Lezione 6 modulo 2

In questo modulo, andremo ad approfondire il discorso della rappresentazione di dati all'interno di un Data Warehouse e come si fa normalmente quando si va a parlare di modellizzazione dei dati, noi avremo questi modelli a due livelli: parleremo di modelli concettuali dei dati e di modelli logici. I modelli logici saranno più legati a delle tipologie di sistemi che io potrò costruire. Il modello concettuale vuole rappresentare quelli che sono i dati principali che noi andiamo a gestire, il fatto che noi vogliamo rappresentare che abbiamo dei fatti, delle dimensioni, e per le dimensioni anche le gerarchie di analisi che possiamo utilizzare come abbiamo visto, per le operazioni di analisi per poter aggregare i dati in certi modi. Cominciamo a vedere quindi un modello concettuale e un modello concettuale che andremo a vedere, che viene utilizzato per Data Warehouse è quello che viene chiamato DFM, Dimensional Fact Model. Noi vogliamo rappresentare, abbiamo detto, il fatto, ad esempio, la vendita, nei suoi elementi, abbiamo detto nell'esempio che abbiamo fatto quando abbiamo mostrato l'ipercubo, abbiamo il costo unità, la quantità, lo sconto e le varie dimensioni di analisi che sono previste per questo tipo di fatto. Abbiamo detto che avremo il tempo e vediamo qua la nostra generica data, mese, trimestre, anno, ma vediamo anche come, ad esempio, può essere rappresentata la settimana. Ovviamente la settimana può andare attraverso i mesi, e quindi noi possiamo avere le settimane numerate all'interno dell'anno, che sono una dimensione di analisi diversa, quindi potrò aggregare per settimane e settimane nell'anno, oppure per mesi, per trimestri - anno e quindi potrò ad esempio passare, poi vedremo nel prossimo modulo, quali sono le operazioni che possiamo fare, da una vista, ad esempio per trimestri, se vogliamo vedere i dati più in dettaglio, posso facilmente passare ad una vista per mesi, quindi col dettaglio per ciascun mese all'interno del trimestre. Vediamo anche altre possibili dimensioni per i prodotti, abbiamo visto tipo categoria, in un esempio precedente, ma ad esempio i prodotti possono essere classificati secondo i marchi, quindi posso analizzare le mie vendite per vedere come vende un certo marchio. Abbiamo detto che un'altra tipica dimensione è quella geografica e dobbiamo decidere quali sono le gerarchie da utilizzare per parlare in modo più in generale della posizione di un punto vendita. Ad esempio, posso definire delle zone all'interno di una città e posizionare le città in regioni e poi in stati, quindi questo mi consente di fare interrogazioni che vanno a vedere, ad esempio, quanto ho venduto all'interno di una certa regione. Altre informazioni che posso avere, ad esempio nel caso io abbia in questo caso un altro dato già disponibile che è quello delle informazioni del cliente, ovviamente un cliente in un punto di vendita non è detto che sia registrato, ma un modo per registrarlo può essere ad esempio l'uso di carte di fidelizzazione per il punto di vendita e posso avere delle informazioni, ad esempio, sesso oppure l'età del cliente. Vediamo che, mettendo in questo modello un trattino su un certo attributo nella gerarchia, definiamo che questo è un attributo opzionale, quindi potrebbe essere che non sia stato raccolto dal sistema di raccolta di dati e quindi non disponibile per alcune analisi. Quindi questo è il punto di vista concettuale, ci dice quali sono i fatti, quali sono le dimensioni, quali sono le gerarchie, abbiamo visto che la gerarchia in realtà può essere più articolata in una struttura tipicamente ad albero oppure a grafo. Come passiamo a un modello logico? Modelli logici per i Data Warehouse sono di tre tipi: chiamati MOLAP, ROLAP e HOLAP. MOLAP sta per Multidimensional OLAP; sostanzialmente in questo caso cosa abbiamo? Abbiamo una rappresentazione nel modello logico di quella che è la struttura di tipo concettuale. Qual è lo svantaggio di modello di questo tipo? (Ovviamente il vantaggio è avere una buona rappresentazione della gerarchia secondo quelle che sono le basi concettuali della descrizione di questa gerarchia). Lo svantaggio è il fatto che dobbiamo avere dei linguaggi di interrogazione ad hoc che possono risultare difficili da gestire da parte dell'utilizzatore. Il secondo tipo di modello logico che viene utilizzato è quello ROLAP e questo è il modello basato su relazioni, Relational OLAP. Vedremo poi che relazione vorrà dire utilizzare le classiche tabelle relazionali, quindi abbiamo una definizione nelle nostre dimensioni secondo una struttura tabellare, ovviamente il vantaggio qui è che noi possiamo usare le solite operazioni SQL e possiamo fare join fra le tabelle per poter analizzare i fatti. Ovviamente nel momento in cui introduciamo dei join, stiamo introducendo delle operazioni di tipo complesso che dovranno poi essere utilizzate. Come vedremo, ci saranno modi diversi di tradurre un

modello concettuale in un modello relazionale a livello logico. Quando abbiamo HOLAP, abbiamo un modello ibrido. Tipicamente avremo un modello di tipo relazionale per il Data Warehouse e poi un modello di tipo multidimensionale per i Data Mart e per consentire di gestire in modo specifico le diverse dimensioni, la navigazione nelle dimensioni, all'interno di uno specifico Data Mart. Vediamo adesso di concentrarci appunto su come tradurre una struttura di tipo multidimensionale in una struttura tabellare. Vedremo che abbiamo due tipi di schemi che possiamo creare in questo caso, che chiameremo a stella e a fiocco di neve per la struttura che avranno, che possiamo rappresentare schematicamente in questo tipo. Il centro sarà sempre la rappresentazione del fatto, quello che ci interessa rappresentare ovviamente sono le dimensioni delle loro gerarchie. Nel modello a stella, che vediamo in un esempio qui, sostanzialmente per ogni dimensione di analisi, creiamo una tabella in cui mettiamo tutti gli aspetti che vogliamo rappresentare della gerarchia. Quindi la data avrà un suo identificatore, vedremo che avremo sempre come base un identificatore che è sostanzialmente quello al livello di granularità più fine. All'interno della tabella che rappresenta una certa dimensione poi mettiamo tutte le varie possibilità di aggregazione, quindi abbiamo detto che le date le possiamo aggregare per giorno, per mese, per anno, per trimestre e per settimane. Un codice di una certa data sarà rappresentativo di un certo giorno all'interno del sistema delle date e poi avrò assegnato a questo dato il mese d'appartenenza, l'anno, il trimestre, la settimana dell'anno, che mi consentono poi di fare l'operazione di aggregazione. Cosa ho fatto? Ho appiattito tutta la gerarchia in un'unica tabella e questo mi consente di inserire l'identificatore della data all'interno della vendita e poi, utilizzando a seconda della richiesta dell'utente, qual è il livello di granularità desiderato, noi possiamo andare a fare le interrogazioni, ad esempio per trovare tutte le vendite all'interno di un certo mese. Stessa cosa viene fatta per gli altri elementi, ad esempio, per il prodotto io poi posso associare al codice del prodotto una serie di informazioni che sono quelle sul prodotto, ma poi abbiamo anche il tipo e vediamo che passiamo di nuovo da codici, quindi un identificatore del tipo, qual è la sua descrizione, avevamo detto il marchio, quindi identificatori del marchio con la sua descrizione, eventualmente il logo e la categoria col nome della categoria. Quindi, vediamo che noi andiamo a Inserire, all'interno della descrizione del prodotto, tutte le informazioni che sono nella gerarchia e che consentono di fare delle interrogazioni sulla gerarchia. Cosa metteremo nel fatto? Metteremo l'ID del prodotto. L'ID del prodotto ci dice di quale prodotto stiamo parlando in un certo momento. Stessa cosa ovviamente possiamo fare per i punti di vendita. Vediamo di nuovo che abbiamo sia ID, per tutto quello che abbiamo rappresentato nella gerarchia, e una descrizione. Come abbiamo detto, tipicamente in un Data Warehouse passo da codifiche, quindi l'associare un ID ad un punto di vendita anziché direttamente, ad esempio, un nome, un indirizzo, mi consente di eliminare eventuali ambiguità. Ad esempio prendiamo il nome della città: ci sono tante città che hanno lo stesso nome e quindi un identificatore della città mi consente di dire che io sto parlando di una certa specifica città e quindi di eliminare quelle che potrebbero essere le ambiguità nel semplice nome. Questo schema a stella quindi ha il vantaggio di darmi tutte le informazioni necessarie ma di appiattire le gerarchie. Ovviamente l'altro tipo di rappresentazione che è quello a fiocco di neve, invece, cerca di preservare il fatto che queste dimensioni sono organizzate in modo gerarchico e quindi vediamo la descrizione del punto vendita, qui, che mi dice dove sta un certo punto di vendita, ad esempio, in una certa zona, ma poi le informazioni sulla città sono rappresentate in una tabella diversa. Ovviamente, questo mi consente di usare questa rappresentazione della città anche all'interno di altri ipercubi perché avrò una descrizione standardizzata della città come anche della regione, che vediamo qua descritta come regione all'interno di un certo Stato, e quindi questo mi consente, diciamo, di utilizzare dimensioni in più ipercubi. Mi rappresenta meglio la gerarchia, ovviamente lo svantaggio è il fatto che dovrò fare più join per ricostruire, ad esempio, quali sono le vendite all'interno di una certa regione, perché dovrò andare a vedere quali sono le città all'interno della regione, quali sono i punti di vendita e quindi potrò andare ad esaminare le vendite e quindi dovrò fare più operazioni di join rispetto a quelle che sono le operazioni di join che sono fatte nello schema a stella.