

# Datalight: un progetto di data warehousing e data mining per piccole e medie imprese

Virginia Gallo  
Olivetti Ricerca – Gruppo Getronics  
Contrada la Marchesa S.S. 271  
70020 Bitritto (Bari) Italy  
+39 80 6352063  
virginia.gallo@getronics.com

Donato Malerba  
Dipartimento di Informatica  
University of Bari  
Via Orabona, 4 – 70126 Bari , Italy  
+39 80 5443269  
malerba@di.uniba.it

## SOMMARIO

In questo articolo si descrivono lo scenario, gli obiettivi, le attività e l'applicazione pilota di un progetto a finanziamento regionale, denominato Datalight, mirante a fornire soluzioni di business intelligence adatte a piccole e medie imprese (PMI). Le soluzioni proposte permettono di coprire l'intero processo di scoperta della conoscenza nei dati, a partire dalla selezione e pulizia dei dati caricati in un data warehouse, per passare alla loro trasformazione e analisi mediante strumenti di data mining, per giungere infine alla visualizzazione di risultati e alla generazione di report. Il settore merceologico sul quale è condotta la sperimentazione è quello agroalimentare, caratterizzato da una forte presenza di PMI operanti nella regione di interesse, ovvero la Puglia.

## 1. LO SCENARIO

Negli ultimi decenni il ciclo di vita dei processi decisionali nelle aziende è andato accorciandosi sempre più. La tempestività nell'individuare nuovi segmenti di mercato, nello scoprire preferenze e comportamenti da parte di clienti, nel ridurre eventuali sprechi nella produzione o nel razionalizzare altri processi aziendali, è diventata vitale per la sopravvivenza delle aziende. Questo è tanto più vero quanto più piccola è la dimensione della realtà aziendale: oggi giorno le Piccole e Medie Imprese (PMI) sono costrette a competere su mercati globali pur non disponendo né delle strutture né delle risorse disponibili invece alle grandi aziende. Tale tempestività, tuttavia, a volte contrasta con la mole dei dati da elaborare per estrarre le informazioni necessarie a supportare il processo decisionale. I dati sono spesso difficilmente recuperabili perché sommersi nell'insieme di informazioni ospitate dal sistema informativo. Il ricorso alle tecnologie dell'informazione è quindi un passo obbligato. Con il termine *Business Intelligence* (BI) si suole indicare l'insieme dei processi, delle tecniche e degli strumenti basati sulla tecnologia dell'informazione, che supportano i processi decisionali di carattere economico. Il problema fondamentale nel BI è quello di disporre di sufficienti informazioni in modo tempestivo e fruibile e di analizzarle cosicché da poter avere un impatto positivo sulle strategie, le tattiche e le operazioni aziendali. Le informazioni possono riguardare la specifica impresa oppure situazioni più generali di mercato, o ancora la concorrenza (*competitive intelligence*). Nel BI è possibile individuare tre attività importanti: raccolta dei dati, analisi dei dati e proposta di consigli.

Per quanto riguarda l'attività di raccolta dei dati è importante che questa non si limiti ai soli dati *transazionali*, generati e usati nei processi produttivi o operativi di un'impresa (processi di più basso livello), ma che sia orientata anche ai dati *decisionali* (o *business data*), caratterizzati da una natura aggregata, una struttura flessibile, un uso non ripetitivo, un orizzonte temporale ampio, e una proprietà di staticità (Inmon, 1992). L'idea alla base del concetto di *data warehouse* (DW) è che le esigenze dei processi operativi sono completamente diverse da quelle dei processi decisionali, e quindi i sistemi informativi utilizzati in ambito operativo devono essere distinti da quelli adoperati in ambito decisionale (Kimball, 1996). Il DW è quindi un ambiente specializzato a sé stante, in cui far periodicamente convergere tutti i dati utili ai processi decisionali, prelevandoli dai dati di produzione, eventualmente arricchendoli con dati provenienti da fonti esterne (Internet, Banche Dati, etc.), e infine organizzandoli in una forma comprensibile per chi in azienda deve prendere decisioni tattiche e/o strategiche.

I DW sono spesso considerati una buona tecnologia per supportare soluzioni di *knowledge discovery in databases* (KDD). Questo è il processo di esplorazione e analisi di grandi quantità di dati, condotto in modo automatico o semiautomatico, al fine di scoprire delle regolarità (*pattern*) nei dati, che siano considerabili nuovi elementi di conoscenza (Fayyad et al., 1996). Tale processo prevede, fra l'altro, delle fasi di selezione e pulizia dei dati, al fine di rimuovere inconsistenze, trattare dati mancanti, e determinare il giusto livello di aggregazione. Tutto ciò è già assicurato da un DW, che diventa quindi una tecnologia di supporto, ma non necessaria, al KDD.

In un contesto aziendale, la conoscenza scoperta può avere un valore perché consente di aumentare i profitti riducendo i costi oppure aumentando le entrate. Questo spiega l'importanza di soluzioni KDD nel BI. Il processo KDD include un passo di *data mining* (DM), consistente nell'applicazione di un particolare algoritmo di analisi dei dati, anche se nel seguito i termini KDD e DM saranno considerati sinonimi, seguendo un uso ormai invalso nel mondo della *information technology*.

Le soluzioni di data warehousing attualmente sul mercato sono costosissime. Uno studio condotto da Watson e Haley (1997) riporta una spesa media per DW di circa 2.3 milioni di dollari, sicuramente non affrontabile dalle PMI che pure necessitano di tali tecnologie almeno quanto le grandi organizzazioni. Le soluzioni di DM sono certamente meno costose ma non sempre adeguate alle esigenze reali di un'azienda.

Allo scopo di fornire a PMI una soluzione “praticabile” di data warehousing insieme ad alcune applicazioni di BI, Olivetti Ricerca del Gruppo Getronics (OR), tramite il centro operante nella sede di Bitritto (Bari), partecipa con il Dipartimento di Informatica dell’Università di Bari (DIB) e il Dipartimento di Progettazione e Produzione Industriale del Politecnico di Bari (DPPI), al progetto POP “Datalight: uno strumento di innovazione per le Piccole e Medie Imprese in Puglia” finanziato dalla Regione Puglia. Il DW proposto sarà “light” e cioè caratterizzato da costi ragionevoli, tempi di realizzazione ridotti, e minor mole di dati immagazzinati. Sulla base delle strutture informative rese disponibili dal DW, questo progetto offrirà all’utente finale anche alcuni strumenti di analisi dei dati a supporto del processo decisionale aziendale.

In questo contributo si descrive inizialmente l’architettura del sistema prototipale prodotto dal progetto. In seguito si spiegano le attività di progetto, in parte legate al particolare settore merceologico di sperimentazione, quello agroalimentare. Si mostrerà come l’analisi preliminare dei processi decisionali delle PMI operanti in Puglia nel settore agro-alimentare abbia sia influenzato la scelta operata fra le varie soluzioni disponibili sul mercato per la realizzazione del DW, e sia guidato l’attività di modellazione concettuale e logica dei dati. Nel lavoro si descriverà anche KDB2000, il tool di data mining che si intende integrare nel prototipo di progetto. Infine si descriveranno alcuni problemi di analisi dei dati posti dall’azienda pilota coinvolta nel progetto.

## 2. L’ARCHITETTURA DEL SISTEMA

L’architettura funzionale del prototipo di sistema in fase di realizzazione nel progetto è mostrata in figura 1.

L’interfaccia dati esterni è quella componente contenente i driver verso i dati esterni (i dati del sistema informativo aziendale, altri dati per esempio acquisiti periodicamente dall’Azienda e memorizzati in formato elettronico mediante opportune procedure, i dati di banche esterne, private o siti Internet, etc...). Gli strumenti di ETL (Extraction, Transformation and Loading) servono ad estrarre i dati esterni dal sistema informativo aziendale o da altre sorgenti, a pulirli in modo da renderli consistenti con lo schema, a validarli, a riepilarli e a convertirli nel formato appropriato ed infine a caricarli nelle tabelle del DW. La componente di Data Management ospita i dati del DW e costituisce il cuore dell’intera soluzione architeturale. Essa comprende un prodotto Olap Server, opzionale, per la creazione e gestione delle aggregazioni multidimensionali dei dati, e un DBMS relazionale (RDBMS) per la gestione delle tabelle del DW.

Esistono, poi, due componenti funzionali che sono a supporto trasversale dell’intera architettura e sono:

- il *Metadata Management*: gestisce i *metadati*, vale a dire le strutture informative usate dagli amministratori e sviluppatori del DW per descrivere le sorgenti dei dati, le regole per effettuare il mapping dei dati dalla forma operativa a quella del DW, i termini aziendali utilizzati per descrivere i dati, i termini tecnici corrispondenti ai termini aziendali che possono essere utilizzati per accedere ai dati, le regole utilizzate per effettuare il riepilogo di alcuni dati;
- i *tool amministrativi*: permettono la corretta gestione del DW.

Infine, la soluzione “Datalight” mette a disposizione dell’utente strumenti di front-end, come:

- strumenti di query e reporting per poter leggere sotto forma di “business report” alcune informazioni aggregate permettendo confronti strategici;
- strumenti di On-Line Analytical Processing (OLAP) per permettere un’analisi multidimensionale delle informazioni ([www.olapreport.com](http://www.olapreport.com));
- strumenti di data mining.

La realizzazione del sistema utilizzando le soluzioni software attualmente offerte dal mercato è descritta nel paragrafo successivo.

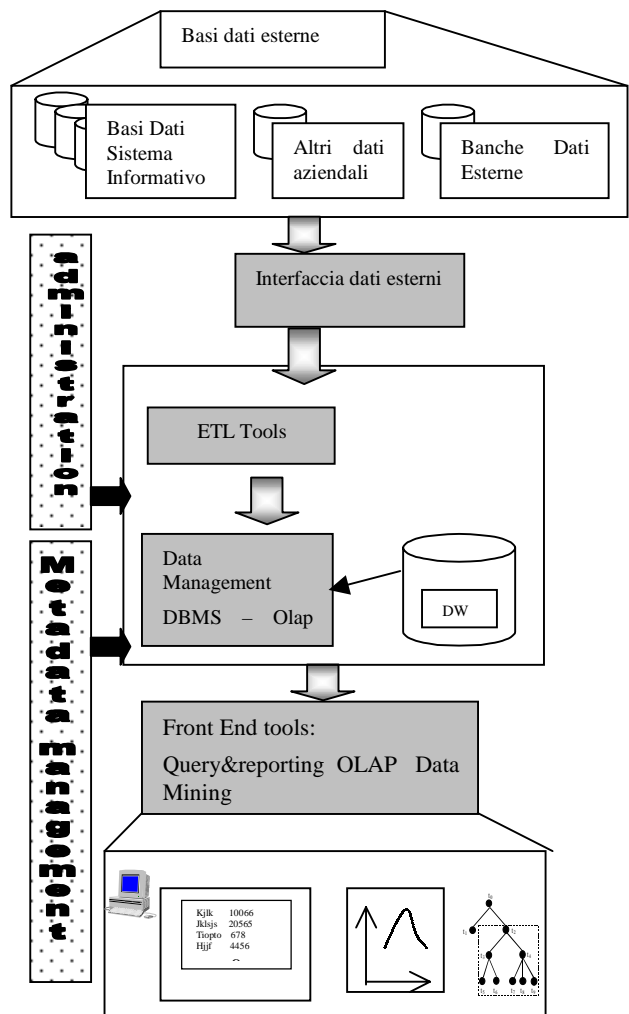


Figura 1. Architettura funzionale di “Datalight”.

## 3. LE ATTIVITÀ DI PROGETTO

Gli obiettivi che si sono perseguiti e si stanno perseguendo nel corso del progetto sono i seguenti:

### OB1. Analisi dei dati e delle procedure di un campione di aziende agroalimentari

Come settore merceologico di sperimentazione della soluzione progettata si è scelto il settore agro-alimentare. Quello agroalimentare è in Puglia, la Regione in cui agiscono sia OR che DIB e DPPI, un settore trainante con notevoli potenzialità di crescita. Inoltre, le imprese agroalimentari del territorio pugliese sono in gran parte PMI, e generalmente non utilizzano nel loro approccio al mercato gli strumenti tipici di marketing, tantomeno se basati su tecnologie informatiche innovative. L'esperienza degli ultimi anni consente di affermare, invece, che il successo del mercato agroalimentare è strettamente connesso proprio ad una migliore organizzazione del marketing e della commercializzazione, grazie anche ad un corretto e diffuso utilizzo di adeguate tecnologie informatiche.

Le attività pertinenti a questo obiettivo, la cui realizzazione è stata affidata al DPPI, sono consistite, quindi, nella raccolta di informazioni su un significativo numero di PMI operanti in Puglia nel settore agroalimentare per acquisire i principali dati di business su cui le aziende operano e le modalità dei processi aziendali propri di questa tipologia di impresa. Il report prodotto grazie alla collaborazione di quaranta aziende pugliesi con fatturato superiore ai 5 miliardi di lire (Dioguardi *et al.*, 2000) ha avuto la finalità di fornire un input per la definizione della soluzione generalizzata di DW idonea alle esigenze delle PMI pugliesi del settore agro-alimentare.

### OB2. Progettazione ed implementazione della soluzione "light" di data warehousing

Definita l'architettura di sistema, si è operata un'attenta valutazione su come essa possa essere realizzata utilizzando le soluzioni di software per implementare progetti di DW attualmente offerte dal mercato (Sybase, Informix, Oracle, Microsoft, IBM). Il confronto tra le diverse opportunità, operato in termini di costi e prestazioni, ci ha suggerito come migliore risposta ai nostri obiettivi progettuali di orientarci alla scelta della piattaforma Microsoft SQL Server 7.0 (2000), che pur non offrendo tool e componenti sofisticati rispetto ad altre scelte più costose, certamente assicura facilità di realizzazione, basso costo nell'investimento iniziale in modo da rendere accessibile la soluzione ad aziende di qualsiasi dimensione, un facile accesso a software di supporto ai processi decisionali di altri venditori, e la scalabilità al crescere del volume dei dati. Infine, i risultati dell'indagine condotta dal DPPI ha rivelato che nel 60% delle imprese viene utilizzato Access di Microsoft, nel 15% SQL Server ed in percentuali notevolmente inferiori (intorno al 7%) Informix, Oracle e DB2. Ciò non fa che corroborare l'opportunità della scelta operata. In Tabella 1 si riporta la corrispondenza che si può tracciare fra alcune componenti previste per Datalight e i prodotti offerti dalla soluzione Microsoft. La scelta di utilizzare Excel 2000 come front-end di sistema è anch'essa spiegabile con la diffusione di questo foglio elettronico.

### OB3. Modellazione dei Dati

Questa attività, affidata al DIB, comprende la modellazione concettuale e logica dei dati di business che devono popolare una soluzione globale e generalizzata di DW a partire dai risultati dell'analisi dei dati e dei processi decisionali indagati nell'obiettivo OB1. Il modello *multidimensionale* è il più consono per la modellazione concettuale di un DW (Agrawal *et al.*, 1995). In un modello multidimensionale i dati sono organizzati in uno o

più ipercubi o cubi n-dimensionali, come si spiegherà in seguito. Le dimensioni di un cubo definiscono lo spazio multidimensionale (Kini, 1996). A differenza dei classici array, nei quali gli indici caratterizzati da un ordine lineare (tipicamente gli indici sono valori interi), in un modello multidimensionale sugli indici potrebbe non essere definito un ordine, o comunque si potrebbe definire un ordinamento parziale (gerarchie di valori). Sfortunatamente non esiste ancora un modello formale universale per i dati multidimensionali, per quanto sia stato costituito un consorzio di aziende allo scopo di definire, fra l'altro, uno standard per sistemi basati su rappresentazioni multidimensionali ([www.olapcouncil.org](http://www.olapcouncil.org)). Nella progettazione ci si è quindi attestati sul modello offerto dal tool ERWIN ([www.cai.com](http://www.cai.com)).

I dati che le aziende del settore agroalimentare analizzano per condurre il proprio business e le rispettive dimensioni di analisi sono stati inizialmente estratti dalla *Sezione Dati* del questionario inviato alle aziende. Mediante un clustering dei dati in base alle potenziali dimensioni di analisi, si è potuto individuare in modo naturale sei cubi multidimensionali, rispettivamente quello delle vendite, degli acquisti, dei costi, dei tassi di cambio, degli indici fornitori e degli indici clienti. Il numero delle dimensioni di analisi varia da due (tassi di cambio) a sette (costi). Questo modello concettuale iniziale è stato poi raffinato tenendo conto dei risultati ottenuti dall'indagine di cui all'obiettivo 1. Il risultato, riportato in Figura 2, mostra un modello concettuale sostanzialmente semplificato, nel quale si individuano come attributi di alcune dimensioni di analisi (fornitori e clienti) gli indici inizialmente ritenuti oggetto di analisi (Esposito & Malerba, 2000).

Tabella 1. Corrispondenza tra i framework "Datalight" e Microsoft per la realizzazione del prototipo.

Framework Datalight	Framework Microsoft
Interfaccia dati operativi	Operational data: OLE/DB, ODBC
DW/DM modelling	Data Warehouse Design and Data Mart design: Visual Data tools
ETL tools	Data transformation and cleansing: DTS
Data Management	OLAP Services - SQL server 7.0
Metadata Management	Microsoft repository
Administration	Data Warehouse Management: Microsoft Management Console, SQL server Enterprise manager, SQL Agent, SQL Profiler, Tuning expert
Front End tools	End User tools: Excel 2000, English query, applicazioni terze parti

In un'applicazione di DW sarebbe naturale tradurre un modello concettuale multidimensionale in un modello logico con proprietà analoghe. Questo tuttavia presuppone l'uso di un DBMS che supporti tale modello logico, cioè un *MDDBMS* (*Multidimensional DBMS*), che per l'appunto memorizzano dati numerici o quantitativi categorizzati su diverse dimensioni

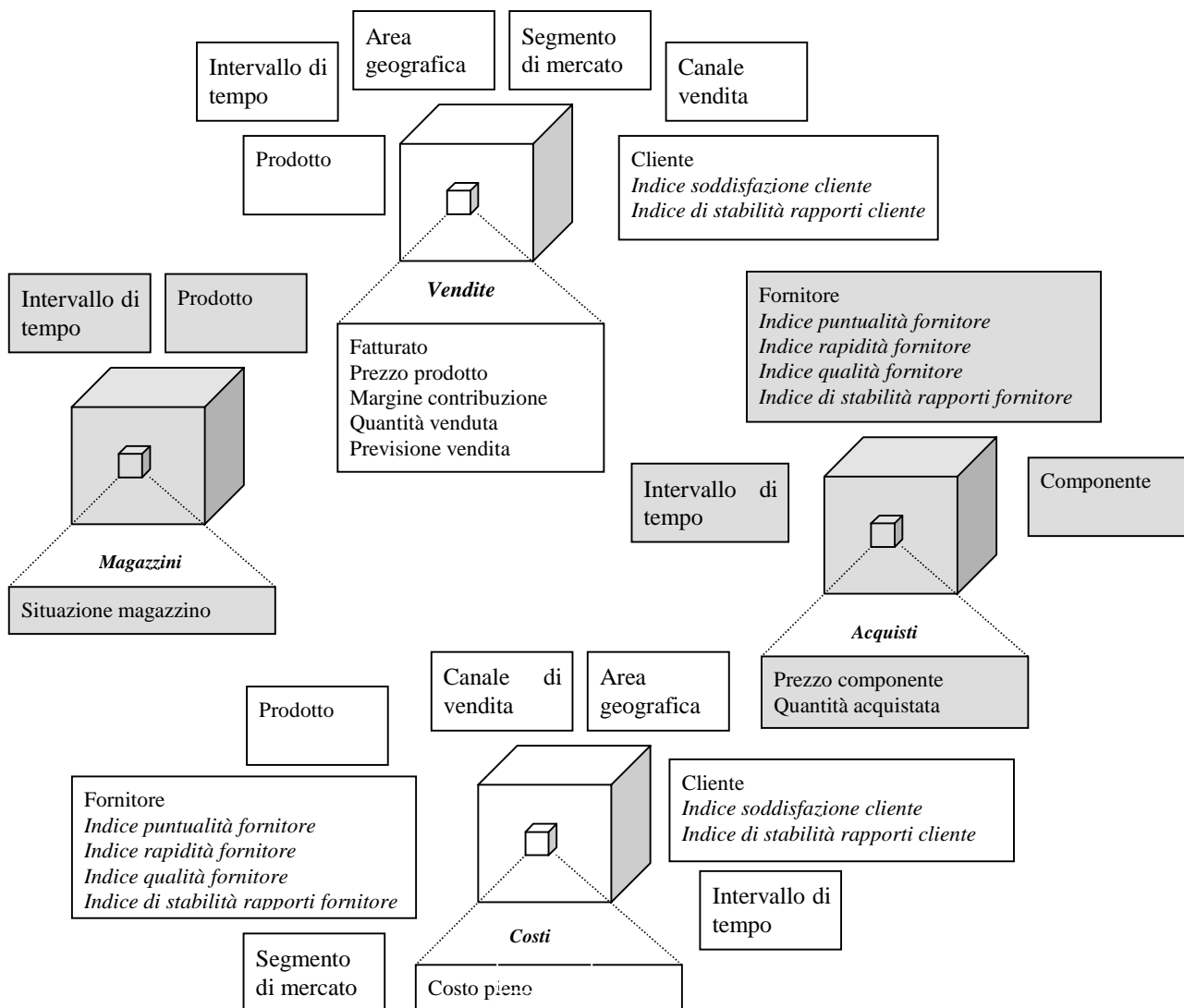


Figura 2. Modello concettuale finale del DW. Le gerarchie di aggregazione delle dimensioni di analisi non sono riportate.

qualitative (Finkelstein, 1995). Gli MDDDBMS non sono nuovi. Per circa vent'anni, il pacchetto software EXPRESS della IRI Software Inc. (Burlington, MA), ora di proprietà della Oracle, ha incluso un MDDDB. Il loro vantaggio è che sono ottimizzati per velocizzare e semplificare le interrogazioni, grazie a operazioni di pre-elaborazione. Il loro svantaggio principale, per quello che è lo stato dell'arte, è che non possono memorizzare grandi quantità di dati (attualmente dell'ordine dei 100GB). Se si tiene conto che un file di 200MB in ingresso a un MDDDBMS può occupare fino a 5GB per via delle pre-elaborazioni compiute, si comprende come questo sia un grosso limite di estendibilità (*scalability*) della soluzione proposta per un DW (Meyer & Cannon, 1998). A questi problemi di natura tecnica, si aggiungono la disponibilità presso le aziende di solo basi di dati relazionali, e la competenza limitata al modello relazionale da parte del personale tecnico che cura il sistema informativo aziendale preesistente. D'altronde, la ricerca condotta dal DPPI sulle aziende del settore agro-alimentare

pugliese ha rilevato che motivi principali dello scarso interesse e della scarsa propensione agli investimenti nel settore delle tecnologie dell'informazione risultano essere la necessità di possedere competenze specifiche per la gestione delle tecnologie o, parallelamente, la necessità di ricorrere ad ingenti investimenti complementari, ad esempio, in formazione del personale. Tutto ciò porta a prendere in considerazione la possibilità di trasformare un modello concettuale multidimensionale in un modello logico di tipo relazionale. Naturalmente la scelta di un DBMS relazionale per la modellazione di un DW, porta ad affrontare i problemi di simulazione di un approccio multidimensionale. Fra i vari schemi proposti in letteratura, come lo *schema a stella*, a *costellazione* e a *fiocco di neve*, si è preferito il secondo per via della condivisione di dimensioni di analisi fra gli ipercubi.

#### OB4. Progettazione, integrazione e/o implementazione di applicazioni a supporto delle decisioni

Le attività previste, ancora in corso, sono la definizione e la realizzazione di strumenti a disposizione dell'utente manager o del suo staff per analizzare i dati disponibili al fine di raggiungere un determinato tipo di business aziendale.

Gli strumenti sperimentati sono di tipo Decision Support ed in particolare:

- di *EIS (Executive Information System)* con generazione di rappresentazioni tabellari e grafiche dei dati in base a interrogazioni e
- di *data mining* con generazione di pattern mediante strumenti di analisi dei dati.

L'applicazione EIS, fortemente orientata al reporting di specifiche informazioni aziendali, consiste nella integrazione e personalizzazione di strumenti a basso costo presenti sul mercato in grado di garantire l'accesso alle informazioni di business utili. Le informazioni aziendali necessarie alla decisione saranno selezionate tramite delle richieste interamente guidate dall'utente manager, che conosce *in toto* come impostare la richiesta per avere determinati risultati. L'applicazione EIS è sviluppata da OR.

Lo strumento di data mining che si intende integrare è KDB2000, un sistema in corso di sviluppo presso il DIB che integra funzioni di preparazione, mining, e post-processing dei dati (Figura 3). In particolare le funzioni di preparazione dei dati consentono la connessione a un data base mediante lo standard ODBC, l'import di dati da file, la selezione di dati da elaborare mediante la formulazione guidata di una query SQL, la trasformazione di dati mediante operatori di discretizzazione (supervisionata e non) e binarizzazione. Le funzioni di mining supportate dal sistema sono l'induzione di alberi di decisione (Esposito *et al.*, 1997) e di regressione (Esposito *et al.*, 1998), come pure la scoperta di regole di associazione. Infine, le funzioni di post-processing prevedono la trasformazione di alberi in regole e la visualizzazione dei risultati.

Una componente del tool verrà poi ritagliata sulla particolare applicazione, che, come si spiegherà nel prossimo paragrafo, richiede tecniche adatte a trattare dati spaziali o geo-referenziati.

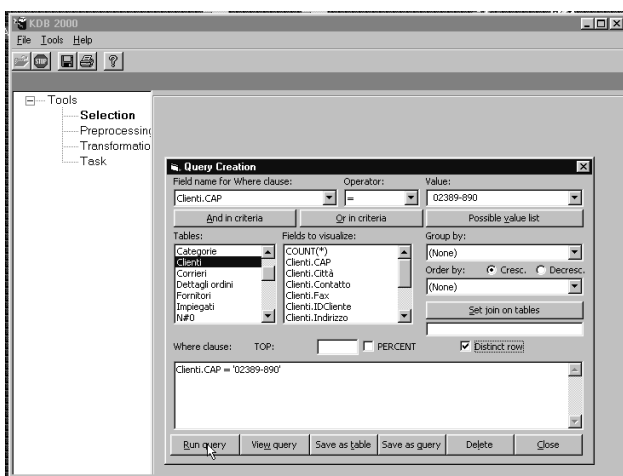


Figure 3. Selezione dei dati in KDB2000.

#### OB5. Sperimentazione della soluzione

Questa attività già in corso prevede l'installazione e la sperimentazione della soluzione presso l'Azienda Pilota "Gruppo Molino-Pastificio Tandoi Pellegrino" di Corato (BA).

## 4. IL PILOTA

Il Gruppo Molino-Pastificio Tandoi Pellegrino è l'azienda che sta sperimentando con OR e con DIB le tecniche di data warehousing, decision support system (DSS) e data mining nell'ambito della propria area "gestione clienti".

Dovendo perseguire delle linee guida che suggeriscono scelte tecniche semplici ma consistenti e con buone prestazioni, la soluzione che si sta sperimentando è quella di implementare un DW specializzato su di una singola area funzionale aziendale. Da un punto di vista tecnico, quindi, parleremo più correttamente di implementazione di un *data mart* (Gray & Watson, 1998).

La soluzione sistemica che progettiamo è tale però che successivamente altri data mart indipendenti specializzati su altre differenti e specifiche aree aziendali possano essere implementati e aggiunti al sistema in modo semplice e non traumatico.

Tutti i data mart implementati saranno integrati utilizzando quale filo conduttore i "dati" a cui fanno riferimento tramite le tecniche del data modeling.

Questo significa che il primo data mart e quelli successivi avranno in comune:

- alcune dimensioni e misure per i fatti analizzati. I fatti sono tutti gli eventi significativi individuabili in un'area di business mentre le dimensioni sono i parametri che si utilizzano per analizzare questi fatti ed infine le misure sono gli attributi numerici del fatto;
- alcune regole di business;
- una semantica;
- un'identificazione dei dati sorgenti;
- un'architettura;
- un DW administration centralizzato.

In particolare la soluzione che si sta implementando per costruire un data mart per la gestione clienti è tale che, seguendo proprio l'attuale modello di gestione clienti usato in azienda, renda fruibili informazioni già disponibili elettronicamente su basi di dati integrandole fra di loro, le arricchisca con altre informazioni, attualmente disponibili solo in formato cartaceo, ed infine supporti il manager con applicazioni decisionali (EIS e DM) che lo aiutino ad analizzare il patrimonio informativo relativo ai clienti.

Alla fine del progetto questo patrimonio informativo dovrà raccogliere dati sui clienti dell'azienda, sui prodotti, sul volume delle vendite e sulla loro redditività; raccoglierà, inoltre, dati sulle potenzialità del mercato di riferimento dell'azienda.

Le informazioni gestite provengono:

- dal sistema informativo transazionale aziendale,
- da banche dati esterne,
- da informazioni acquisite da campagne specifiche di rilevamento o da canali aziendali (per esempio tramite agenti o tramite personale del Molino preposto periodicamente ad una attività di rilevamento dati)
- da dati disponibili in azienda e non ancora integrati con i dati del sistema informativo (come ad esempio i dati di laboratorio).

Quando sarà completamente disponibile (giugno 2001), questo patrimonio informativo costituirà la base dati di un sistema di supporto alle decisioni per le attività commerciali e di marketing di gestione dei clienti e della loro fidelizzazione, di definizione delle strategie di vendita e di pianificazione di nuovi prodotti.

L'azienda avrà, così, la possibilità di analizzare una considerevole quantità di dati, avvalendosi nel futuro ed in prospettiva per implementare delle applicazioni che effettuino varie tipologie di analisi, quali:

- analisi del rischio;
- analisi del posizionamento dei clienti;
- analisi della redditività dei clienti e dei prodotti;
- analisi del valore del cliente;
- analisi dei differenti segmenti di mercato e delle linee di offerta;
- analisi delle potenzialità di mercato;
- analisi sulla propensione all'abbandono da parte dei clienti.

Tra tutte le possibili applicazioni di supporto alle decisioni per la gestione clienti, le analisi direzionali che si stanno implementando nell'ambito del progetto Datalight, sono le seguenti:

- Analisi del posizionamento dei clienti, in termini di fatturato realizzato con l'azienda Tandoi Pellegrino e rapporto del fatturato realizzato con le capacità produttive del cliente. Quest'analisi consentirà di rilevare se la Tandoi è un fornitore primario o di emergenza per i suoi clienti. Fornirà, inoltre, indicatori utili per meglio indirizzare le attività commerciali e di marketing verso clienti già acquisiti che dimostrano di possedere le maggiori potenzialità di crescita. I dati relativi al fatturato ed alle capacità produttive dei clienti saranno aggregati per prodotto e per famiglia di prodotti. Sarà possibile analizzare l'andamento nel tempo del fatturato dei clienti nel corso degli ultimi tre esercizi finanziari dell'azienda. Questa tipologia di analisi potrà essere effettuata per singolo cliente, per tipologia di cliente, per zona di appartenenza, per agente, per prodotti e famiglia di prodotti.

- Analisi dell'affidabilità dei clienti. Questa funzionalità fornirà un report contenente indicatori utili per valutare l'affidabilità dei clienti, sulla base di dati provenienti da fonti diverse, come:
  - dati acquisiti mediante la compilazione della scheda informativa clienti (per esempio le dimensioni aziendali)
  - dati circa la situazione patrimoniale ed economica dei clienti ed altre informazioni (come informazioni sulla proprietà di macchinari e strutture) acquisiti o da canali diretti con i clienti ovvero da fonti esterne se possibile (Camera di Commercio, società di recupero crediti)
  - dati del sistema informativo (informazioni sulla puntualità di pagamento)

Analisi del costo globale della gestione dei clienti. Il sistema consentirà di controllare

il costo di gestione dei clienti appartenenti alla grande distribuzione, fornendo dati riassuntivi relativi alle promozioni ed ai premi di fine anno loro assegnati, ai volumi di fatturato ed ai margini con loro realizzati.

- Analisi delle potenzialità del mercato. Ricorrendo a dati disponibili presso le camere del commercio e/o presso organizzazioni di categoria, si dovranno individuare i potenziali clienti (panetterie, industrie dolciarie, pasticci, etc.) dell'azienda pilota. I clienti potranno essere classificati, oltre che per il tipo di attività, anche per fatturato complessivo e fatturato nei confronti dell'azienda pilota. Mediante una visualizzazione spaziale (Figura 4), sarà possibile comprendere la loro distribuzione geografica, stabilire se ci sono zone sottodimensionate nel numero di agenti, stimare la fetta di potenziale mercato coperta in ogni zona, valutare la distribuzione degli agenti per potenziale volume d'affari stimato e in ultima analisi aiutare a ridefinire le zone da affidare ai vari agenti. L'individuazione di aree omogenee per presenza di potenziali clienti e per percentuale di copertura potrà essere raggiunta mediante l'applicazione di algoritmi di clustering spaziale.

## 5. RINGRAZIAMENTI

Questa attività di ricerca è condotta nell'ambito del progetto POP "Datalight: uno strumento di innovazione per le Piccole e Media Imprese in Puglia", finanziato dalla regione Puglia. Si ringrazia il Dott. Luca Sanarico per la collaborazione al progetto durante il periodo di permanenza presso il laboratorio LACAM del Dipartimento di Informatica di Bari.

## 6. BIBLIOGRAFIA

Agrawal, R., Gupta, A., & Saawagi, S. (1995). Modeling Multidimensional Databases, <http://citeseer.nj.nec.com/agrawal95modeling.html>

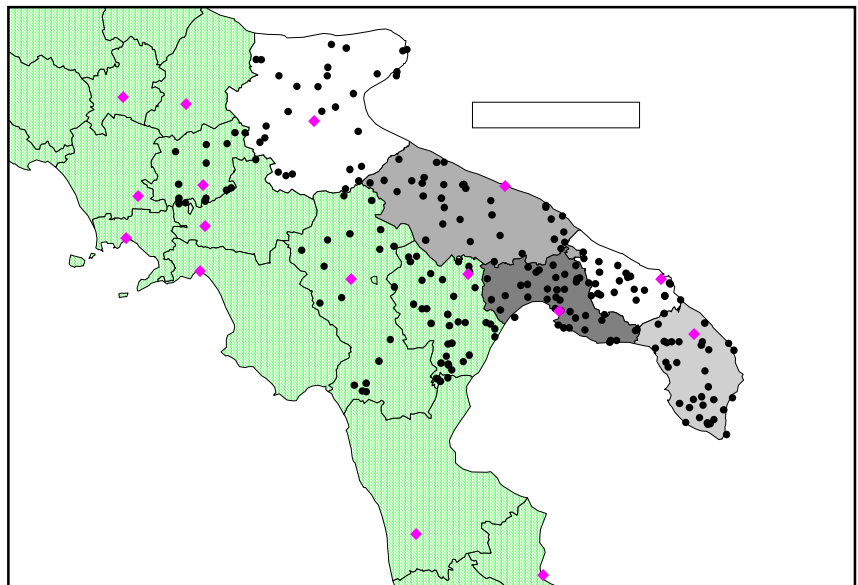


Figura 4. Una mappa di distribuzione di potenziali clienti sul territorio.

- Dioguardi, G., Gorgoglione, M., & Pontrandolfo, P. (2000). Analisi dei processi decisionali delle PMI operanti in Puglia nel settore agro-alimentare. Rapporto progetto POP "Datalight: uno strumento di innovazione per le Piccole e Medie Imprese in Puglia".
- F. Esposito, D. Malerba, & G. Semeraro (1997). A Comparative Analysis of Methods for Pruning Decision Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-19, 5, 476-491.
- F. Esposito, D. Malerba, & V. Tamma (1998). Efficient Data-Driven Construction of Model-Trees. Proceedings of NTTS'98, Int. Seminar on New Techniques & Technologies for Statistics, 163-168.
- Esposito, F., & Malerba, D. (2000). Modellazione dei dati di un data warehouse per il settore agro-alimentare. Rapporto progetto POP "Datalight: uno strumento di innovazione per le Piccole e Medie Imprese in Puglia".
- Fayyad, U. , Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery: an overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, pp. 1-35, 1996.
- Finkelstein, R. (1995). MDD: Database reaches the next dimension. *Database Programming and Design*, pp. 27-38, April.
- Gray, P., Watson, H.L. (1998). *Decision support in the Data Warehouse*, Prentice Hall PTR, New Jersey.
- Inmon, W.H. (1992). *Building the data warehouse*. Wellesley, MA: QED Tech. Pub. Group.
- Kimball R. (1996). *The data warehouse toolkit*, John Wiley & Sons.
- Kini, A. (1996). *A Multidimensional Data Model For Data Warehousing Applications*, <http://hsb.baylor.edu/html/ramsower/ais.ac.96/papers/KINI.htm>
- Meyer, D., & Cannon, C. (1998). *Building a better Data Warehouse*, Prentice Hall PTR, New Jersey.
- SQL Server 7.0 Data Warehousing training. Microsoft Press, (2000).
- Watson, H.J., and Haley, B.J. (1997). Framework for developing a Data Warehouse," *Journal of Data Warehousing.*, 2(1), 10-17.