DENALI: il *Data Warehouse* di Sanità Pubblica della Regione Lombardia

Giancarlo Cesana [1], Carla Fornari [1], Virginio Chiodini [1], Fabiana Madotto [1], Luca Merlino [2], Lorenzo G. Mantovani [1,3]

[1]CESP, Università di Milano Bicocca, Monza
[2]Direzione Generale Sanità, Regione Lombardia, Milano
[3]CIRFF, Università degli Studi di Napoli Federico II, Napoli

Le nostre capacità di raccogliere e generare dati sono aumentate rapidamente negli ultimi decenni. Fattori determinanti sono l'esteso uso di codici a barre per i prodotti commerciali, l'informatizzazione delle aziende e delle transazioni di lavoro e la disponibilità sempre maggiore di strumenti di raccolta dati, dagli scanner alle piattaforme per immagini ai sistemi di registrazione satellitare. Inoltre non bisogna dimenticare la diffusione ormai veramente popolare di Internet. Questa crescita esplosiva degli archivi di dati ha prodotto un'esigenza sempre più pressante di tecniche e strumenti automatici che possano con intelligenza assistere all'opera di trasformazione di questa quantità enorme di dati in informazioni e conoscenza utilizzabili.

Il *Data mining,* noto anche come *Knowledge Discovery in Databases (KDD)*, è l'estrazione automatica od opportunistica di conoscenze implicitamente contenute nei grandi database o in altri "depositi" di grandi masse di informazioni. *Data mining* è un campo multidisciplinare, che utilizza metodi e contenuti derivati dalla tecnologia dei database, dall'intelligenza artificiale, dalle tecnologie di apprendimento automatico (*Machine Learning*), dalle reti neurali, dalla statistica, dal riconoscimento di schemi, dai sistemi esperti, dalle tecnologie di reperimento di informazioni da dati non strutturati e dalle tecnologie di elaborazione ad alte prestazioni e di visualizzazione dei dati. Il *Data mining*, emerso verso la fine degli anni Ottanta, ha compiuto progressi notevoli negli anni Novanta e certamente continuerà a fiorire negli anni futuri.

IL DATABASE DI SANITÀ PUBBLICA DELLA REGIONE LOMBARDIA

Il successo delle tecniche di *Data mining* ha un fondamento: i dati. Il database di Sanità Pubblica della Regione Lombardia contiene tre archivi principali dove vengono registrati eventi di assistenza sanitaria:

- AMBULATORIALE – Visite specialistiche, analisi ambulatoriali e trattamenti terapeutici eseguiti presso strutture sanitarie;
- FARMACEUTICA – Prescrizioni farmaceutiche erogate da farmacie in Lombardia;
- RICOVERI – Ricoveri ospedalieri, diagnosi e interventi chirurgici;

Contiene inoltre archivi di riferimento per l'interpretazione dei dati contenuti negli archivi principali:

- ASSISTIBILI – Archivio anagrafico storico dei residenti e degli assistiti in Lombardia;
- MEDICI – Anagrafe dei medici;
- STRUTTURE – Archivio delle strutture sanitarie lombarde;
- FARMACIE – Archivio delle farmacie;
- FARMACI and ATC (*Anatomical Therapeutic Chemical Classification*) – Archivio farmaci;
- CODICI DIAGNOSI – Codici diagnostici fissati dal comitato internazionale;
- CODICI DI PROCEDURE CHIRURGICHE;
- DRG & MDC - *Diagnosis-Related Groups and Major Diagnostic Categories*;
- altri 40 archivi contenenti descrizioni del territorio e di codici usati negli archivi principali.

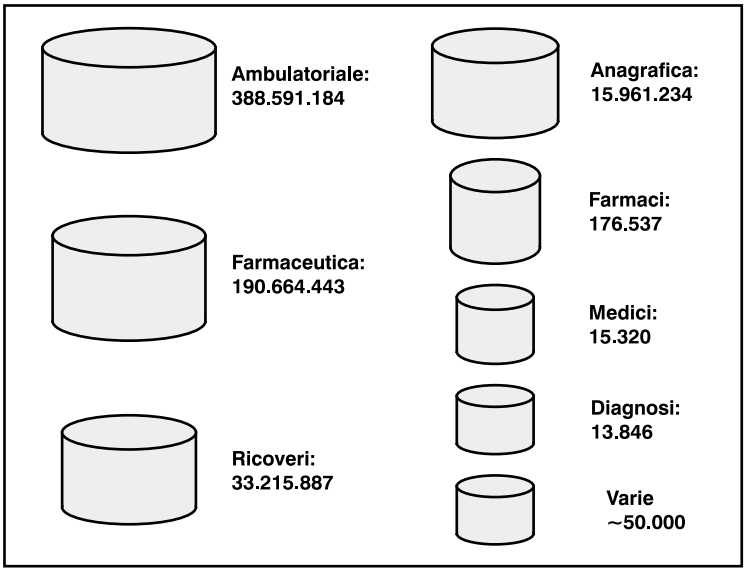

Figura 1
Principali tabelle dati del database sanitario lombardo e relativo numero di record (righe). Le dimensioni delle tabelle si riferiscono ai dati del periodo 2003-2005

I dati registrati nei tre archivi principali si riferiscono a eventi relativi al periodo 2003-2005. Gli altri archivi si basano su dati del 2002 aggiornati periodicamente negli anni successivi.

Il database di Sanità Pubblica della Regione Lombardia è processato da un database relazionale Oracle. I dati sorgenti usati per la costruzione del *Data Warehouse* contengono oltre 700 milioni di record. Nella Figura 1 sono illustrate le principali tabelle dati del database sanitario lombardo e il relativo numero di record (righe). Le dimensioni delle tabelle si riferiscono ai dati del periodo 2003-2005.

Il database sanitario lombardo è affetto da tutti i problemi tipici di qualità dei dati:

- strutture di supporto eterogenee:
  - tabelle relazionali con criteri di normalizzazione disparati;
  - nomenclatura dei campi disomogenea. Per esempio, il campo contenente il codice fiscale è denominato, nelle tre tabelle principali, "Codice_Fiscale", "Cod_Fisc_Assto", "Cd_Fisc_Cit";
- formati disomogenei: per esempio, le date sono espresse a volte in formato Oracle standard ("TO_DATE ('28-11-2005'...) e a volte in formato stringa di caratteri: "28-11-2005";
- incoerenza: i costi sono espressi a volte in Euro e a volte in centesimi di Euro; gli anni sono stati registrati su 4 o 2 caratteri all'interno dello stesso archivio: TO_DATE ('28-11-05'...) anziché TO_DATE ('28-11-2005'...);
- duplicazione dei dati: la gravità del problema della duplicazione è illustrata dall'esempio di 10 codici farmaco duplicati all'interno dello stesso gruppo ATC. Un'analisi che utilizzasse una giunzione relazionale su 2.000.000 di prescrizioni all'interno di questo gruppo ATC produrrebbe 20.000.000 di eventi duplicati. Al costo di € 2.00 per prescrizione l'errore causerebbe una sovrastima di costo pari a € 40.000.000;
- errori di digitazione. Per esempio: Codice Regionale 4LM**J**45VFR invece di 4LM**K**45VFR**.**; Codice Fiscale GRTNNT25R56G273**F** invece di GRTNNT25R56G273**R**; Codici diagnostici: 1142 (Coccidioidal Meningitis) anziché 01142 (Lung Tubercular Fibrosis);
- sconnessione tra i dati: la mancata registrazione o la digitazione incorretta di codici di concatenamento (per esempio, Codice Regionale) tra tabelle del database rende impossibile il reperimento di informazioni chiave nelle tabelle di riferimento (per esempio, l'archivio anagrafico) causando una perdita di dati utilizzabile per l'analisi. Si assume comunemente che una perdita di dati superiore al 20% possa influire sulla correttezza dell'analisi introducendo un *bias* statistico.

DENALI

La creazione di un *Data Warehouse* utilizza un approccio statico o "update driven" integrando anticipatamente informazioni provenienti da sorgenti molteplici ed eterogenee e "immagazzinandole" per supportare interrogazioni dirette e analisi dei dati. Diversamente dalle basi di dati processate con transazioni in tempo reale, il *Data Warehouse* non contiene informazioni aggiornate. La funzione del *Data Warehouse* è di offrire elevate prestazioni di accesso al database risultante dall'integrazione di sorgenti disomogenee. Per questo i dati vengono estratti, preprocessati, integrati e ristrutturati in un unico deposito. Il *Data Warehouse* integra dati estratti da molteplici sorgenti eterogenee in una struttura unificata ordinata secondo un modello semantico di dati particolarmente adatto a supportare interrogazioni strutturate o estemporanee, generazione di rapporti analitici e processi decisionali.

L'obiettivo del *Data Warehouse* di Sanità Pubblica è di consolidare i dati di sanità pubblica raccolti dalla Regione Lombardia in un unico deposito coerente e accurato, adeguato a essere messo a disposizione di molteplici comunità di ricerca per diversi tipi di analisi e di supporto decisionale.

DENALI STUDIO è un sistema software disegnato e sviluppato per la creazione e la manutenzione del *Data Warehouse* della Sanità Pubblica della Regione Lombardia.

DENALI supporta il lavoro di creazione del *Data Warehouse* offrendo un'estesa interfaccia grafica che facilita l'analisi e la gestione dei dati (Figura 2). DENALI offre due funzionalità qualificanti:

- il *Probabilistic Record Matching* (PRM) che permette la ricostruzione di connessioni tra righe di tabelle diverse corrotte da errori di digitazione. Questa funzionalità elimina o quantomeno riduce notevolmente il *bias* statistico prodotto dalla perdita di dati. I risultati preliminari indicano che il PRM è stato in grado di recuperare oltre il 30% di connessioni danneggiate;
- la definizione grafica di "viste" standard e lineari. Queste viste nascondono la complessità strutturale della *Data Warehouse* e permettono ai ricercatori statistici di analizzare i dati senza dover affrontare la sintassi di espressioni di interrogazione complesse.

L'accesso al *Data Warehouse* di Sanità Pubblica è strettamente controllato dall'Authority della Regione preposta alla Privatezza per prevenire accessi non autorizzati a dati

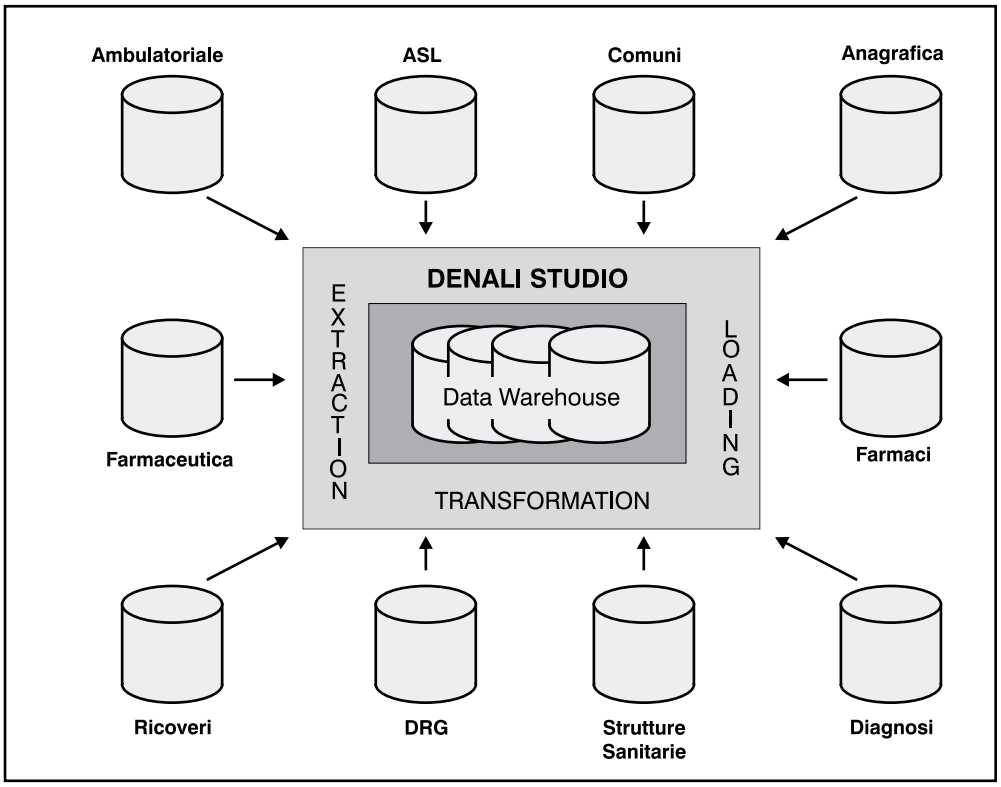

Figura 2
Struttura di DENALI

sensibili. Gli standard di privatezza sono regolati da apposite leggi regionali e nazionali. DENALI STUDIO offre le seguenti funzionalità:

- valutazione preliminare della qualità dei dati, identificando duplicazioni, inconsistenze, incompletezze e sconnessioni;
- integrazione dei dati, omogeneizzazione delle strutture e degli standard di nomenclatura dei campi;
- pulizia dei dati, attraverso le funzioni di deduplicazione e di ricostruzione probabilistica di connessioni mancanti tra gli archivi di dati;
- caricamento e aggiornamento del *Data Warehouse*;
- materializzazione delle "viste" per analisi specifiche.

La struttura funzionale di DENALI è illustrata in Figura 3.

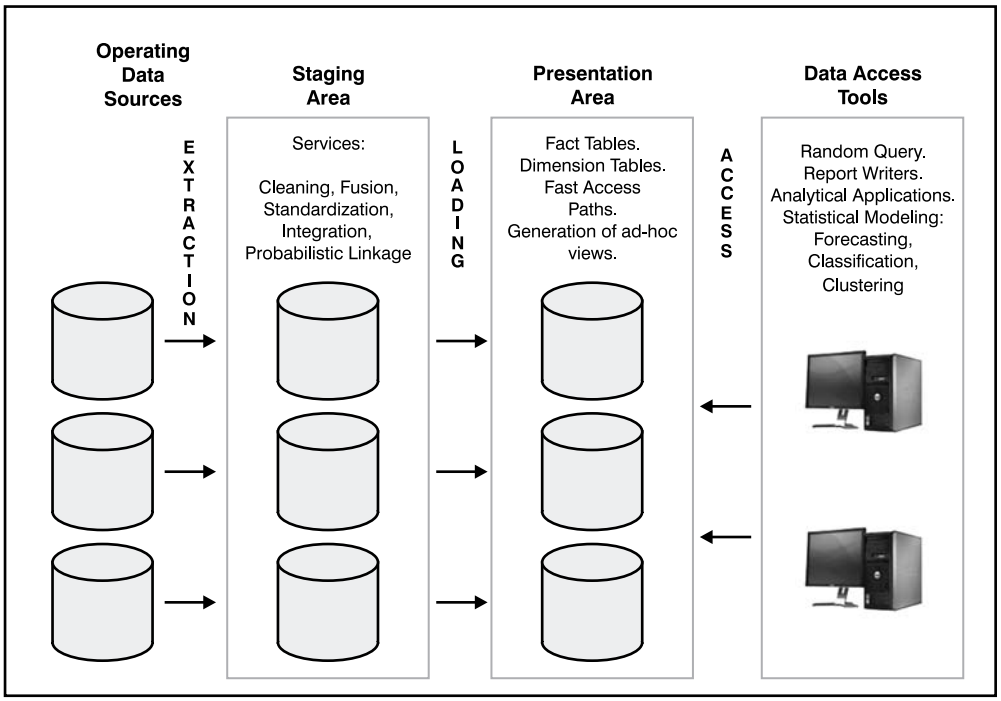

Figura 3
Struttura funzionale di DENALI

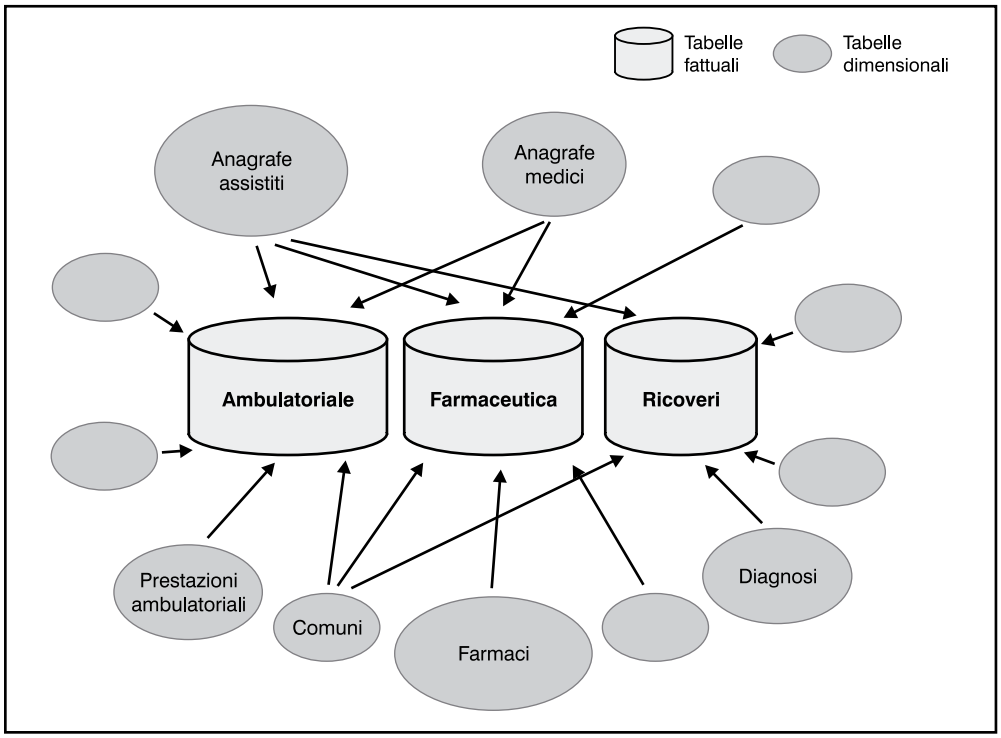

Figura 4
Schema a stella per Data Warehouse

MODELLO SCHEMATICO DEL *DATA WAREHOUSE* DI SANITÀ PUBBLICA

Un *Data Warehouse* richiede il disegno di un modello schematico conciso e specifico che faciliti l'analisi dei dati. Modelli multidimensionali di dati sono i più comuni per un *Data Warehouse*. Tali modelli possono assumere la forma di schemi a stella (*Star Schema*) o schemi a fiocco di neve *(Snowflake Schema*).

In uno schema a stella il *Data Warehouse* contiene un insieme di grandi archivi centrali (tabelle fattuali) che contengono il nucleo principale di dati, senza alcuna ridondanza, e un insieme di archivi ausiliari, uno per ogni dimensione informativa complementare ai dati degli archivi fattuali. La Figura 4 illustra lo schema a stella del *Data Warehouse* di Sanità Pubblica.

Lo schema a fiocco di neve (*Snowflake Schema*) è una variante dello schema a stella in cui alcu-

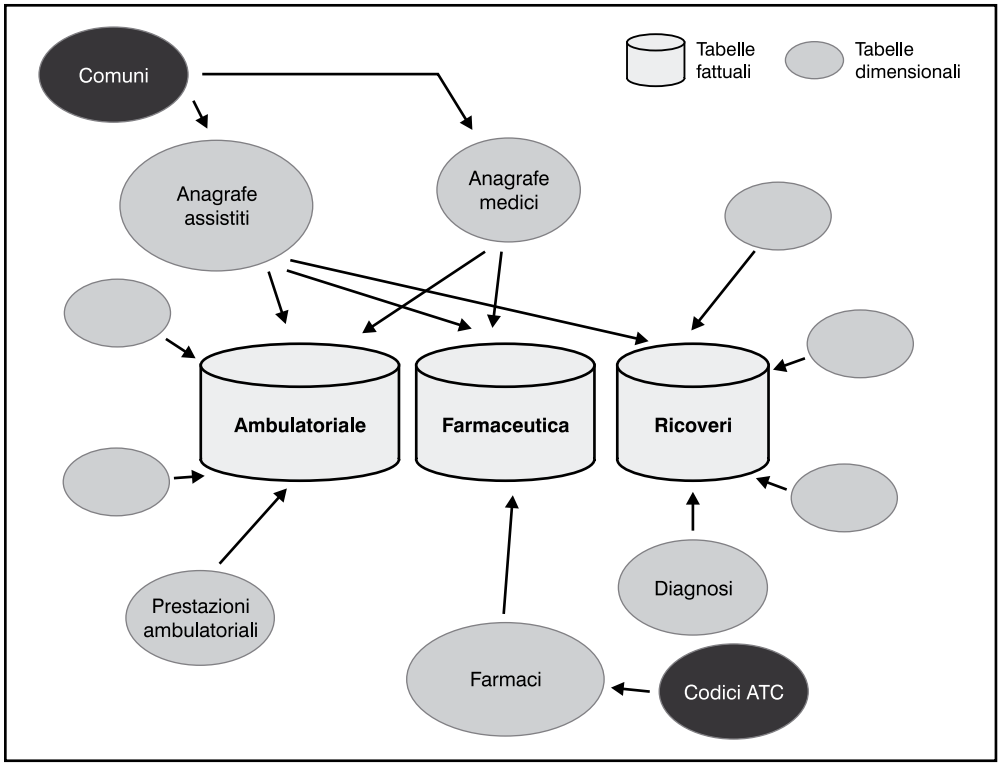

Figura 5
Schema a fiocco di neve per Data Warehouse

ne tabelle dimensionali sono normalizzate, suddividendo ulteriormente i dati in tabelle aggiuntive. Il grafico dello schema risultante ha una forma simile a un fiocco di neve. La differenza principale tra lo schema a stella e quello a fiocco di neve consiste nel fatto che le tabelle dimensionali in uno schema a fiocco di neve vengono normalizzate per ridurre ridondanze. Tabelle in questa forma sono agevoli da aggiornare e risparmiano spazio di registrazione. Tuttavia, il risparmio è trascurabile in confronto alle dimensioni tipiche delle tabelle fattuali. Inoltre lo schema a fiocco di neve può ridurre la velocità di esecuzione delle interrogazioni in quanto causa un aumento del numero di giunzioni di tabelle per ogni interrogazione, producendo un impatto negativo sulle prestazioni.

In Figura 5 è illustrato un ipotetico modello schematico a fiocco di neve del *Data Warehouse* di Sanità Pubblica.

Per il *Data Warehouse* di Sanità Pubblica della Regione Lombardia è stato adottato un modello schematico a stella.

RICOSTRUZIONE PROBABILISTICA DELLE CONNESSIONI

Una delle principali funzionalità che distinguono DENALI STUDIO è la ricostruzione probabilistica delle connessione tra dati degli archivi sorgenti, mancanti in seguito a errori di digitazione. La ricostruzione probabilistica delle connessioni è eseguita in tre fasi:

- stima statistica dei parametri elaborata attraverso un esame dei dati sorgenti da connettere;
- calcolo di valori di soglia per determinare se una coppia di record (righe) sia da connettere;
- calcolo del peso probabilistico di connessione per ciascuna coppia di record proveniente da due tabelle.

Per accelerare l'esecuzione delle intense elaborazioni, in DENALI la funzione di ricostruzione probabilistica delle connessioni è disegnata con un'architettura *multithreading* che, attraverso la generazione di processi paralleli ottimizzati, sfrutta appieno le capacità dell'hardware disponibile.

L'esempio di studio presentato in seguito in questo Supplemento ha utilizzato il *Data Warehouse* 2003-2005. Attualmente DENALI immagazzina i dati 2000-2009, vale a dire oltre 90 milioni di anni persona.