le **misure unitarie** sono valutate in momenti particolari, ma sono espresse in termini relativi (prezzo unitario del prodotto, percentuale di sconto, cambio valuta).

Le misure di flusso possono essere aggregate mediante gli operatori somma, media, minimo e massimo rispetto a tutte le gerarchie. Per le misure di livello vale lo stesso rispetto alle gerarchie non temporali, mentre l'operazione somma non può essere utilizzata per aggregare rispetto a dimensioni temporali. Per le misure unitarie, la somma non produce mai aggregazioni significative.

A seconda dell'operatore utilizzabile le misure sono anche dette **additive** (quando puoi utilizzare l'operatore SUM per aggregarne i valori lungo ciascuna dimensione), **semi-additive** (quando l'operatore SUM può essere utilizzato per alcune dimensioni, ma non tutte), **aggregabili** (quando possono essere utilizzati altri operatori, ma non la somma), **non aggregabili** (quando non è possibile utilizzare alcun operatore).

Slide 14 – Modello multidimensionale – altri attributi

Per caratterizzare completamente le dimensioni possono, inoltre, essere necessari:

- Attributi **descrittivi**: memorizzano informazioni aggiuntive su un attributo dimensionale. Non vengono utilizzati per l'aggregazione perché hanno dominio denso o sono in associazione uno-a-uno (es. indirizzo o telefono di un punto vendita);
- Attributi **cross-dimensionali**: i cui valori sono definiti dalla combinazione di due o più attributi dimensionali, eventualmente appartenenti a diverse gerarchie;
- Attributi **condivisi** da diverse gerarchie o porzioni di gerarchie condivise tra più dimensioni;
- Attributi **opzionali,** gerarchie incomplete;
- Associazioni **molti-a-molti** tra attributi dimensionali.

Slide 17 – ROLAP, MOLAP, HOLAP

I sistemi di Data Warehouse utilizzano diversi modelli.

In particolare, si parla di sistemi **ROLAP** (relational OLAP) se basati su tecnologia relazionale. Il vantaggio principale è che si tratta di un modello molto conosciuto e di una tecnologia consolidata.

Si parla invece di **MOLAP** (multidensional OLAP) se sono basati sul modello multidimensionale (che fa riferimento proprio alle celle del cubo. In questo contesto, la sparsità degli eventi primari può essere un fattore da considerare).

Esistono infine soluzioni ibride, dette **HOLAP** (hybrid OLAP), ad esempio, che utilizzano modelli diversi per cubi primari e quelli secondari.

A seguire, ci concentreremo sugli schemi ROLAP.

Slide 16 – Schema a stella

La modellazione multidimensionale nei sistemi relazionali si basa sul cosiddetto **schema a stella** e sulle sue varianti. Uno schema a stella è costituito da:

- una serie di **tabelle dimensionali** ciascuna corrispondente a una **dimensione**,
- e una **tabella dei fatti**.

Ogni tabella delle dimensioni contiene:

- una **chiave primaria** (tipicamente surrogata) (nell'esempio IDNegozio per la tabella DTNegozio)
- un insieme di **attributi dimensionali** corrispondenti a diversi livelli di aggregazione (nell'esempio negozio, città, regione, responsabile).

La tabella dei fatti include un attributo per ogni misura (nell'esempio Quantità e Ricavo) e una chiave esterna per ogni tabella delle dimensioni, che fa riferimento alla tabella delle dimensioni (nell'esempio, IDNegozio, IDProdotto e IDData).

La chiave primaria della tabella dei fatti è l'insieme delle chiavi esterne.

Slide 17

Le tabelle delle dimensioni sono de-normalizzate, a causa della transitività delle dipendenze funzionali. Per ogni negozio in una certa città viene ripetuto in che regione si trova quella città.

Ogni informazione è recuperabile con un'unica operazione di *join* a partire dalla tabella dei fatti.

Il nome schema a stella è proprio dovuto al fatto che le tabelle delle dimensioni sono disposte (come «raggi») intorno alla tabella dei fatti, che rappresenta il centro della stella.

Slide 18 – Schema a fiocco di neve

Lo schema del **fiocco di neve** riduce la de-normalizzazione delle tabelle dimensionali presente in uno schema a stella rimuovendo alcune delle dipendenze transitive. Le tabelle delle dimensioni in uno schema a fiocco di neve sono caratterizzate da:

- una chiave primaria (tipicamente surrogata);
- il sottoinsieme di attributi dimensionali determinato funzionalmente da tale chiave;
- zero o più riferimenti di chiave esterna ad altre tabelle delle dimensioni, necessari per garantire la ricostruibilità delle informazioni.

Le tabelle delle dimensioni le cui chiavi sono importate nella tabella dei fatti sono dette primarie altrimenti sono dette tabelle dimensionali secondarie.

Nell'esempio sono tabelle delle dimensioni secondarie quelle relative a Città e Tipo.

Slide 19

Lo spazio necessario è ridotto grazie alla normalizzazione, anche se è necessario inserire nuove chiavi surrogate che consentano di determinare le corrispondenze tra tabelle dimensionali primarie e secondarie.

Il tempo di esecuzione delle interrogazioni che coinvolgono gli attributi delle tabelle delle dimensioni secondarie aumenta, perché richiedono un ulteriore join.

L'esecuzione di interrogazioni che coinvolgono solo gli attributi contenuti nella tabella dei fatti e nella tabella delle dimensioni primarie è invece più rapida, perché le tabelle sono più piccole.

La scelta degli attributi su cui effettuare snowflaking è parte della progettazione. Lo **schema a fiocco di neve** può essere particolarmente utile in presenza di **eventi secondari memorizzati**.

Slide 20 – Eventi secondari e viste

Gli eventi secondari possono essere infatti essere precalcolati utilizzando viste materializzate.

Questo permette di eseguire interrogazioni OLAP accedendo a cubi secondari di dimensioni molto minori. La scelta delle viste da materializzare viene effettuata bilanciando lo spazio richiesto (e i relativi tempi di aggiornamento al caricamento di nuovi dati) e i vantaggi in termini di velocizzazione delle interrogazioni OLAP.

Slide 21 – Altri schemi relazionali per Data Warehouse

Un altro schema frequentemente utilizzato è lo **schema a costellazione** che prevede più tabelle dei fatti che possono memorizzare fatti primari (ad esempio: vendite, spedizioni, promozioni, ecc.) o secondari (ad esempio: vendite pre-aggregate per mese).

In alcuni casi possono essere introdotte delle tabelle *bridge table* per gestire associazioni molti a molti o attributi cross-dimensionali.

Infine, per dimensioni degeneri (cioè costituite da un unico attributo dimensionale) si può avere una gestione ad-hoc mediante *junk table*.

Slide 22 – Progettazione di un Data Warehouse

Costruire un Data Warehouse è un'attività molto complessa che coinvolge problematiche e rischi sia di natura organizzativa che architetturale.

Per minimizzare i rischi sono state sviluppate metodologie e sono possibili diversi approcci: sia top down sia bottom-up, a seconda che si parta dalla progettazione dell'intero Data Warehouse per ottenere i Data Mart o che, viceversa, si inizi con la progettazione dei Data Mart.

La progettazione è *supply driven* se parte dall'analisi e dalla riconciliazione delle sorgenti dati operazionali, mentre è *demand driven* se parte con l'analisi dei requisiti utente.

Un aspetto importante è la qualità dei dati di partenza ed è fondamentale coinvolgere nella progettazione chi ha conoscenza delle sorgenti operazionali.

Slide 23 – Riepilogo e conclusioni finali

Bene, siamo giunti alla fine di questa video lezione.

Ti ricordo che abbiamo visto i diversi modelli di Data Warehouse, a partire dal modello multidimensionale, per poi discutere le sue rappresentazioni nel modello relazionale: gli schemi a stella e a fiocco di neve. Abbiamo infine sottolineato l'importanza della progettazione.

Grazie per l'attenzione!