Speciale Database

È possibile utilizzare l'insieme dei dati aziendali per suggerire strategie, confrontare scelte ed in generale supportare le decisioni? Sì, ma andando ben oltre i database relazionali oggi in funzione. In questo articolo vediamo alcuni rudimenti e i concetti base del data warehousing

Alle radici del Data Warehousing

di Marco Angeli

"Un sistema informativo deve provvedere alla raccolta e alla classificazione delle informazioni, da attuarsi con procedure integrate e idonee, al fine di produrre in tempo utile e ai giusti livelli le sintesi necessarie per i processi decisionali, nonché per gestire e controllare l'attività aziendale nel suo complesso."

La definizione di Sistema Informativo in letteratura riporta le funzioni che gli utenti si aspettano da un tale sistema. Esso dovrebbe essere in grado di gestire e controllare le attività operazionali dell'azienda (gestione paghe, magazzino, clienti e fornitori ecc...) ma anche sintetizzare questi dati producendo informazioni strategiche. Sino ad oggi, però, i Sistemi nella maggior parte delle aziende si preoccupano di risolvere solo le problematiche operazionali. Questi sistemi vengono chiamati OLTP (On Line Transaction Processing), in quanto basati sul concetto di transazione ovvero un insieme di più interrogazioni definisce una singola operazione logica.Un sistema OLTP può sollevare i normali impiegati di molti gravosi compiti, ma non aiuta molto i manager aziendali. In questi ultimi anni i problemi che i manager si trovano a dover gestire si sono moltiplicati. La globalizzazione porta a doversi confrontare con mercati su larga scala, ma anche una competizione che si basa sulla tempestività delle decisioni, sulla breve durata dei prodotti, ed ancora il decentramento dell'impresa o la flessibilità del lavoro.E' chiaro che nella lotta alla velocità d'azione, la dirigenza che è in grado di analizzare in maniera più dettagliata i dati in suo possesso, o ancora di prevedere gli andamenti futuri, vince.Nasce allora l'esigenza di utilizzare i dati operazionali storici a fini strategici, analizzando l'andamento di particolari settori dell'azienda, valutando eventuali cambiamenti strutturali ecc....Per fare ciò sono necessari tutta una serie di strumenti *evoluti* in grado, attraverso tecniche statistiche, neurali e quant'altro, di stimare, valutare, approssimare, "prevedere".Le risorse più interessanti sono naturalmente quelle maggiormente dinamiche, dunque a seconda delle tipologie di aziende le risorse umane, il marketing o altro ancora.Infatti è su queste informazioni che si basano le decisioni e la scelta di investimenti. Un esempio può essere la scelta del lavoro flessibile, quando si prevede di averne bisogno, in quale quantità, con quali caratteristiche professionali o attitudinali.

I Sistemi di Supporto alle Decisioni

Le aziende, non solo di grandi dimensioni, necessitano di modi *creativi* di esaminare e valutare il mercato, non solo per prevederne gli andamenti, ma con lo scopo di capirne i meccanismi. Purtroppo però è impensabile l'analisi manuale dei dati operazionali, per la quantità, per l'incompletezza o inesattezza ed ancora troppo ricchi di dettagli marginali che possono distrarre.Quello che serve è un DSS (Sistema di Supporto alle Decisioni). Questi sistemi propongono una serie di metodiche atte all'analisi, al reporting ed alla previsione basate sui dati immagazzinati.Le operazioni fondamentali sono di quattro tipologie **[1]**:

Reporting. Questa tipologia di strumenti, molto semplici eseguono una serie di query, periodicamente creando appunto dei report, tipicamente a fini statistici, per l'analisi grafica degli andamenti aziendali.

Data Mining. Tipicamente strutturato con tecniche neurali il software di data mining scorre la base di dati alla ricerca di pattern, per quelle operazioni che vengono definite di Knowledge Discovery, ovvero la scoperta di correlazioni sconosciute. Un esempio tipico è dato dalle assicurazioni che potrebbero essere interessate alle correlazioni esistenti tra incidenti e colore delle auto.

What if. Questi strumenti basati su sistemi neurali o statistici, tentano di prevedere gli andamenti di alcuni parametri al variare di altri. Un esempio tipico di cosa questi sistemi tentano di risolvere sono interrogazioni del tipo "cosa succede se al posto di vendere cartoline a 500 Lire vendo biglietti di auguri a 300 Lire?".

OLAP (On Line Analytical Process). Probabilmente lo strumento più interessante, più potente che gli utenti hanno al loro servizio. Questi software permettono l'interazione diretta degli utenti sulla base dati. Lo scopo ultimo è permettere la generazione di interrogazioni, in maniera intuitiva e senza preoccuparsi dell'effettiva struttura logica dei dati [4]. Questi obiettivi sono in parte raggiunti attraverso l'utilizzo di operatori, che l'utente può eseguire sui diversi oggetti rappresentanti i dati in suo possesso. Tipici operatori messi a disposizione sono: Drill-Down, Roll-Up, Pivot, Top N. Naturalmente molti altri possono essere definiti in base alle necessità degli utenti. L'operatore Drill-Down eseguito su un attributo di una vista aggregata, mostra i dati della vista stessa ad un maggiore dettaglio rispetto all'attributo prescelto. Al contrario l'operatore di Roll-Up aggrega i dati rispetto all'attributo prescelto. L'operatore di Pivot, risulta necessario per poter cambiare punto di vista, ovvero poter cambiare la dimensione in analisi. Infine l'operatore di Top N restituisce i primi N elementi rispetto ad un attributo.

Un esempio tipico di analisi OLAP può essere definito come segue:

- Una catena di negozi a livello nazionale, per decidere l'apertura di nuovi centri vendita ricerca quali punti vendita sono maggiormente redditizi (Top N). Analizza quindi i dati sui fatturati rispetto alle regioni, quindi (Drill-Down) rispetto alle città o ancora (Drill-Down) rispetto all'ubicazione nei centri cittadini. Ma è fondamentale non solo dove costruire, ma cosa costruire, quindi si può spostare l'attenzione (Pivot) alla tipologia di punto vendita, o ancora al tipo di oggetti venduti ecc....
- L'analisi viene condotta navigando i dati. Le interrogazioni poste dall'utente attraverso gli operatori OLAP partizionano l'insieme dei dati creando opportune viste. Queste possono essere poste a diverse granularità, dipendentemente dall'aggregazione possibile e richiesta. Basse granularità significa una moltitudine di dati spesso inefficace a mostrare la realtà, mentre alte granularità, pochi valori aggregati possono mostrare in maniera più limpida la cosidetta bussins logic. Per poter recuperare la singola unità di informazione, detta *fatto*, vengono definiti tutti i parametri che la definiscono, detti *dimensioni*. Come mostrato in **Figura 1** non viene utilizzata alcuna aggregazione. L'aggregazione di un *range* di fatti avviene imponendo opportune condizioni sulle dimensioni, come mostrato in **Figura 2** e **Figura 3**. In questo caso l'aggregazione avviene in base alle condizioni poste sulle dimensioni stesse.

Una volta accertata la necessità di un sistema decisionale, e la presenza di un sistema informativo già collaudato ed in uso, si valuta l'introduzione di strumenti di analisi, di interrogazione e di previsione. I risultati sono spesso deludenti; il sistema non è in grado di eseguire i compiti richiesti. Le motivazioni sono però chiare e semplici **[4]**:

- problemi di inconsistenza tra dati prelevati da sorgenti eterogenee;
- problemi di integrazioni tra dati presenti in sorgenti diverse;
- evoluzione temporale non prevista e di difficile introduzione;
- enormi quantità di dati storici che riducono le prestazioni del sistema per l'utilizzo operazionale;
- prestazioni insoddisfacenti nel recupero dati.

I problemi di integrazione sono spesso collegati ai problemi di consistenza. L'unione di dati provenienti da sorgenti eterogenee, costruite in genere da personale differente, con diverse specifiche e diverse metodiche sono spesso incompatibili. Pensiamo soltanto a quanti modi vi sono per dichiarare il sesso di una persona: un valore booleano (1 per maschio o per femmina), un carattere (F o M, ma anche f o m, se non F o m).Le problematiche relative alla temporizzazione, ovvero all'inserimento della variabile temporale o storicizzazione del sistema informativo può risultare problema assai complesso. In particolare se la base dati precedente non è ben organizzata, o ben documentata o ancora se si ha a che fare con sistemi non strutturati (file di testo, e-mail, news, html, grafici, immagini, suoni, ...).L'efficienza del sistema preesistente è chiaramente posta in crisi dall'incremento dei dati immagazzinati, che ad ogni aggiornamento si accumulano.

Soluzione Data Warehouse

Tutti questi problemi possono essere risolti attraverso il Data Warehouse. Un *magazzino* in cui i dati estratti dalle diverse fonti, aggregati, omogeneizzati e resi consistenti vengono mantenuti in forma storica. Il sito appositamente strutturato, progettato e costruito, prende il nome di Warehouse (magazzino). Dunque i Data Warehouse sono un ulteriore mezzo di potenziamento del sistema informativo aziendale; non sostitutivo del precedente sistema operazionale ma completamento dello stesso attraverso l'integrazione e la strutturazione apposita, atta al supporto degli ambienti di analisi tipici della scienza decisionale. Un sistema decisionale si

basa su uno o più sistemi operazionali. Tipicamente questi sono Data Base relazionali già in opera, in grado di gestire le operazioni transazionali che avvengono all'interno dell'azienda, ma, anche esterni quali andamenti di borsa, o ancora le informazioni cartacee che tuttora assillano molti impiegati.I sistemi di supporto alle decisioni (DSS, Decision Support System), sono nati anni prima dell'idea dei Data Warehouse. In generale i due concetti possono essere separati. In realtà i Data Warehouse sono una soluzione ai problemi sollevati dai DSS, non l'unica, ma spesso la migliore e la più facile da applicare. I Data Warehouse risolvono i problemi utilizzando una tecnica cosiddetta *in-advance*, ossia prevenendo le richieste dell'utente. Questa tecnica si sviluppa attraverso la prelevazione dei dati dalle fonti distribuite, integrandoli e immagazzinandoli in un deposito con facilitazioni per l'accesso e l'interrogazione. La *soluzione* Data Warehouse è consigliabile quando esistono delle precondizioni che ne facilitano l'utilizzo e l'installazione **[3]**:

- Le interrogazioni poste dagli utenti debbono agire su un'anticipabile porzione delle informazioni raggiungibili, in pratica è necessario sapere quali dati, tra quelli a disposizione sono interessanti per l'utenza.
- Gli utenti necessitano di bassi tempi di risposta.
- Gli utenti non sono interessati ad avere le informazioni aggiornate. Tipicamente non vi è interesse ai fini decisionali alle informazioni immediate, o ancora in fase di esecuzione. Per la maggior parte delle operazioni DSS sono sufficienti i dati relativi al giorno antecedente. In alcuni casi, possono essere sufficienti anche dati aggiornati al mese o all'anno, basti pensare agli indici di bilancio in un sistema finanziario. Questi indici possono essere considerati validi solo dopo che i bilanci stessi sono chiusi ovvero a fine anno; averli a disposizione prima di tale data può essere solo controproducente, a causa della non esattezza.
- Gli utenti desiderano un accesso a copie private e statiche delle informazioni, non soggette a variazioni.

Una volta riunite le informazioni desiderate si possono eseguire tutta una serie di ottimizzazioni in maniera da diminuire i tempi di esecuzione delle interrogazioni. Le ottimizzazioni si possono ottenere applicando svariate tecniche, alcune delle quali già note ed applicate per i VLDB (Very Large DataBase). Su questi ultimi sistemi, infatti, le prestazioni sono rese critiche dalle enormi moli di dati presenti. Proprio la quantità di dati permette però di separare i campi di una tabella per esempio (partizionamento verticale) in base all'utilizzo o meno di questi attributi in certe query. Naturalmente se le interrogazioni accedono a determinate righe della tabella in base al valore di un determinato campo, questo può essere utilizzato per suddividere la tabella in due in base a tale valore (partizionamento orizzontale). Un'altra tecnica *classica* è la materializzazione di viste aggregate. Interrogazioni molto frequenti vengono materializzate come tabelle vere e proprie, a scapito dei tempi di aggiornamento, ma con il vantaggio di poter eseguire la query con la semplice lettura della vista. Tecniche specifiche per il mondo Data Warehouse sono invece: la denormalizzazione e l'indicizzazione effettuata con indici ad-hoc o cablati per le specifiche query. La denormalizzazione della base dati è una pratica molto utilizzata che ha alcuni vantaggi, evita dei costosi join, ed alcuni svantaggi, aumenta lo spazio occupato dai dati. In pratica supponiamo di avere una tabella di persone ognuna delle quali vive in una certa città. Per un normale Data Base relazionale si provvederebbe a definire due tabelle, la prima con i dati delle persone ed il codice della città, la seconda con i dati delle città, nazione, paese e dati di interesse per le città. Per un Data Warehouse si potrebbe costruire una sola tabella con le informazioni delle città ripetute per ogni persona. Chiaramente lo spazio utilizzato dal secondo metodo è molto maggiore del primo poiché molte informazioni saranno ripetute. Nel caso due persone abitino nella medesima città i dati di questa verrebbero ripetuti due volte. Si deve però considerare che per trovare in che regione abita una persona si devono unire le due tabelle nel primo caso, mentre nel secondo è una semplice lettura della tabella.Il Data Warehouse non è solo un magazzino di dati, ma un insieme di strumenti che mettono in grado di gestire il sistema e di riottimizzarlo a seconda delle necessità del momento. Questi tools appartengono a categorie ben definite di programmi software:

- Strumenti per l'estrazione/trasformazione dei dati e popolamento della base dati. In funzione del numero di fonti dati, della etereogeneità e dei problemi connessi con i dati stessi è possibile inserire degli strumenti di estrazione e popolazione generici o implementare dei programmi *ad-hoc*.
- Il DBMS, il contenitore dei dati, che deve garantire livelli di prestazioni, scalabilità e disponibilità definiti in fase di progetto.
- Il motore per i processi di analisi. Sostanzialmente un sistema in grado di interfacciarsi con il DBMS che conosca la disposizione dei dati e traduca attraverso opportune metodiche le interrogazioni degli utenti in una o più query SQL.
- L'interfaccia utente utilizzata per l'accesso alle informazioni contenute nel Data Warehouse. Questi programmi non sono però i soliti sistemi di interfacciamento ai dati ma sono sistemi evoluti che

debbono permettere la definizione di query da parte dell'utente o ancora di navigare i dati in maniera intuitiva.

Un'implementazione di Data Warehouse è veramente efficace solo se fornisce agli utenti finali gli strumenti utili per ottenere informazioni specifiche, necessarie a sostenere i vari processi decisionali aziendali. In altre parole, il Data Warehouse si impone come strumento strategico se le varie direzioni funzionali alle quali è destinato possono accedere ai dati secondo principi di analisi multidimensionale. In generale si può affermare che una misura di bontà di una tale soluzione non è data soltanto dalle prestazioni in termini di tempo di esecuzione, ma di quanto facilmente gli utenti finali possono ricevere conoscenza dai dati. Non solo rispetto alle informazioni registrate (ad esempio i fornitori o i clienti) ma sul business e sulla sua natura.

Gli strumenti di analisi **OLAP** pongono l'accento sull'intuitività dell'interfaccia utente, poiché debbono permettere l'impostazione di interrogazioni, spesso complesse. Spesso l'utilizzo degli strumenti di OLAP da parte del management viene intesa come una sorta di validazione o ricerca manuale dei dati di interesse, dunque, è fondamentale la risposta in tempi adeguati al mantenimento dell'attenzione sulla problematica in analisi. I concetti di OLAP e Data Warehouse sono complementari. Un Data Warehouse immagazzina i dati e ne controlla il flusso mentre i tool OLAP trasformano i dati estratti in informazioni strategiche. Si parla di trasformazione, non di navigazione, poiché enfasi particolare è posta sulla visualizzazione di queste e sui metodi di analisi a disposizione, spesso basati su modelli estremamente complessi.Il vantaggio delle tecniche OLAP sta soprattutto nell'interattività e nella flessibilità caratteristiche direttamente correlate con la fantasia e la preparazione di chi le utilizza.

A chi serve un Data Warehouse

Un management che si domandi *"cosa è successo?",* "*perché è successo?*" e *"come posso ripeterlo?"* oppure *"come posso evitarlo?"*, è un buon candidato all'installazione di un sistema decisionale, se poi risultano verificate anche le condizioni di applicabilità come definite sopra allora il Data Warehouse è la soluzione adeguata. In questo caso si possono integrare i dati provenienti da fonti sparse nell'intera azienda, fornendo un quadro completo delle informazioni presenti, consentendo di condurre in modo estremamente efficace iniziative di marketing, attività di fidelizzazione della clientela o ancora ridefinizione dei prodotti. La soluzione Data Warehouse garantisce un notevole vantaggio competitivo anche da un punto di vista finanziario, in quanto consente ai sistemi di fornire la maggior quantità di informazioni possibile nel formato più adatto a prendere le migliori decisioni operative, questo infine permette di mantenere sotto controllo specifici costi. Le fonti di informazione possono essere le più varie, non solo interne all'azienda, pensiamo solo ai dati di borsa, ai bancomat ecc...I vantaggi correlati con l'utilizzo di queste tecnologie, sono derivanti dalla codifica di un unico standard tecnologico per l'analisi dei dati. Inoltre le informazioni possono essere utilizzate molto prima riducendo così il cosiddetto time-to-market e dando la possibilità di gestire il business quotidianamente. Inoltre si deve considerare l'autonomia guadagnata dagli utenti nelle loro analisi poiché i dati provengono dalla stessa origine, nella quale viene mantenuta la consistenza.Questo tipo di soluzione si adatta particolarmente bene anche a chi necessita di un'organizzazione finanziaria efficientissima, in grado di fornire la massima assistenza in fase decisionale, e non soltanto nella semplice elaborazione delle transazioni **[9]**.Dunque chi abbisogna dei dati di un Data Warehouse sono varie tipologie di utente: responsabili marketing, product manager, responsabili commerciali. Ognuno di questi con proprie esigenze applicative: dal semplice query e reporting fino alle applicazioni di analisi economica. I sistemi di analisi economica erano un tempo dominio degli uffici studi che poche aziende di grosse dimensioni o enti di categoria potevano permettersi. Oggi queste applicazioni, sono svolte al proprio computer poiché esiste una tecnologia che permette di manipolare in modo semplice ed intuitivo i dati di complesse realtà aziendali, i Data Warehouse.

Come costruire un Data Warehouse

La prima fase di studio, che può durare da pochi mesi all'anno se non si utilizzano le opportune evolute metodiche, comporta l'esame dell'organizzazione in tutti i suoi vari aspetti, e la definizione dei traguardi da raggiungere, in termini di prestazioni e centri di interesse da analizzare.Effettuata questa prima fase di analisi dei requisiti, è fondamentale una conoscenza dei dati, delle incompatibilità tra essi, delle eventuali necessità di integrazione ecc...Nei casi in cui i tempi di costruzione si rivelino troppo alti si prende in considerazione la costruzione incrementale del magazzino. I vari segmenti in cui viene suddiviso il lavoro sono legati alla struttura dell'azienda. Ad esempio se si è particolarmente interessati alla gestione finanziaria, si può costruire un *piccolo* Data Warehouse che immagazzini solo i dati relativi a tale area. Questo sito, prende il nome di Data Mart [6]. Naturalmente le varie sezioni devono essere strutturate

appositamente per facilitare la connessione con altri successivi Data Mart. Per permettere questa futura integrazione è necessaria una fase di progettazione concettuale. La fase di progettazione produce la prima documentazione di confronto con il committente e di pianificazione per le fasi successive. Scelto un modello dei dati, progettazione logica, inizia un ciclo di ottimizzazione che verrà ripetuto ad intervalli regolari per mantenere il Data Warehouse sempre in perfetta efficienza. Questo ciclo prevede le operazioni di caricamento dei dati ovvero di aggiornamento degli stessi, con il prelevamento dai dati operazionali; il partizionamento, la creazione di viste e la ristrutturazione degli indici. Questa fase risulta generalmente inglobata in un software (loader) che si preoccupa di tutte queste operazioni.

Conclusioni

Il mondo Data Warehouse ha visto negli anni una serie di fallimenti clamorosi, sia per la quantità di risorse spese, sia per l'inutilità del prodotto finale. I fallimenti sono dovuti ad un numero notevole di fattori, il primo dei quali sicuramente un analisi dei requisiti effettivi non consona, o una progettazione concettuale, logica e fisica non adeguati. Una progettazione concettuale approssimativa se non inesistente può portare a sistemi caotici di difficile mantenimento ed utilizzo, sistemi che in generale degradano molto velocemente senza alcuna possibilità di riorganizzazione, essendo inutilizzabile la documentazione. Una progettazione fisica non adeguata può rendere difficoltoso l'accesso ai dati a causa delle prestazioni deludenti. I dati sporchi o incompleti o ancora non aggiornati rendono, di fatto, il sistema non utilizzabile a causa dell'inutilità dei dati ivi contenuti. Infine la mancanza di connessione tra le diverse parti rende inutilizzabile il sistema impedendo quella facile navigazione dei dati fondamentale per il problema che si vuole risolvere con questo mezzo.

Bibliografia

[1] *"Essbase Arbor, The Role of the OLAP Server in a Data Warehousing Solution"*, Arbor Software White Paper, C/O www.essbase.com;

[2] *"INFORMIX-OnLine Extended Parallel Server and INFORMIX-Universal Server (a new generation of Decision. Support Indexing for Enterprise Data Warehouses"*, Technical note from Informix;

[3] J.Widom, *"Research Problems in Data Warehousing"*. Conference on Information and Knowledge Management, 1995;

[4] *"Olap Council White Paper, Lessons From the experts"*;

[5] *"The Data Warehouse Toolkit: Pratical Techniques for Building Dimensional Data Warehouse"*, Ralph Kimball;

[6] *"The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses"* Ralph Kimball, ed al;

[7] *"Sybase IQ, delivering Interactive Performance to the Data Warehouse"*, Sybase White Paper;

[8] *"Data Warehouse Technical Guide"*, Sybase White Paper, C/O www.sybase.com, T.Flanagan, E. Safdie;

[9] *"Oracle Warehouse"*, Oracle White Paper, C/O www.oracle.com.

Marco Angeli è laureato in Scienze dell'informazione e si occupa attivamente di progettazione di basi di dati e sistemi DSS.